Learning Identifiable Representations: Independent Influences and Multiple Views

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Luigi Gresele

aus Perugia/Italien

Tübingen
2022

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	19.06.2023
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Bernhard Schölkopf
2. Berichterstatter:	Prof. Dr. Philipp Hennig
3. Berichterstatter:	Prof. Dr. Ferenc Huszár

Learning Identifiable Representations: Independent Influences and Multiple Views

Luigi Gresele

Submitted in November 2022.
Defended in June 2023.

To Ada Rossi and Fernanda Simoncini.

Abstract

Alternative representations may differ in what entities or types of information they make explicit. As suggested by David Marr, this can be illustrated by the following example: the number thirty-seven can be represented as 37 in the decimal numeral system, and as 100101 in the binary one. What is made explicit in the decimal representation is the number's decomposition into powers of ten; in contrast, the binary representation makes explicit its decomposition into powers of two.

Information about the external world is often available to learning systems, both biological and artificial, only in an unstructured form: artificial networks trained for object recognition take collections of pixels as inputs; visual information processing in biological systems starts in photoreceptors, where incoming light is converted into biological signals. In both cases, some complex processing is required in order for certain aspects (e.g., the position, dimension and colour of objects in an image) to be made explicit and easily accessible. Central questions are then what information should be made explicit, and how to do so.

We consider two problems in representation learning. The first one is the *cocktail-party problem*, where a number of conversations happen in parallel in a room, and the task is to recover (or *separate*) the voices of the individual speakers from recorded mixtures—also termed *blind source separation*. The second one is what we call the *independent-listeners problem*: given two listeners in front of some loudspeakers, the question is whether, when processing what they hear, they will make the same information explicit, identifying similar constitutive elements. Rather than the reconstruction of a ground truth, what interests us here is that both process the same auditory signals, and we want to compare their representations thereof.

These questions can be studied with the approach of independent component analysis (ICA). This entails establishing whether, under some technical assumptions (most importantly statistical independence of the latent components, either unconditional or conditional on some other variable), representations can be *uniquely* specified—up to some ambiguities deemed tolerable, and except for a small number of corner cases. In technical terms, this corresponds to characterising *identifiability* of the model: in ICA, this is a central theoretical question and a prerequisite to the practical estimation of representations from data.

A key result of ICA theory is that, when the mixing is nonlinear, the model is provably nonidentifiable: in other words, the cocktail-party problem cannot be solved. A first question is therefore under what additional assumptions (ideally as mild as possible) the problem becomes identifiable; the following one is what estimation algorithms should be used.

The contributions presented in this thesis address these questions, and revolve around two main principles.

The first principle is to learn representation where the latent components influence the observations *independently*. Here the term "independently" is used in a non-statistical sense—which, inspired by causal inference and the principle of independent causal mechanisms (ICM), can be loosely thought of as absence of fine-tuning between distinct elements of a generative process. In the context of the cocktail-party problem, our independence postulate amounts to stating that the speakers' positions are not fine-tuned to the room acoustics and placement of the recording devices, or to each other.

Firstly, we formalise this principle as the condition that the columns of the Jacobian of the mixing function (which represent *influences* of the corresponding latent components on the observed mixtures) should be orthogonal. We call this independent mechanism analysis (IMA), and provide theoretical and empirical evidence that our approach circumvents a number of nonidentifiability issues arising in nonlinear blind source separation. Whereas ICA had already been useful in the context of causal inference, providing the backbone for successful causal discovery algorithm, this is to the best of our knowledge the first attempt to use ideas from causality to make progress in the difficult task of nonlinear blind source separation.

We then study a popular approach to unsupervised learning, variational autoencoders (VAEs), through the lens of independent mechanism analysis. VAEs provide an efficient way to train deep latent variable models by maximising a tractable, approximate variational approximation of the intractable, exact likelihood. While

VAEs are commonly used for representation learning, it is unclear why maximisation of this variational objective (the evidence lower bound, or ELBO) would be useful in that context, since maximising the exact likelihood corresponds to estimating a provably nonidentifiable model. We show that, in a regime which we term near-deterministic, Gaussian VAEs perform independent mechanism analysis: the difference between the exact likelihood and the ELBO (or ELBO gap) equals a regularisation term which favours VAE decoders with column-orthogonal Jacobians. We formally prove this for the near-deterministic regime, and show in experiments on synthetic and image data that VAEs uncover the true latent factors when the data generating process satisfies the IMA principle.

The IMA principle is expressed as a constraint on the Jacobian of the mixing function, and optimisation of functions of a Jacobian is a central problem in probabilistic modelling: for example, in deep density models, where the likelihood includes the log-determinant of the Jacobian. Because of this term, their likelihood-based training is computationally expensive. We propose a new approach for exact training of a class of deep density models. Based on relative gradients, we exploit the matrix structure of neural network parameters to compute updates efficiently even in high-dimensional spaces: the computational cost of the training is quadratic in the input size, in contrast with the cubic scaling of naive approaches. This is achieved without constraining the Jacobian to be triangular, in contrast to autoregressive normalizing flows.

Whereas in the first part of this thesis observations are modelled as independent and identically distributed (i.i.d.) draws from a given distribution, in the second part we investigate a different setting, based on the following principle: representations can be learned from *paired observations* or *views*, where mixtures of the same latent variables are observed, and they (or a subset thereof) are perturbed in one of the views. We call this the *multi-view setting*.

Our first result establishes identifiability for a multi-view nonlinear ICA model, where views are nonlinear mixtures of component-wise corruptions of the same latent sources. We present novel identifiability proofs showing that the mixing can theoretically be undone under the assumption of *sufficiently distinct views*: intuitively, the two views should be sufficiently different from one another, resulting in more information being available in totality than from each view individually. In contrast with the previous part of this thesis, which relied on constraints on the mixing function, the setting with paired observations allows identifiability for *any* invertible, nonlinear mixing, provided that multiple, sufficiently different noisy views are available. To the best of our knowledge, this is the first identifiability result for the nonlinear multi-view setting.

We then apply multi-view ICA to model group studies in neuroimaging: we consider settings where multiple subjects are exposed to the same experimental stimulus. Data from each subject are then modelled as mixtures of shared components, representing responses evoked by the common stimulus, plus subject-specific noise, accounting for each individuals' deviation from the shared response. Unlike in the previous contribution, the mixing is assumed to be linear and, contrary to most group-ICA procedures, the likelihood of the model is available in closed form. We develop an alternate quasi-Newton method for maximizing the likelihood, and demonstrate the usefulness of our approach on fMRI and magnetoencephalography (MEG) data, where our model demonstrates better sensitivity in identifying common sources among subjects than alternative methods, as well as lower between-session variability.

Finally, we study a widespread and successful approach to self-supervised learning, where representations are learned from the original images together with *augmentations*, where hand-crafted transformations are intended to leave the semantics of the data invariant.We formulate the augmentation process as a latent variable model by postulating a partition of the latent representation into a content component, which is assumed invariant to augmentation, and a style component, which is allowed to change. Unlike prior work on disentanglement and independent component analysis, we allow for both non-trivial statistical and causal dependencies in the latent space. We study the identifiability of the latent representation based on pairs of views of the observations and prove sufficient conditions that allow us to identify the invariant content partition up to an invertible mapping.

In the conclusion, we discuss the connections between identifiability in representation learning and causal inference; we comment on the significance of identifiability theory for current empirical practice in machine learning; and outline some potential directions to extend the works presented in this thesis.

Zusammenfassung

Alternative Darstellungen können sich darin unterscheiden, welche Einheiten oder Arten von Informationen sie deutlich werden. Wie von David Marr eindrücklich beschrieben, lässt sich das durch den Vergleich verschiedener Zahlensysteme veranschaulichen. Nehmen wir die Zahl *Siebenunddreißig*: Im binären Zahlensystem kann sie als 100101 dargestellt werden, im dezimalen als 37. Während die dezimale Darstellung die Zerlegung der Zahl in Zehnerpotenzen ausdrückt, stellt die binäre Darstellung ihre Zerlegung in Zweierpotenzen dar. Beide Versionen sind valide Darstellungen derselben Information aber für verschiedene Anwendungen ist die eine oder die andere zu bevorzugen.

Informationen über die Außenwelt stehen lernenden Systemen—sowohl künstlichen als auch biologischen—oft nur in unstrukturierter Form zur Verfügung: Künstliche neuronale Netze, die für die Objekterkennung trainiert werden, nehmen Informationen in Form von Pixeln aus Bildern auf; die visuelle Informationsverarbeitung in biologischen Systemen beginnt in den Photorezeptoren, wo das einfallende Licht in biologische Signale umgewandelt wird. In beiden Fällen ist eine komplexe Verarbeitung dieser Eingaben erforderlich, damit bestimmte Aspekte (z.B. die Position, Größe und Farbe von Objekten in einem Bild) offengelegt und leicht zugänglich gemacht werden können. Die zentrale Frage ist daher, welche Informationen extrahiert werden sollen und wie dies geschehen kann.

Wir betrachten zwei Probleme des Repräsentationslernens. Das erste ist das sogenannte Cocktail-Party-Problem, bei dem eine Reihe von Gesprächen parallel in einem Raum stattfinden und die Aufgabe darin besteht, die Stimmen der einzelnen Sprecher aus einer Aufzeichnung zu rekonstruieren (oder zu trennen) - auch als blinde Quellentrennung bezeichnet. Das zweite ist das so genannte Problem der unabhängigen Zuhörer: Bei zwei Zuhörern, die sich vor einer Reihe von Lautsprechern befinden, stellt sich die Frage, ob sie bei der Verarbeitung des Gehörten die gleichen Informationen herausfiltern, indem sie ähnliche Grundbausteine identifizieren. Im Gegensatz zum vorigen Problem geht es dabei nicht um die Rekonstruktion des wahren Signals, sondern um den Vergleich der Repräsentationen desselben auditiven Signals zwischen den beiden Zuhörern.

Diese Fragen können mit dem Ansatz der unabhängigen Komponentenanalyse (*engl.* independent component analysis - ICA) untersucht werden. Dies bedeutet, dass unter bestimmten technischen Annahmen (vor Allem der statistischen Unabhängigkeit der latenten Komponenten, möglicherweise bedingt durch eine andere Variable), Repräsentationen eindeutig spezifiziert werden können—bis zu einer gewissen Mehrdeutigkeit, die als tolerierbar erachtet wird, und mit Ausnahme einer kleinen Anzahl von Grenzfällen. Technisch gesehen entspricht dies der Charakterisierung der Identifizierbarkeit des Modells. Dies ist eine zentrale theoretische Frage der ICA und eine Voraussetzung für das praktische Lernen von Repräsentationen aus Daten.

Ein zentrales Ergebnis der ICA-Theorie ist, dass das Modell nachweislich nicht identifizierbar ist, wenn die Mischung nichtlinear ist. Dies bedeutet beispielsweise, dass das Cocktail-Party-Problem nicht gelöst werden kann. Die Frage ist daher, unter welchen zusätzlichen Annahmen (idealerweise so wenige wie möglich) das Problem identifizierbar wird und welche Lösungsmethoden verwendet werden können.

Die in dieser Arbeit vorgestellten Beiträge befassen sich mit diesen beiden Fragen und drehen sich um zwei Hauptprinzipien.

Das erste Prinzip besteht darin, Repräsentationen zu lernen, in denen die einzelnen latenten Komponenten einen voneinander unabhängigen Einfluss auf die Beobachtungen haben. Der Begriff unabhängig bezieht sich hierbei nicht auf statistische Unabhängigkeit sondern Unabhängigkeit im Sinne des Prinzips der unabhängigen kausalen Mechanismen (*engl.* independent causal mechanisms - ICM) aus dem Feld der kausalen Inferenz—als Abwesenheit von Feinabstimmungen zwischen verschiedenen Elementen eines generativen Prozesses. Im Kontext des Cocktail-Party-Problems bedeutet dieses Unabhängigkeitspostulat, dass die Positionen der Sprecher weder aufeinander, noch auf die Raumakustik oder die Platzierung der Aufnahmegeräte abgestimmt sind.

Zunächst formalisieren wir dieses Prinzip als die Bedingung, dass die Spalten der Jacobi-Matrix der Mischfunktion, die die Einflüsse der entsprechenden latenten Komponenten auf die beobachteten Mischungen darstellt, orthogonal sein sollten. Wir nennen dies unabhängige Mechanismusanalyse (*engl.* independent mechanism analysis - IMA) und liefern theoretische und empirische Beweise dafür, dass unser Ansatz eine Reihe von Problemen der Nicht-Identifizierbarkeit umgeht, die bei der nichtlinearen blinden Quellentrennung auftreten. Die ICA hat sich bereits im Kontext der kausalen Inferenz als nützlich erwiesen und bildet das Rückgrat für erfolgreiche Algorithmen zur kausalen Entdeckung. Dies ist unseres Wissens nach der erste Versuch, Ideen aus dem Feld der Kausalität zu nutzen, um Fortschritte bei der herausfordernden Aufgabe der nichtlinearen blinden Quellentrennung zu erzielen.

Anschließend untersuchen wir einen beliebten Ansatz für unüberwachtes Lernen, den Variational-Auto-Encoder (VAE), aus dem Blickwinkel der Analyse unabhängiger Mechanismen. VAEs bieten eine effiziente Möglichkeit, tiefe neuronale Netze als Modelle der latenten Variablen zu trainieren, indem sie eine lösbare, Variationsapproximation der nicht zugänglichen, exakten Likelihood maximieren. Obwohl VAEs häufig für das Lernen von Repräsentationen verwendet werden, ist unklar, warum die Maximierung dieser Zielfunktion (die untere Schranke der Evidenz oder ELBO) in diesem Zusammenhang nützlich sein sollte, da die Maximierung der exakten Likelihood dem Lernen eines nachweislich nicht identifizierbaren Modells entspricht. Wir zeigen, dass Gaußsche VAEs in einem System, das wir als nahezu deterministisch bezeichnen, eine unabhängige Komponentenanalyse durchführen: Die Differenz zwischen der exakten Likelihood und der ELBO entspricht einem Regularisierungsterm, der VAE-Decoder mit spaltenorthogonalen Jacobi-Matrizen begünstigt. Wir beweisen dies formal für das nahezu deterministische Regime und zeigen in Experimenten mit synthetischen (Bild-)Daten, dass VAEs die wahren latenten Faktoren identifizieren, wenn der Datengenerierungsprozess das IMA-Prinzip erfüllt.

Das IMA-Prinzip wird als Einschränkung der Jacobi-Matrix der Mischfunktion ausgedrückt. Die Optimierung von Funktionen einer Jacobi-Matrix ist ein zentrales Problem bei der probabilistischen Modellierung, beispielsweise bei tiefen Density-Modellen, bei denen die Likelihood die Log-Determinante der Jacobi-Matrix enthält. Aufgrund dieses Zusammenhangs ist Likelihood-basiertes Training im Allgemeinen rechenintensiv. Wir entwickeln einen neuen Ansatz für das exakte Training einer Klasse von tiefen Density-Modellen. Auf Grundlage relativer Gradienten nutzen wir die Matrixstruktur der Parameter des neuronalen Netzes, um Aktualisierungen selbst in hochdimensionalen Räumen effizient zu berechnen: Die Rechenkosten für das Training sind quadratisch mit der Größe der Eingaben, im Gegensatz zur kubischen Skalierung naiver Ansätze. Dies wird erreicht, ohne Triagonalität der Jacobi-Matrix zu erfordern, wie es bei autoregressiven Normalizing-Flows der Fall ist.

Während im ersten Teil dieser Arbeit die Beobachtungen als unabhängige und identisch verteilte (*engl.* independent and identically distributed - i.i.d.) Ziehungen aus einer gegebenen Verteilung modelliert werden, untersuchen wir im zweiten Teil eine andere Situation, die auf folgendem Prinzip beruht: Repräsentationen können aus gepaarten Beobachtungen oder Perspektiven gelernt werden, bei denen Mischungen der gleichen latenten Variablen beobachtet werden und diese (oder eine Teilmenge davon) in einer der Perspektiven gestört werden. Wir nennen dies die Multi-View-Situation.

Unser erstes Ergebnis beweist die Identifizierbarkeit eines nichtlinearen ICA-Modells mit mehreren Perspektiven, bei dem die Perspektiven nichtlineare Mischungen von komponentenweisen Störungen der selben latenten Variablen sind. Wir präsentieren neuartige Identifizierbarkeitsbeweise, die zeigen, dass die Vermischung theoretisch rückgängig gemacht werden kann, wenn hinreichend unterschiedliche Perspektiven angenommen werden: Intuitiv sollten sich die Perspektiven hinreichend voneinander unterscheiden, was dazu führt, dass in der Gesamtheit mehr Informationen zur Verfügung stehen als aus jeder einzelnen Perspektive. Im Gegensatz zum vorangegangenen Teil dieser Arbeit, der sich auf Beschränkungen der Mischfunktion stützte, ermöglicht die Betrachtung von gepaarten Beobachtungen die Identifizierbarkeit für jede invertierbare, nichtlineare Mischung, vorausgesetzt, dass mehrere, hinreichend unterschiedliche, verrauschte Perspektiven verfügbar sind. Soweit wir wissen, ist dies das erste Identifizierbarkeitsergebnis für die nichtlineare Multiview-Situation.

Anschließend wenden wir die Multiview-ICA an, um Gruppenstudien im Neuroimaging zu modellieren: Wir betrachten Situationen, in denen mehrere Probanden demselben experimentellen Stimulus ausgesetzt sind.

Die Daten der einzelnen Versuchspersonen werden dann als Mischungen aus gemeinsamen Komponenten modelliert, die die durch den gemeinsamen Stimulus hervorgerufenen Reaktionen darstellen, sowie aus subjektspezifischem Rauschen, das die Abweichungen der einzelnen Personen von der gemeinsamen Reaktion berücksichtigt. Anders als im vorherigen Beitrag wird die Mischung als linear angenommen und im Gegensatz zu den meisten Gruppen-ICA-Verfahren ist die Likelihood des Modells in geschlossener Form verfügbar. Wir entwickeln eine alternative Quasi-Newton-Methode zur Maximierung der Likelihood und demonstrieren die Nützlichkeit unseres Ansatzes anhand von fMRI- und Magnetoenzephalographie (MEG)-Daten, bei denen unser Modell eine bessere Sensitivität bei der Identifizierung gemeinsamer Quellen zwischen den Probanden, sowie eine geringere Variabilität zwischen den Sitzungen zeigt als alternative Methoden.

Schließlich untersuchen wir einen weit verbreiteten und erfolgreichen Ansatz des selbstüberwachten Lernens, bei dem Repräsentationen aus Originalbildern zusammen mit Augmentationen gelernt werden, bei denen die von Hand vorgenommenen Transformationen dazu dienen, die Semantik der Daten unverändert zu lassen. Im Gegensatz zu früheren Arbeiten zu Disentanglement und unabhängigen Komponentenanalyse erlauben wir sowohl nicht-triviale statistische als auch kausale Abhängigkeiten im latenten Raum. Wir untersuchen die Identifizierbarkeit der latenten Repräsentation anhand von Paaren von Beobachtungen und beweisen hinreichende Bedingungen, die es uns erlauben, die invariante inhaltliche Partition bis zu einer invertierbaren Abbildung zu identifizieren.

Im Schlussteil diskutieren wir die Verbindungen zwischen Identifizierbarkeit beim Repräsentationslernen und kausaler Inferenz. Wir kommentieren die Bedeutung der Theorie der Identifizierbarkeit für die aktuelle empirische Praxis des maschinellen Lernens und skizzieren einige mögliche Richtungen, um die in dieser Arbeit vorgestellten Arbeiten zu erweitern.

Acknowledgments

The past years have been a fascinating, long journey, full of intellectual adventures and emotional moments: it seems justified to have a long list of people I would like to thank.

Bernhard Schölkopf, for pushing me to investigate fundamental questions in machine learning; for many thought-provoking conversations and enriching idea exchanges; and for his support and trust.

Aapo Hyvärinen, for hosting me in Paris; for his clarity, expressed both in his work and in the research discussions we had during the time we collaborated; and for his patience and understanding.

Philipp Hennig, Robert C. Williamson and Ferenc Huszár, for the time they dedicated to reviewing my work, and for their interesting comments and questions on my thesis.

Michel Besserve, for helping me right from the beginning of my PhD, for some great collaborations and for his unselfish dedication to research, which I find nothing short of heroic.

Dominik Janzing, for hosting me during my internship at Amazon, and for many insightul conversations and discussions on causality.

Julius von Kügelgen, for sharing a wonderful experience of learning and scientific discovery, a productive collaboration and a great friendship.

Francesco Locatello, Paul K. Rubenstein, Patrik Reizinger, Jack Brady, Yash Sharma, Wieland Brendel, Arash Mehrjou, Vincent Stimper, Giambattista Parascandolo, Alex Neitz, Antonio Orvieto, Jonas M. Kübler, Jonas Dehning, Viola Priesemann, Armin Kekić, Zhijing Jin, Yuen Chen, Felix Leeb. Thank you for your trust and dedication, it was great to collaborate with you.

Joanna Sliwa, Daphna Keidar, Shubhangi Ghosh, Liang Wendong, for giving me the chance to learn how to supervise someone, and for collaborating with me.

The Empirical Inference department (2017-2023)—including Timmy Gebhard, Heiner Kremer, Yassine Nemmour, Jonas M. Kübler, Dieter Büchler, Vincent Berenz, Matthias Bauer, Carl Johann Simon-Gabriel, Sergios Gatidis, Niki Kilbertus, Philipp Geiger, Sebastian Gomez-Gonzalez, Muhammad Waleed Gondal, Amir Hossein Karimi, Hamza Keurti, Shashank Singh, Si Kai Lee, Eduardo Pérez-Pellitero, Patrick Putzky, Sebastian Weichwald, Marina Munkhoeva, Nasim Rahaman, Frederick Träuble, Siyuan Guo, Sergio Garrido, Matej Zečević, Max Dax, Jonas Wildberger, Alex Immer, Giovanni Visona, Max Mordig, Ronan Perry, Cian Eastwood, Manuel Wüthrich, Sabrina Rehbaum, Ann-Sophie Bähr, Lidia Pavel... And many others, apologies to those I'm forgetting.

I am grateful I could meet many amazing people in the MPI for Biological Cybernetics—Gabriele Lohmann, Klaus Scheffler, Kai Buckenmaier, Jonas Bause, Alexander Loktyushin, Johannes Stelzer, Ju Young Lee, Eric Lacosse.

The Causality Team at Amazon (2021), including Atalanti Mastakouri, Elke Kirschbaum and all the others I met during my stay there.

Bertrand Thirion and the Parietal Team at INRIA (2019), where I spent an amazing time. I am grateful I could collaborate with Hugo Richard, Alexandre Gramfort and Pierre Ablin, from whom I learned a lot.

Many other researchers I am grateful I had the chance to meet during the last years—Isabel Valera, Antonio Vergari, Sara Magliacane, Kun Zhang, Uri Shalit, Taco Cohen, Ben Poole, Elias Bareinboim, Kartik Ahuja, Yixin Wang, Sébastien Lachapelle, Tristan Deleu, Hiroshi Morioka, Atticus Geiger, Riccardo Massidda, Fabio Massimo Zennaro, Tom Beer, Daniel Greenfeld, Bar Eini-Porat, Ilya Tolshtikhin, Mijung Park, Krikamol Muandet, Georgios Arvaniditis, Wittawat Jitkrittum, Caterina de Bacco, Martina Contisciani, Andy Keller, Imant Daunhawer, Alexander Marx, Alice Bizeul, Alizée Pace, Vitória Barin Pacela, Antonia Machlouzarides-Shalit, Hubert Banville, Omar Chehab, Teodora Pandeva, Eddie Cunningham, Ilyes Khemakhem, Ricardo Pio

Monti, Hermanni Hälva, Marco Fumero, Antonio Norelli, Luca Moschella, Irene Cannistraci and Emanuele Rodolà. I would also like to thank Matteo Marsili, for supporting me in the process of applying for the PhD, and for keeping in touch during last years and offering me new opportunitites.

There are a number of people who made life better through their invaluable friendship, including Cri, Diego, Sofia, Susi, John, Dani, Eric, Niklas, Sergey, Alex Gessner, Julius and Carina, Luca, Nicco, Ruggi, Margherita, Alexis, Claudio, Gianca, Simo, Paolino, Benedetta, Filippo, Marie Baur, Micheal, Marie and Isaac, Adrián, Emilio and many others.

I would also like to thank again those who proofread various parts of this thesis or gave me suggestions on the presentation—Fred Träuble, Daniela Leite, Cristina Pinneri, Heiner Kremer, Armin Kekić, Jack Brady, Patrik Reizinger, Julius von Kügelgen, Simon Buchholz, Choo Xianjun Davin, Timothy Gebhard.

My parents, Paolo and Cristina, my sister, Gemma, and my aunt Gabriella, for their love and support.

Finally, thank you, Sabrina and Viola, *"for making everything beautifuller"*.

Contents

Motivation: Learning Representations

1

In this chapter, we introduce and discuss two metaphors: the *cocktail-party problem* (§ 1.1) and the *independent-listeners problem* (§ 1.2). These are meant to illustrate without much technical overhead some of the questions studied in the rest of this thesis. In § 1.3, we will discuss the significance of the two metaphors in the context of representation learning. The terminology used in this chapter is largely nontechnical: in Chapter 2, we present a framework to formalise and solve these problems under suitable technical assumptions.

1.1 The Cocktail-Party Problem

Cocktail parties are usually noisy (successful ones at least), with multiple conversations taking place simultaneously in the same room. To successfully engage in social interaction, attendees should be able to discern the voices of individual speakers.

In its simplest version, the *cocktail-party problem* features two speakers talking simultaneously in a room. Two recording devices (microphones, or the ears of a listener) are placed in the same room. Due to the physics of sound propagation, the shape of the room and arragement of objects therein, and influenced by the position of the speakers relative to the recording devices, these pick up mixtures of the two voices.

If we call $\mathbf{s}(t) = (s_1(t), s_2(t))^\top$ the vector containing the soundwaves $s_i(t)$ emitted by the speakers at a given time t,[1] and denote the mixing process by $\mathbf{f}$, we may write the mixtures picked up by the microphones as the vector $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t))$.[2]

1: we take the transpose to stress that here and in the rest of this thesis we use column vectors.

2: In the following we will neglect the temporal structure of the signals and drop the time index t for simplicity.

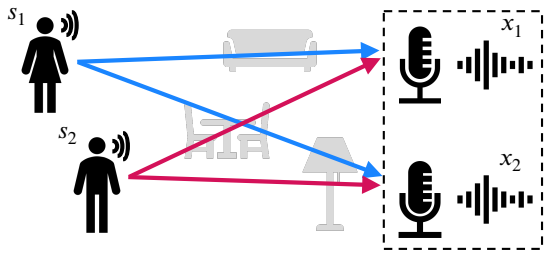


Figure 1.1: A visualisation of the cocktail-party problem.

We consider the following problem: **Given only the recorded mixtures $\mathbf{x}$, can we recover the individual voices of the speakers s_1 and s_2?**

This is also referred to as the problem of *blind source separation*: the voices of the individual speakers s_i are called *sources* and we want to separate them in the sense that we want to transform the mixtures $\mathbf{x}$ into $\mathbf{y} = (y_1, y_2)^\top$ such that each y_i component only retains one of the two voices.

Some comments on the metaphor above may be necessary:

(i) While the number of speakers and the number of recording devices in the example is arbitrary, we deliberately chose the two to be equal and we will assume that the mixing process involves no loss of information. See for more comments on this § 1.3.3.

(ii) The example is heavily stylised and neglects many realistic aspects of audio mixing. For example, the mixing of audio signals is often not instantaneous. See [1, Sec 24.2] for some of the complexities presented by the problem of audio separation uncaptured in the stylised metaphor above.

1.2 The Independent-Listeners Problem

Consider now two attendees of the cocktail party, Alice and Bob. Both of them are exposed to the same audio signals, e.g., music played over some loud-speakers. Suppose that each of them (or rather each brain or auditory pathway) tries to solve the blind source separation problem of § 1.1. Each will then produce their own attempt at source separation—$\mathbf{y}_{\text{Alice}}$ for Alice and $\mathbf{y}_{\text{Bob}}$ for Bob.

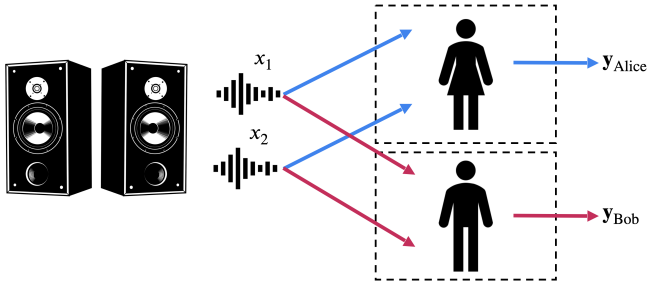


Figure 1.2: A visualisation of the independent-listeners problem.

We then ask: **Are the two vectors $\mathbf{y}_{\text{Alice}}$ and $\mathbf{y}_{\text{Bob}}$ necessarily equal? If not, how are they related?**

This question focuses on the *listeners*, not on the speakers—unlike in § 1.1. As long as we are solely interested in the $\mathbf{y}_{\text{Alice}}$ and $\mathbf{y}_{\text{Bob}}$ vectors, Alice and Bob may in fact be performing a different task from blind source separation. For the question to be interesting, trivial transformations of the input signals (e.g., mapping to a constant vector regardless of the input) should be ruled out though, and the task of blind source separation ensures this; different tasks, for example related to classification, may also work under suitable conditions. We will come back to this in § 2.4.2.

In fact, in this context we would like to abstract away the true generating process of the auditory signals (and wheteher blind source separation is achieved) as much as possible, and just think of $\mathbf{x} = (x_1, x_2)^\top$ as sounds emitted by some stereo loud-speakers. The main aspect that is of interest to us here is that both Alice and Bob engage in the same signal processing task, and we want to study and compare the outputs of these processes.

1.3 Learning Representations

We will now discuss the problems in § 1.1 and § 1.2 in the context of machine learning, and particularly representation learning.

Why learning? In blind source separation, listeners may not know in advance (may be blind to) what the speakers will say, what their locations are and what the precise arrangement of objects in the room and the mechanisms of sound propagation are: if these mechanisms were known precisely in advance, the listeners would simply have to invert it. However, due to their ignorance about these aspects, *learning* strategies not requiring prior specification of such details may be helpful to solve the problem.

1.3.1 What is a Representation?

According to David Marr [2],

> A *representation* is a formal system for making explicit certain entities or types of information, together with a specification of how the system does it.

[2]: Marr (2010), *Vision: A computational investigation into the human representation and processing of visual information*

If we want to interpret this definition based on the metaphors in § 1.1 and § 1.2, we could think of the vectors $\mathbf{y}$ and $\mathbf{y}_{\text{Alice}}$, $\mathbf{y}_{\text{Bob}}$ as *representations* of the observations, or data, $\mathbf{x}$. For example, in the context of § 1.1, the $\mathbf{x}$ vector may contain the same information as the $\mathbf{s}$ vector does,[3] but in $\mathbf{s}$ some information is made explicit (the separation between voices of the two speakers) with respect to $\mathbf{x}$. In blind source separation, we would like this same information to be made explicit in $\mathbf{y}$. Specifying *"how the system does it"* could then be interpreted as specifying (how the system learns) a function $\mathbf{g}$ such that $\mathbf{y} = \mathbf{g}(\mathbf{x})$ makes the kind of information we are interested in explicit (i.e., it separates the sources).

3: this is true in a formal sense if, for example, the transformation $\mathbf{f}$ is invertible, which we will often assume in this work.

Marr additionally argues that alternative representations may differ in what information they make explicit:

> The Arabic, Roman, and binary numeral systems are all formal systems for representing numbers. [...] What [the Arabic numeral system] makes explicit is the number's decomposition into powers of 10. The binary numeral system's description of the number thirty-seven is 100101, and this description makes explicit the number's decomposition into powers of 2.

For example, the decomposition of a number into powers of 2 (respectively 10) can be discussed based on either representation, but it is made explicit in the binary (respectively Arabic) numeral system.

Representations and downstream tasks. This naturally leads to the following consideration: How information is represented can greatly affect how *easy* it is to do different things with it [2].

> This is evident even from our numbers example: It is easy to add, to subtract, and even to multiply if the Arabic or binary representations are used, but it is not at all easy to do these things–especially multiplication–with Roman numerals. This is a key reason why the Roman culture failed to develop mathematics in the way the earlier Arabic cultures had.

In other words, different representations may be more or less suited for different things we may want to do with them, termed *downstream tasks*. These may or may not require source separation [3]: in the cocktail party, source separation may not be required to answer certain questions—for example, what is the overall perceived volume of the voices of the two speakers. In answering this question, and many others, some information may safely be discarded.

[3]: Eastwood et al. (2022), 'On the DCI Framework for Evaluating Disentangled Representations: Extensions and Connections to Identifiability'

However, the downstream task may be a priori unclear: cocktail party attendees may be unable to predict exactly what questions they will have to ask or answer.[4] It has therefore been argued that for an intelligent system in a complex environment (be it an artificial neural network or a biological party attendee) a good strategy could be to learn representations that *"discard as little information about the data as is practical"* while still *"disentangl[ing] as many factors [of variation] as possible"* [4]. In the context of § 1.1, if we think of the speakers' voices as factors of variation, we may then say that solving blind source separation (thereby *"disentangling"* the voices of the two speakers) can be a helpful first step to answer many questions about their voices or what they say (although not always a *necessary* one).

4: such unpredictability may be part of what makes social interactions interesting.

[4]: Bengio et al. (2013), 'Representation learning: A review and new perspectives'

In a different spirit, we could otherwise study the problem of representations in a task-dependent manner: What are the characteristics of, and how to learn, good or optimal data representations for a given task? In this context, one might want to compare representations extracted by similar systems performing the same task, much in the spirit of § 1.2, see also [5].[5] The degree of similarity between these representations may also affect how easy it is for intelligent agents to communicate about them: loosely speaking, if Alice and Bob later want to have a conversation about the $\mathbf{x}$ signals, communication would be easier if they *heard* similar things (although the *interpretation* or meaning they assign to what they hear may be quite different).[6]

5: If the data Alice and Bob apply their algorithms to, the initial conditions and the algorithms they use are equal and deterministic, no difference is to be expected among $\mathbf{y}_{\text{Alice}}$ and $\mathbf{y}_{\text{Bob}}$. However, in some cases, the problem may admit many equally good outputs or solutions, and slight perturbations of the aforementioned conditions may lead to large differences. We will articulate this more formally in § 2.4.1.

6: This difference between *hearing* and *interpreting* auditory signals mirrors a similar distinction in visual perception, where Kanizsa posited the separation between a low-level visual information processing termed *seeing* and a high-level one termed *thinking* [6].

1.3.2 Representations in Machine Learning

From a machine learning perspective, all of this might sound unnecessarily complicated. If we are interested in a downstream task, one might wonder why is it even necessary to talk about representations:[7] Should we not simply focus on developing algorithms which are good at solving the task we are interested in [8]?

7: Such focus on representations might appear as a violation of the so-called Vapnik principle *"never to solve a problem which is more general than the one we actually need to solve"* [7].

Deep learning and representation learning. One possible answer is that many successful machine learning algorithms—in particular deep neural networks [9]—are characterised by learning representations of the data which enable solving tasks of interest efficently. This takes inspiration from computational theories of the brain [10]:

[10]: Hinton (2007), 'Learning multiple layers of representation'

> To achieve its impressive performance in tasks such as speech perception or object recognition, the brain extracts multiple levels of representation from the sensory input.

Similarly, deep learning models learn representations of data with *"multiple levels of abstraction"* [9], and have the capability to learn representations which *"can entangle and hide more or less the different explanatory factors of variation behind the data"* [4].

[9]: LeCun et al. (2015), 'Deep learning'

[4]: Bengio et al. (2013), 'Representation learning: A review and new perspectives'

A metaphor such as the one in § 1.1 provides a way to interpret the sentence above: in the cocktail-party example, it is tempting to consider the voices of individual speakers as factors of variation. A representation which uncovers or "disentangles" them may be one which solves the problem of blind source separation.[8] So the study of representations and representation learning is related to deep neural networks and may suggest how to improve them, and it may in turn inspire computational neuroscience models.[9]

8: However, many other aspects of the perspective on representation learning presented in [4] are uncaptured by § 1.1, most prominently the so called "manifold hypothesis" which we will discuss in § 1.3.3. We also note that many other interpretations of disentanglement exist, see, e.g., [11–13] for some different perspectives.

9: for an example in the context of how representation learning has provided computational models for early vision, see, e.g., [14].

1.3.3 *What* is Represented and *How*

In the cocktail-party example of § 1.1, we considered equal number of microphones and speakers and and no information loss. However, in many settings, representation learning may involve different dimensionalities of the observed and latent variables, and information may be compressed or discarded in the representations. This is true for representation learning both in the brain and in machine learning. For example, in the context of human perception, there is evidence that a lot of information is discarded by the visual system.[10] In machine learning, the *manifold hypothesis* [4, 17, 18] posits that in many problems of interest observations lie in the vicinity of a low-dimensional manifold embedded in a higher-dimensional space; see also [19] for a related hypothesis in the context of physics and biochemistry. Classic statistical methods suchs as Principal Component Analysis (PCA) [20, 21] can be used to find a low-dimensional representation while preserving as much information as possible, in a task-agnostic manner.

10: at the same time, in visual perception a lot of information may be missing (e.g., due to occlusions) and must be guessed by the brain, see [15] and [16, Chapter 10].

Representations may therefore also differ in the information they *discard*, or in their effectiveness at *compressing* information. This may be described as the problem of *what* information present in the raw data (e.g., pixel level information for images) is captured or retained in the representation, as opposed to the problem of *how* it is represented.[11]

11: Lossless compression arguably still falls into the *how* category.

***What* and *how* in deep learning.** To connect to our discussion of deep learning in § 1.3.2, we note that the *what* and *how* categories also play a role in some proposed theories of generalisation for deep learning. In the context of supervised learning, a view is that the success of deep learning may be studied through the process of discarding information which is irrelevant to the downstream task of interest [22], [23, 24]. The problem of optimal information compression in learning machines has also been studied based on intrinsic properties of the data [25–28]. Overall, according to these works, *"learning entails extracting a compressed representation [...] of the structure of statistical dependencies of the data"* [29]. A different line of work focuses instead on *how* information is encoded:

[22]: Tishby et al. (2015), 'Deep learning and the information bottleneck principle'

[29]: Xie et al. (2022), 'A random energy approach to deep learning'

[30]: Saxe et al. (2019), 'On the information bottleneck theory of deep learning'

for example, [30] argued that networks that do not compress may still be capable of generalization; and [31] showed that invertible (up to the penultimate layer) neural networks can be trained to solve large-scale supervised problems.[12]

While both *what* and *how* are interesting aspects of representation learning,[13] this thesis will mainly focus on the former, although we will also discuss the latter in one of the original contributions presented in Chapter 8.

In Chapter 2, we will introduce a way to formalise and study the cocktail-party problem: independent component analysis. An important aspect we will focus on, and of crucial importance in the context of the problems § 1.1 and § 1.2, is the notion of *identifiability*, which we will introduce formally in § 2.3.

12: While not the main focus of this thesis, the formalism introduced in Chapter 2 has also been used to study the *how* question in supervised deep learning [5].

13: Note that while throughout this thesis we mainly consider the problem of dimensionality reduction as separate from blind source separation, dimensionality reduction may be a necessary component of certain identifiability results, see, e.g., [32].

Independent Component Analysis: Identification and Estimation

2

Independent component analysis (ICA) provides a way to formalise and (under suitable assumptions) solve the blind source separation problem § 1.1. A key assumption in ICA is that the unobserved sources should be statistically independent, thus justifying the method's name. We will start by introducing the generative model postulated in ICA.

2.1 Independent Component Analysis (ICA)

We assume the following data-generating process

$$\mathbf{x} = \mathbf{f}(\mathbf{s}), \qquad \mathbf{s} \overset{\text{i.i.d.}}{\sim} p_{\mathbf{s}}, \qquad p_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^{n} p_{s_i}(s_i), \tag{2.1}$$

where the *observed mixtures* $\mathbf{x} \in \mathbb{R}^n$ result from applying a *smooth and invertible mixing function* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ to a set of *unobserved, independent signals or sources* $\mathbf{s} \in \mathbb{R}^n$ with smooth density $p_{\mathbf{s}}$. Due to joint unconditional independence of the latent components, $p_{\mathbf{s}}$ can be expressed as a product of the univariate densitites p_{s_i}.[1]

The goal of ICA is to learn an *unmixing function* $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathbf{y} = \mathbf{g}(\mathbf{x})$ has independent components. *Blind source separation* (BSS), on the other hand, aims to recover the true unmixing $\mathbf{f}^{-1}$ and thus the true sources $\mathbf{s}$ (up to *tolerable ambiguities*, as we will see below).

Whether performing ICA corresponds to solving BSS is related to the concept of *identifiability* of the model class $(\mathbf{f}, p_{\mathbf{s}})$. Intuitively, identifiability is the desirable property that *all models which give rise to the same mixture distribution should be "equivalent" up to certain ambiguities*.

1: The terms "latent" and "unobserved" are sometimes use to refer to variables of which only *corrupted* observations are available (e.g., due to noise in the generative process). Here, we also use them to refer to variables when we only observe mixtures thereof, even when there is no loss of statistical information between sources and mixtures.

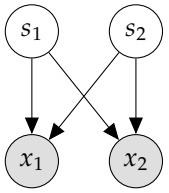


Figure 2.1: ICA setting with $n = 2$ sources (shaded nodes are observed, white ones are unobserved). Here arrows indicate deterministic relations.

2.2 Linear ICA

We start by considering a well studied subcase of the model in (2.1), where the mixing function is linear—i.e., linear ICA.

Our brief account of linear ICA is not meant as a comprehensive treatment of the topic: we refer the reader to [1, 33, 34] for comprehensive introductions. Rather, our main aim is to present ICA as a *case study in identifiability for latent variable models*. In the following, we:

(i) introduce the model and discuss what ambiguities should a priori be considered as tolerable, § 2.2.1;
(ii) characterise its identifiability, including corner cases where it is not achievable, § 2.2.2.

Finally, we briefly discuss estimation in § 2.2.3. Linear ICA thus provides a way to illustrate the main features of our approach to the study of latent variable models, in particular with respect to identifiability. Later the main focus will be on nonlinear ICA, and we introduce a more formal language to articulate identifiability results in § 2.3.

2.2.1 The Model

Linear ICA corresponds to the setting in which a linear mixing is applied to the independent sources, i.e.,

$$\mathbf{x} = \mathbf{A}\mathbf{s}\,, \qquad \mathbf{s} \overset{\text{i.i.d.}}{\sim} p_{\mathbf{s}}, \qquad p_{\mathbf{s}}(\mathbf{s}) = \textstyle\prod_{i=1}^{n} p_{s_i}(s_i)\,, \tag{2.2}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is an invertible mixing matrix.

Tolerable ambiguities. Recall that in ICA the $\mathbf{x}$ variables are observed, whereas both the matrix $\mathbf{A}$ and the vector $\mathbf{s}$ are not (see also Fig. 2.1). In the specification of these unobserved quantities, two ambiguities will necessarily hold:

(i) **The variance of the latent components cannot be determined.** In fact, since both $\mathbf{s}$ and $\mathbf{A}$ are unknown, the effect of multiplying a scalar times the i-th source s_i can always be cancelled by multiplying the inverse scalar times the i-th column of the $\mathbf{A}$ matrix,

$$\mathbf{x} = \sum_i \left(\frac{1}{\alpha_i}\mathbf{a}_i\right)(s_i\alpha_i)\,,$$

where α_i is a scalar and $\mathbf{a}_i$ denotes the i-th column of $\mathbf{A}$. It is therefore customary to fix the source variances, e.g., to $\mathbb{E}[\mathbf{s}\mathbf{s}^\top] = \mathbf{I}$. This still leaves the sign indeterminacy; for most applications, this is inconsequential.

(ii) **The ordering of the sources cannot be determined.** It is easy to see that we can permute the ordering of the sources and simultaneously permute the columns of the matrix $\mathbf{A}$ without affecting the observations. Given a permutation matrix $\mathbf{P}$, we have $\mathbf{x} = \mathbf{A}\mathbf{s} = \mathbf{A}\mathbf{P}^{-1}\mathbf{P}\mathbf{s}$, with $\mathbf{A}\mathbf{P}^{-1}$ (resp. $\mathbf{P}\mathbf{s}$) a new unmixing matrix (resp. a new set of sources).

In short, scale (and sign) and ordering of the sources are two unresolvable ambiguities in the context of linear ICA. Note that as long as our objective is to separate the sources, ordering and scale are both inconsequential: we will therefore also refer to them as "tolerable" ambiguities.

Additional assumptions w.l.o.g. When studying model (2.2) it is customary to make some additional simplifying assumptions. Firstly, the source variables $\mathbf{s}$ can be assumed to have zero mean without affecting estimation of the mixing matrix.[2] Additionally, it can be shown that without loss of generality one can assume that the mixing matrix is orthogonal, $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$. This is because we can always *whiten* $\mathbf{x}$ first through an invertible linear transformation and obtain an orthogonal mixing. We explain this in more detail in Appendix A.1.

As shown in Fig. 2.2, this reduces the problem of linear ICA to the problem of resolving a rotation.[3] The key result of linear ICA identifiability is that statistical independence of the sources, together with minimal assumptions on their distributions, is sufficient to resolve this rotation.

2: If this is not true, one can always apply some simple preprocessing to make it hold, see [1, Sec. 7.2.4].

3: Here we are using the terms "rotation" and "orthogonal matrix" interchangeably. Note also that this reduces the number of parameters required to specify the mixing: in fact, orthogonal $n\times n$ matrices only have $n(n-1)/2$ degrees of freedom.

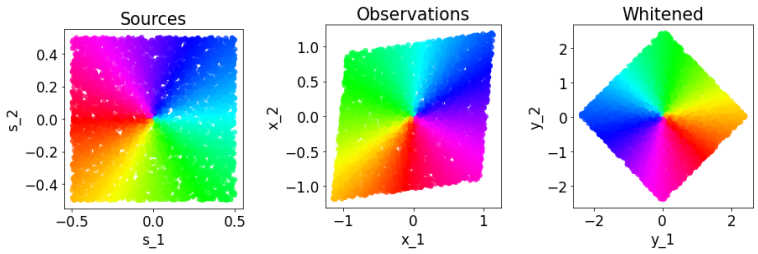


Figure 2.2: Whitening does not separate the independent components: **Left**: Uniform, independent sources **s**; **Center**: Linear mixtures, i.e., observations **x**; **Right**: Whitened (decorrelated) mixtures. We still have to resolve an orthogonal matrix, i.e., a rotation.

2.2.2 Identification

We will assume w.l.o.g. that $\mathbf{A}$ is an orthogonal matrix. Now suppose that we are given an orthogonal matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$ such that the vector

$$\mathbf{y} = \mathbf{Bx} = \mathbf{BAs} \tag{2.3}$$

has independent components. Then $\mathbf{C} = \mathbf{BA}$ is also orthogonal and the following type of identifiability holds [35–37].

Theorem 2.2.1 (Identifiability of linear ICA; based on Thm. 11 of [37]) *Let* $\mathbf{s}$ *be a vector of* n *independent components, of which at most one is Gaussian and whose densities are not reduced to a point mass. Let* $\mathbf{C} \in \mathbb{R}^{n\times n}$ *be an orthogonal matrix. Then* $\mathbf{y} = \mathbf{Cs}$ *has (mutually) independent components iff.* $\mathbf{C} = \mathbf{DP}$*, with* $\mathbf{D}$ *a diagonal matrix and* $\mathbf{P}$ *a permutation matrix.*

In short, Thm. 2.2.1 shows that the two ambiguities deemed unresolvable in § 2.2.1 (i.e., scale and ordering of the sources) are, in fact, the only ambiguities, as long as at most one of the s_i is Gaussian. That is, linear ICA is identifiable up to rescaling and permutation of the sources.

The consequence of identifiability of the linear ICA model is that *ICA can solve blind source separation*: if the model in Equation 2.2 holds, linearly transforming the observations $\mathbf{x}$ into independent components is equivalent to separating the sources.

Note that the terms ICA and blind source separation are sometimes used as synonyms and interchangeably in the literature. Here, we chose to present blind source separation as the overarching goal of separating some latent sources when only given mixtures thereof, and ICA as a specific formalisation of this problem in which transforming observations into independent components is equivalent to separating the true sources (also assumed to be statistically independent) as a consequence of identifiability.[4]

4: See [38] and [1, Section 1.4] for some historical notes on ICA and blind source separation.

Gaussianity and independence. One might wonder why nongaussianity is sufficient to resolve the rotation ambiguity left by whitening. To provide some intuition, we will briefly mention two ways to think about the importance of nongaussianity, see also [39].

[39]: Pavan et al. (2018), 'On the Darmois-Skitovich Theorem and Spatial Independence in Blind Source Separation'

(i) The relationship between Gaussianity and independence has been studied by several authors. J. C. Maxwell investigated this in the context of studying molecule velocity distributions in three-dimensional space [40], and proved that the multivariate normal distribution with identity covariance matrix and zero mean *is the only spherically symmetric*

distribution with independent components. Therefore, all other distributions where the components are independent will not be spherically symmetryc (or rotationally invariant). This gives some hope that rotation can be resolved in ICA.[5]

5: a rigorous proof that this is possible is however a bit more involved [37, 39].

(ii) Loosely speaking, an implication of the central limit theorem (see, e.g., [1, Section 2.5.2]) is that summing (non-Gaussian) independent random variables makes the resulting variables more Gaussian than the original ones. Therefore, if we assume the model in (2.2) and define a variable y as a linear combination of the observed components, $y = \sum b_i x_i$, it will be maximally nongaussian if it equals one of the original independent components. In fact, many methods for linear ICA work by maximising the nongaussianity of the estimated components. It can be shown that, under certain conditions, this objective is also related to those of projection pursuit and sparse coding, see [1, Sec. 1.3.3].

Beyond nongaussianity. In the context of linear ICA, other deviations from a Gaussian i.i.d. setting can also lead to identifiability: for example, nonstationarity [41] and time correlation [42]. A general information-geometric framework links these three different routes to identifiability [43]. A different line of research employs a tensorial framework to achieve source separation [44].[6]

6: Interestingly, in the context of nonlinear ICA, nonstationarity and time correlation are special cases of a more general framework based on *auxiliary variables*, see also § 2.4.2.

2.2.3 Estimation

Once the identifiability of the ICA model is established, the problem remains how to estimate the sources from data.

There is a multitude of approaches for estimation in linear ICA. To provide an example, we will briefly mention one method: maximising the likelihood under the generative model in Equation 2.2.

The likelihood of the ICA model. As a starting point, we write the likelihood of a single datapoint $\mathbf{x}$ under the model in (2.2). Define $\mathbf{B} = \mathbf{A}^{-1}$; through a change of variable, we find

$$p_{\mathbf{x}}(\mathbf{x}) = |\det \mathbf{B}| p_{\mathbf{s}}(\mathbf{s}) = |\det \mathbf{B}| \prod_i p_{s_i}(s_i) \,. \tag{2.4}$$

It is useful to express the likelihood as a function of the observed variables $\mathbf{x}$ and the unmixing matrix $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_n)^\top$, where $\mathbf{b}_i$ is the i-th row of $\mathbf{B}$, which yields

$$p_{\mathbf{x}}(\mathbf{x}) = |\det \mathbf{B}| \prod_i p_{s_i}\left(\mathbf{b}_i^\top \mathbf{x}\right). \tag{2.5}$$

Maximum likelihood estimation. A natural approach for estimation is then to maximise the likelihood above as a function of the parameters of a tentative unmixing matrix $\mathbf{W}$, assuming some factorised distribution $p(\mathbf{y}) = \prod_i p_{y_i}(y_i)$ for the sources, where the p_{y_i} are nongaussian. Similar to the expression of the true log-likelihood (2.5), we get

$$\log p_{\mathbf{W}}(\mathbf{x}) = \sum_{i=1}^{n} \log p_{y_i}\left(\mathbf{w}_i^\top \mathbf{x}\right) + \log |\det \mathbf{W}| \,. \tag{2.6}$$

Here we indicated the log-likelihood on the LHS with $p_{\mathbf{W}}$ to stress that it is the likelihood of the data *under our model* (specified by the parameters $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)^\top$—as well as by the choice of p_{y_i}, which we will get back to later), and to distinguish it from the true likelihood (2.5).

Maximising this log-likelihood w.r.t. the parameters of an unmixing $\mathbf{W}$ is equivalent to

$$\begin{aligned}\mathbf{W}^* &= \arg\max_{\mathbf{W}} \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}} \left[\log p_{\mathbf{W}}(\mathbf{x})\right] \\ &= \arg\max_{\mathbf{W}} \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}} \left[\log p_{\mathbf{y}}(\mathbf{W}\mathbf{x}) + \log|\det \mathbf{W}|\right] \\ &= \arg\max_{\mathbf{W}} \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}} \left[\sum_i \log p_{y_i}(\mathbf{w}_i^\top \mathbf{x}) + \log|\det \mathbf{W}|\right]. \end{aligned} \tag{2.7}$$

In practice, the expectation in (2.7) is substituted with a finite sample average: given a collection of N samples $\{\mathbf{x}^{(j)}\}_{j=1}^N$, we seek to determine

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \frac{1}{n} \sum_{j=1}^{N} \left[\sum_i \log p_{y_i}(\mathbf{w}_i^\top \mathbf{x}^{(j)}) + \log|\det \mathbf{W}|\right]. \tag{2.8}$$

Choice of the latent distribution. One issue with the approach above is how to choose p_{y_i}. The true likelihood in (2.5) is expressed in terms of the univariate source distributions p_{s_i}. In practice, as the sources are unobserved, their distributions might be unknown a priori. This would complicate the problem, as one should simultaneously (i) estimate the unmixing matrix $\mathbf{W}$ and (ii) estimate the source density p_{s_i} (in principle a nonparametric, hard problem).

However, it turns out that (small) misspecifications of the source densities p_{s_i} are inconsequential for source separation: this can be shown rigorously, see, e.g., [34, Sec. 1.4.2] and [1, Sec. 9.1.2], as well as [45]. We therefore want to stress that in practice, for source separation, it may not be necessary to guess or estimate the true source densities, which simplifies the problem. In practice, many algorithms work by either fixing a nongaussian form for the source densities, or by adopting more sophisticated approaches.

A gradient-based algorithm. How do we solve the problem in (2.8) in practice? A natural way to do so is by gradient ascent. We start by taking the gradient of the log-likelihood in (2.2) with respect to the parameters $\mathbf{W}$ (here we compute the gradient for a single sample, corresponding to online or stochastic gradient descent):

$$\frac{\partial \log p_{\mathbf{W}}(\mathbf{x})}{\partial \mathbf{W}} = \left[\mathbf{W}^\top\right]^{-1} + \mathbf{g}(\mathbf{W}\mathbf{x})\mathbf{x}^\top, \tag{2.9}$$

where $\mathbf{g}(\mathbf{y}) = (g_i(y_i), \ldots, g_n(y_n))$ is a component-wise vector function defined as

$$g_i = \left(\log p_{y_i}\right)' = \frac{p'_{y_i}}{p_{y_i}}.$$

This corresponds to the Bell-Sejnowski algorithm for source separation [46, 47].

Note that computing the right hand side of (2.9) involves a matrix inversion, which has cubic cost in the dimensionality n of the observations. This may be computationally inefficient, and many gradient-based algorithms involve ways to sidestep this expensive computation.We will come back to this in Chapter 5.

2.2.4 Concluding Remarks on Linear ICA

We summarise some take home messages we consider relevant beyond the specific form of the model in Equation 2.2, and which will recur in our study of different and possibly nonlinear latent variable models.

- Certain ambiguities are unavoidable from the outset (in linear ICA, permutation and scale) and may be deemed tolerable given the problem we want to solve (as they would still allow source separation).
- An identifiability result defines under which conditions the only unresolvable ambiguities are those deemed tolerable from the outset. This usually holds *apart from a (hopefully small) number of corner cases* which can be characterised (in linear ICA, when there are more than two Gaussian components).
- Certain assumptions can be made without loss of generality given the assumed model (e.g., orthogonality of the mixing matrix). This might simplify both the theoretical analysis (see § 2.2.2) and the estimation (see note 3 and, e.g., [48]).
- Estimation of an identifiable model comes with its own problems (e.g., statistical, computational efficiency), which can be addressed separately from the problem of identification.[7]

7: Identifiability of a model may however be insufficient to ensure that it can be estimated from data [49, 50]. We will come back to this in Chapter 9.

An additional remark is that we introduced (linear) ICA as the estimation of an (identifiable) generative model. This is motivated by blind source separation and the metaphor we introduced in § 1.1. However, ICA may also be motivated differently. For example, it may sometimes be unrealistic to assume that the model in Equation 2.2 holds exactly. It might however still be useful to find a (linear) map which transforms the obervations into a vector whose components are as independent as possible. This can be motivated as a good sensory coding strategy [1, Chapter 10]. We might then be interested in comparing the representations extracted by two learning systems adopting the same sensory coding strategy, as in the metaphor of § 1.2, regardless of the ground truth generative process.

Finally, while we focused on introducing the model and formalism, we want to mention that applications of linear ICA are ubiquitous, including signal processing [51], text mining [52], financial time series analysis [53] and neuroimaging [54]. Another example is in the context of astronomy [55] and separation of astrophysical emissions [56–58]. The application to neuroimaging of a model based on linear ICA will also be the topic of Chapter 6.

2.3 Identifiability

Now that we presented our case study based on linear ICA, we will define identifiability more formally.

Identifiability in statistics. Traditionally [59–61], identifiability for a class of models p_θ for observed data $\mathbf{x}$ parametrised by $\theta \in \Theta$ is expressed as the condition that there needs to be a one-to-one mapping between the space of models and the space of parameters, i.e., the model class p_θ is said to be identifiable if[8]

$$\forall \theta, \theta' \in \Theta : \qquad p_\theta(\mathbf{x}) = p_{\theta'}(\mathbf{x}) \quad \forall \mathbf{x} \quad \Longrightarrow \quad \theta = \theta'. \tag{2.10}$$

However, the equality on the right hand side of (2.10) is a very strong condition which makes this type of identifiability impractical for many settings. For example, as we mentioned in section 2.2, for linear ICA the ordering of the sources cannot be determined, so identifiability in the sense of (2.10) is infeasible for the parameters of the unmixing matrix.

Identifiability in terms of equivalence relations. The equality in parameter space on the right hand side of the implication in (2.10) is therefore sometimes replaced by an equivalence relation denoted with $\sim$ [62].

Definition 2.3.1 *An equivalence relation $\sim$ on a set A is a binary relation which satisfies the following three properties:*

1. *Reflexivity:* $a \sim a, \forall a \in A$.
2. *Symmetry:* $a \sim b \implies b \sim a, \forall a, b \in A$.
3. *Transitivity:* $(a \sim b) \wedge (b \sim c) \implies a \sim c$.

An equivalence relation on a set A imposes a partition into disjoint subsets. Each such subset corresponds to an equivalence class, i.e., the collection of all elements which are $\sim$-related to each other. For example, $[a] = \{b \in A : a \sim b\}$ denotes the equivalence class containing the element a.[9]

Given a suitable equivalence relation, we can then define the following notion of identifiability: The model class p_θ is $\sim$-identifiable if

$$\forall \theta, \theta' \in \Theta : \qquad p_\theta(\mathbf{x}) = p_{\theta'}(\mathbf{x}) \quad \forall \mathbf{x} \quad \Longrightarrow \quad \theta \sim \theta' . \tag{2.11}$$

Defining an appropriate equivalence class for the problem at hand therefore allows us to specify exactly the type of indeterminancies which cannot be resolved and up to which the true generative process can be recovered.

Separate constraints on the mixing and on the latent variables. Since the generative process of nonlinear ICA (2.1) is determined by the choice of mixing function and source distribution, the space Θ from (2.11), in this case, corresponds to the product space of the space of mixing functions $\mathcal{F}$ and source distributions $\mathcal{P}$. Moreover, the pushforward density $\mathbf{f}_* p_\mathbf{s}$ denotes the density of the transformed variable $\mathbf{x} = \mathbf{f}(\mathbf{s})$, corresponding to the density of the observed mixtures—$p_\theta(\mathbf{x})$ in (2.10). Based on this observation, we now provide an alternative formulation of the identifiability defined in (2.11).[10]

[59]: Wasserman (2004), *All of statistics: a concise course in statistical inference*
[60]: Lehmann et al. (2006), *Theory of point estimation*
[61]: Casella et al. (2021), *Statistical inference*

8: While this definition is based on the likelihood of two different models, a similar definition could in principle be given substituting p_θ with a function different from the likelihood.

9: A trivial example of an equivalence relation is equality (=), which would lead back to (2.10). In the context of linear ICA, one can consider equivalence up to permutation and scale of the columns of the mixing matrix.

10: Defn. 2.3.2 was originally introduced in [63], on which the contribution in Chapter 3 is based.

Definition 2.3.2 (∼-identifiability) *Let $\mathcal{F}$ be the set of all smooth, invertible functions $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, and $\mathcal{P}$ be the set of all smooth, factorised densities $p_\mathbf{s}$ with connected support on $\mathbb{R}^n$. Let $\mathcal{M} \subseteq \mathcal{F} \times \mathcal{P}$ be a subspace of models and let $\sim$ be an equivalence relation on $\mathcal{M}$. Denote by $\mathbf{f}_* p_\mathbf{s}$ the push-forward density of $p_\mathbf{s}$ via $\mathbf{f}$. Then the generative process* (2.1) *is said to be* ∼-identifiable *on $\mathcal{M}$ if*

$$\forall (\mathbf{f}, p_\mathbf{s}), (\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}}) \in \mathcal{M} : \qquad \mathbf{f}_* p_\mathbf{s} = \tilde{\mathbf{f}}_* p_{\tilde{\mathbf{s}}} \quad \Longrightarrow \quad (\mathbf{f}, p_\mathbf{s}) \sim (\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}}) \,. \tag{2.12}$$

Here we used a similar notation to [62, Sec. 2.2]: we use $\mathbf{f} \in \mathcal{F}$ to refer to functions, not parameters. For example, in the case of a neural network, $\mathbf{f}$ refers to the input-output mapping implemented by the network, and not to the parameters (e.g., weights and biases) in which one usually performs gradient ascent.

[62]: Khemakhem et al. (2020), 'Variational Autoencoders and Nonlinear ICA: A Unifying Framework'

The equality $\mathbf{f}_* p_\mathbf{s} = \tilde{\mathbf{f}}_* p_{\tilde{\mathbf{s}}}$ requires that the two models give rise to the same observational distribution: for any modification $\mathbf{f} \to \tilde{\mathbf{f}}$, the distribution $p_\mathbf{s}$ needs to be modified accordingly $p_\mathbf{s} \to p_{\tilde{\mathbf{s}}}$ such that the observed distributions $\mathbf{f}_* p_\mathbf{s}$ and $\tilde{\mathbf{f}}_* p_{\tilde{\mathbf{s}}}$ are the same.

Identifiability of linear ICA. As a matter of example, we will restate the result on identifiability of linear ICA in terms of our notation.

Firstly, in (2.2), the mixing is an invertible matrix, so we take $\mathcal{F}_{\text{LIN}}$ to be the space of invertible $n \times n$ matrices.[11] Moreover, since identifiability holds if at most one of the latent components is Gaussian, we take and $\mathcal{P}_{\text{LIN}}$ as the space of source distributions $p_\mathbf{s} = \prod_i p_{s_i}$ with at most one Gaussian p_{s_i}. The subspace of models we consider in linear ICA is therefore $\mathcal{M}_{\text{LIN}} = \mathcal{F}_{\text{LIN}} \times \mathcal{P}_{\text{LIN}}$.

11: Together with the operation of ordinary matrix multiplication, $\mathcal{F}_{\text{LIN}}$ defines the *general linear group* or order n.

We discussed how linear ICA is identifiable up to permutation and rescaling of the sources. In terms of an equivalence relation, linear ICA is $\sim_{\text{LIN}}$-identifiable on $\mathcal{M}_{\text{LIN}}$, where $\sim_{\text{LIN}}$ is defined as

Definition 2.3.3 ($\sim_{\text{LIN}}$) *The equivalence relation $\sim_{LIN}$ on $\mathcal{F}_{LIN} \times \mathcal{P}_{LIN}$ defined as in Defn. 2.3.2 is given by*

$$(\mathbf{A}, p_\mathbf{s}) \sim_{LIN} (\tilde{\mathbf{A}}, p_{\tilde{\mathbf{s}}}) \iff \exists \mathbf{D}, \mathbf{P} \quad s.t. \quad (\mathbf{A}, p_\mathbf{s}) = (\tilde{\mathbf{A}}\mathbf{D}\mathbf{P}, [\mathbf{P}^{-1}\mathbf{D}^{-1}]_* p_{\tilde{\mathbf{s}}}) \,, \tag{2.13}$$

with $\mathbf{D}$ a diagonal matrix and $\mathbf{P}$ a permutation matrix.

As we discussed in § 2.2.2, linear ICA is identifiable up to $\sim_{\text{LIN}}$ on the subspace $\mathcal{M}_{\text{LIN}}$ of pairs of invertible matrices (constraint on $\mathcal{F}$) and factorizing densities for which at most one s_i is Gaussian (constraint on $\mathcal{P}$).

Motivating our notation. We deliberately choose to define identifiability and to express the observed distribution in terms of the source distribution and the mixing function—as opposed to in terms of the observed distribution and the unmixing function as in some prior work [64–66]—because this is aligned with the causal direction of data generation, something we will exploit in Chapter 3 and come back to in Chapter 9.

We also believe that, in this framework, separate constraints on the space of mixing functions $\mathcal{F}$ and source distributions $\mathcal{P}$ are expressed more naturally. As we showed, the former may, for example, suitably encode that the considered models in (2.2) are linear, whereas the latter may be used to specify the class of latent distribution (e.g., independent components) and articulate exceptions (e.g., more than two Gaussian components). The equivalence class "$\sim$" specifies then "up to" what class of unresolvable ambiguities we can achieve identification.

Where did the ground truth end up? The formulation in Defn. 2.3.2 does not mention a *true* generative process: by comparing two models $(\mathbf{f}, p_{\mathbf{s}}), (\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}})$, without explicit reference to a ground truth, the definition may sound closer in spirit to the independent-listeners problem § 1.2 than to the one in § 1.1. One way to view the connection to § 1.1 is as follows: suppose that the model in Equation 2.2 corresponds to the ground truth, denoted by $(\mathbf{A}, p_{\mathbf{s}})$, and we consider a learnt linear model $(\tilde{\mathbf{A}}, p_{\tilde{\mathbf{s}}}) \in \mathscr{M}_{\text{LIN}}$ such that the true distribution and the model distribution for the observed mixtures, respectively $p_{\mathbf{x}} = \mathbf{A}_* p_{\mathbf{s}}$ and $p_{\tilde{\mathbf{x}}} = \tilde{\mathbf{A}}_* p_{\tilde{\mathbf{s}}}$, are equal.[12] Identifiability of linear ICA guarantees that, under the assumptions in Thm. 2.2.1, the learnt model recovers the ground truth sources up to the ambiguities in Defn. 2.3.3—i.e., for § 1.1, it solves blind source separation.[13] In practice, we may be unable to know whether the model distribution matches the ground truth distribution, since the latter may be unknown to us. Nevertheless, if we believe that the linear ICA model (2.2) holds, and under the assumptions of Thm. 2.2.1, it is sufficient to know that the learnt model belongs to $\mathscr{M}_{\text{LIN}}$—i.e., that $\tilde{\mathbf{A}}$ is an invertible mixing matrix and that $p_{\tilde{\mathbf{s}}}$ is a distribution with independent components. The model is then guaranteed to separate the sources.

12: Note that the model that maximises the likelihood is also the one which minimises the Kullback-Leibler divergence between the model distribution and the true data distribution, see, e.g., [67, 68].

13: We however remark that, as we mentioned in § 2.2.3, source separation can be achieved even if the likelihood under the learnt model does not match the true likelihood, e.g., if the source density is (slightly) misspecified. Note also that all above considerations are in the population limit

2.4 Nonlinear ICA

Whereas we presented linear ICA as a case study in identifiability, nonlinear ICA can be considered as a case study in *nonidentifiability*.

However, as we will discuss, the implications of nonidentifiability of nonlinear ICA are broader. In fact, the proof technique used to show that nonlinear ICA is not identifiable can also be used to show nonidentifiability of any nonlinear latent variable model in the i.i.d. setting. The results therefore have very broad implications as an impossibility result for representation learning (specifically for blind source separation) without strong assumptions and constraints.

We will start by analysing the nonlinear ICA model and show that it is not identifiable in § 2.4.1. We will then move on (§ 2.4.2) to defining a set of additional assumptions under which it can become identifiable.

2.4.1 Nonidentifiability

Model and tolerable ambiguities. The nonlinear ICA model is given by the equation (2.1) (i.e., no constraints on $\mathcal{F}$ beyond smoothness and invertibility).

Similar to what we did for linear ICA, it is natural to define what ambiguities should be deemed unresolvable a priori, and whether they are tolerable for blind source separation. An important observation is that if s_i and s_j are independent random variables, then so are $h_i(s_i)$ and $h_j(s_j)$ for any functions h_i and h_j. Therefore, in addition to the permutation ambiguity we encountered in § 2.2.1, such *element-wise nonlinear transformations* $\mathbf{h}(\mathbf{s}) = (h_1(s_1), \dots, h_n(s_n))$ cannot be resolved either.[14]

14: these can be considered as generalisations of the scale ambiguity in linear ICA.

For nonlinear ICA, the desired notion of identifiability—in the sense of the strongest feasible type of identifiability that is possible without further assumptions—is captured by $\sim_{\text{BSS}}$ defined as follows.

Definition 2.4.1 ($\sim_{\text{BSS}}$) *The equivalence relation $\sim_{BSS}$ on $\mathcal{F} \times \mathcal{P}$ defined as in Defn. 2.3.2 is given by*

$$(\mathbf{f}, p_{\mathbf{s}}) \sim_{BSS} (\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}}) \iff \exists \mathbf{P}, \mathbf{h} \quad s.t. \quad (\mathbf{f}, p_{\mathbf{s}}) = (\tilde{\mathbf{f}} \circ \mathbf{h}^{-1} \circ \mathbf{P}^{-1}, (\mathbf{P} \circ \mathbf{h})_* p_{\tilde{\mathbf{s}}}) \tag{2.14}$$

where $\mathbf{P}$ is a permutation and $\mathbf{h}(\mathbf{s}) = (h_1(s_1), \dots, h_n(s_n))$ is an invertible, element-wise function.

While the ambiguity defined in Defn. 2.4.1 is larger than that in linear ICA Defn. 2.3.3, it is still *tolerable* in the context of blind source separation: if we were able to achieve it, we would still *separate* the sources in the sense that, while they may be nonlinearly distorted with respect to the true ones, the would not be mixed in our representation.

However, as we will now show, a fundamental obstacle—and a crucial difference to the linear problem—is that in the nonlinear case, different *mixtures* of s_i and s_j can be independent even in the nongaussian case: i.e., solving nonlinear ICA is *not* equivalent to solving blind source separation. Therefore we cannot even guarantee reconstruction of the sources up to these ambiguities. This can be shown through suitably constructed counterexamples or "spurious solutions".

The Darmois construction. A prominent example of this is given by the *Darmois construction* [69, 70].

Definition 2.4.2 (Darmois construction) *The* Darmois construction $\mathbf{g}^{\text{D}} : \mathbb{R}^n \to (0,1)^n$ *is obtained by recursively applying the conditional cumulative distribution function (CDF) transform:*

$$g_i^{\text{D}}(\mathbf{x}_{1:i}) := \mathbb{P}(X_i \leq x_i | \mathbf{x}_{1:i-1}) = \int_{-\infty}^{x_i} p(x_i' | \mathbf{x}_{1:i-1}) dx_i' \qquad (i = 1, \dots, n). \tag{2.15}$$

The resulting *estimated* sources $\mathbf{y}^{\text{D}} = \mathbf{g}^{\text{D}}(\mathbf{x})$ are mutually-independent uniform r.v.s by construction: by applying a change of variables, we can see that the transformed variables in (2.15) are uniformly distributed in the open unit cube, thereby corresponding to independent components [71, § 2.2].[15] However, they need not be meaningfully related to the *true* sources $\mathbf{s}$, and will, in general, still be a nonlinear mixing thereof [70]. Consider, e.g., a mixing $\mathbf{f}$ with full Jacobian which yields a contradiction to Defn. 2.4.1, due to Remark 2.4.1.

15: Note that by applying the construction in Equation 2.15 to the first component of the observed vector, x_1, one can transform its distribution into a uniform distribution through the (unconditional) CDF transform.

Remark 2.4.1 $\mathbf{g}^{\mathrm{D}}$ has lower-triangular Jacobian, i.e., $\partial g_i^{\mathrm{D}}/\partial x_j = 0$ for $i < j$. Since the order of the x_i is arbitrary, applying $\mathbf{g}^{\mathrm{D}}$ after a permutation yields a different Darmois solution. Moreover, (2.15) yields independent components $\mathbf{y}^{\mathrm{D}}$ even if the sources s_i were not independent to begin with.

Denoting the mixing function corresponding to (2.15) by $\mathbf{f}^{\mathrm{D}} = (\mathbf{g}^{\mathrm{D}})^{-1}$ and the uniform density on $(0,1)^n$ by $p_{\mathbf{u}}$, the *Darmois solution* $(\mathbf{f}^{\mathrm{D}}, p_{\mathbf{u}})$ thus allows construction of counterexamples to $\sim_{\text{BSS}}$-identifiability on $\mathcal{F} \times \mathcal{P}$.

Measure-preserving automorphisms. Another well-known obstacle to identifiability are *measure-preserving automorphisms* (MPAs) of the source distribution $p_{\mathbf{s}}$: these are functions $\mathbf{a}$ which map the source space to itself without affecting its distribution, i.e., $\mathbf{a}_* p_{\mathbf{s}} = p_{\mathbf{s}}$ [70].

A particularly instructive class of MPAs is the following [62, 72].

Definition 2.4.3 ("Rotated-Gaussian" MPA) *Let* $\mathbf{R} \in O(n)$ *be an orthogonal matrix, and denote by* $\mathbf{F}_{\mathbf{s}}(\mathbf{s}) = (F_{s_1}(s_1), \ldots, F_{s_n}(s_n))$ *and* $\mathbf{\Phi}(\mathbf{z}) = (\Phi(z_1), \ldots, \Phi(z_n))$ *the element-wise CDFs of a smooth, factorised density* $p_{\mathbf{s}}$ *and of a Gaussian, respectively. Then the "rotated-Gaussian" MPA* $\mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}})$ *is*

$$\mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}}) = \mathbf{F}_{\mathbf{s}}^{-1} \circ \mathbf{\Phi} \circ \mathbf{R} \circ \mathbf{\Phi}^{-1} \circ \mathbf{F}_{\mathbf{s}} \,. \tag{2.16}$$

$\mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}})$ first maps to the (rotationally invariant) standard isotropic Gaussian (via $\mathbf{\Phi}^{-1} \circ \mathbf{F}_{\mathbf{s}}$), then applies a rotation, and finally maps back, without affecting the distribution of the estimated sources. Hence, if $(\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}})$ is a valid solution, then so is $(\tilde{\mathbf{f}} \circ \mathbf{a}^{\mathbf{R}}(p_{\tilde{\mathbf{s}}}), p_{\tilde{\mathbf{s}}})$ for any $\mathbf{R} \in O(n)$. Unless $\mathbf{R}$ is a permutation, this constitutes another common counterexample to $\sim_{\text{BSS}}$-identifiability on $\mathcal{F} \times \mathcal{P}$.

Other spurious solutions. Despite their prominent role in the literature, the two classes of spurious solutions we just presented do not necessarily exhaust *all* possible spurious solutions: we are not aware of an analytical characterisation of all possible solutions yielding independent components.[16] An implication is that ruling out these constructions by suitable restrictions on the model class or latent distribution does *not* correspond to proving identifiability of a model.

16: However see [73], eq. (7), for the differential functional equation solutions to nonlinear ICA need to satisfy. See also [74] for a recent example where a spurious solution was used to illustrate nonidentifiability issues in the context of causal representation learning.

Spurious solutions vs. identifiability "except when". We remark that there is a fundamental difference between the counterexamples we presented and the "exception" constituted by Gaussian components in the identifiability of linear ICA. In fact, both counterexamples in Defn. 2.4.2 and Defn. 2.4.3 can be constructed for any choice of $p_{\mathbf{s}} \in \mathcal{P}$ and $\mathbf{f} \in \mathcal{F}$: unlike the *exception* of Gaussian in linear ICA, they cannot be simply ruled out by choosing a restricted class of models $\mathcal{M} \subset \mathcal{F} \times \mathcal{P}$.

A first implication of the counterexamples presented above is that nonlinear ICA is not identifiable: *solving nolinear ICA does not solve blind source separation*.

Implications for unsupervised machine learning. Note that Defn. 2.4.2 does not require that the observations are generated by the model in Equation 2.1: the Darmois construction can be applied even if the observed distribution is not generated by mixing independent components to begin with.[17]

Therefore, an even more radical implication of the Darmois construction is that it can be used to show that unsupervised representation learning is impossible (in the setting with i.i.d. observations) without additional constraints: in fact, this is true *even if we were to postulate a different model wherethe latent distribution has non-independent components*.[18]

To show this for a general prior on $\mathbf{z}$ (i.e., one with non-independent components), it is enough to point out that we can transform any variable into independent (Darmois) Gaussian (element-wise inverse Gaussian CDF transform) variables; apply a rotation $\mathbf{O}$ to the resulting Gaussian variables; transform them back to uniform random variables (element-wise Gaussian CDF transform) and finally invert the Darmois construction. By doing so, we get a nonlinear transformation $\mathbf{z}' = \mathbf{g}_{\text{CDF}}^{-1}(\mathbf{O} \circ \mathbf{g}_{\text{CDF}}(\mathbf{z}))$, where $\mathbf{z}'$ has exactly the same distribution as $\mathbf{z}$ but is a complex nonlinear transformation (and mixture) thereof.[19]

To summarise, *by looking at the (i.i.d.) data alone and without further assumptions, it is not possible to recover the true latents, no matter what the prior may be*. Or in other terms: representations extracted by two different models fitting the data equally well may be arbitrarily entangled with respect to one another.[20]

Is all hope for identifiability in nonlinear representation learning lost? One is then left wondering whether there is any hope for identifiability in (nonlinear) representation learning.

The case of linear ICA, together with our formulation in Defn. 2.3.2, suggests that a path to recover identifiability could be to constrain the class of the mixing functions $\mathcal{F}$. As we saw, if the class is constrained to only include linear invertible functions, we recover the useful notion of identifiability up to Defn. 2.3.3 under mild conditions. It therefore seems reasonable to believe that it should be possible to define a broader class of (nonlinear) functions such that some interesting notion of identifiability might still be achievable. We will review some of the work on this in Chapter 3, where we additionally present one of the original contributions of this thesis, where we defined a function class based on constraints inspired by principles of causal inference.

2.4.2 Nonlinear ICA with Auxiliary Variables

In addition to adding assumptions on $\mathcal{F}$, another possible avenue, which we explore below, would be to add assumptions to the source distribution $p_{\mathbf{s}}$. In this spirit, a modification of the basic generative model (2.1) which has received significant attentionis to consider settings where an *auxiliary variable* $\mathbf{u}$ renders the sources *conditionally* independent [64–66]:[21] the assumption on $p_{\mathbf{s}}$ in (2.1) is therefore replaced with

$$\mathbf{s} \overset{\text{i.i.d.}}{\sim} p_{\mathbf{s}|\mathbf{u}}, \qquad p_{\mathbf{s}|\mathbf{u}}(\mathbf{s}|\mathbf{u}) = \prod_{i=1}^{n} p_{s_i|\mathbf{u}}(s_i|\mathbf{u}) \,. \tag{2.17}$$

17: The converse implication is that any smooth distribution which is fully supported on $\mathbb{R}^n$ can be written as the mixture of n independent components through the Darmois construction.

18: The observation that the Darmois construction implies that not only factorised priors, but *any* unconditional prior is insufficient for identifiability can be found, e.g., in [62, App. D.2].

19: where $\mathbf{g}_{\text{CDF}} = \mathbf{g}^D \circ \mathbf{h}_{\text{CDF}}$, and $\mathbf{h}_{\text{CDF}}$ is the element-wise unconditional CDF of a zero mean standard normal distribution.

20: The two formulations here are intended to mirror the cocktail-party problem in § 1.1 and the independent-listeners problem in § 1.2, and imply an impossibility for both.

21: as we will see, in some cases $\mathbf{u} \in \mathbb{R}^d$, where d does not need not be equal to the data dimensionality n; in some other cases it is a categorical variable.

Under some technical assumptions, identifiability up to suitable equivalence classes, and in some cases even up to Defn. 2.4.1, can then be achieved *without further restrictions on the nonlinear mixing* $\mathbf{f}$.

It may sound surprising that such a strategy could succeed at all: after all, we remarked in Subsection 2.4.1 that the Darmois construction provides a way to show nonidentifiability *even beyond the setting where* $p_{\mathbf{s}} \in \mathcal{P}$ *has independent components*. So it might appear as if additional conditions on $\mathcal{P}$ in Defn. 2.3.2 would not help us recover interesting notions of identifiability.

However, the Darmois construction rests on the assumption that the datapoints are *independent and identically distributed* samples from a distribution $p_{\mathbf{s}}$, i.e.,

$$\mathbf{s} \overset{\text{i.i.d.}}{\sim} p_{\mathbf{s}} \,.$$

It turns out that the auxiliary variable setting can in fact be interpreted as a deviation from the i.i.d. assumption. A special case where this deviation has a clear interpretation is the one where the observations are nonlinear mixtures of latent variables whose distribution is *nonstationary* [64]: observations can then be grouped into time segments indexed by time-segment labels $\mathbf{u}$. Within a time segment, the latent vector $\mathbf{s}$ is sampled i.i.d.; but the distribution changes across time segments, giving rise to nonstationarity. A different criterion relies on a different kind of temporal structure, autocorrelation [65]: the sources are sampled from a stochastic process where $p_{\mathbf{s}(t)|\mathbf{s}(t-1)}$ has independent components.[22]

In both cases, the variable $\mathbf{u}$ indicates a deviation from the i.i.d. setting: any two $\mathbf{s}$ variables may be nonidentically distributed (in the nonstationary case their distribution changes depending on the time-segment label $\mathbf{u}$), or not independent (due to autocorrelation). Deviations from the i.i.d. assumption therefore help solving the problem of identifiability.[23]

A possible interpretation of the conditional independence statement in (2.17) is that it posits that $\mathbf{u}$ is a parent of the sources $\mathbf{s}$, see also Fig. 2.3.

Other settings with auxiliary variables. The results based on autocorrelation and nonstationarity are special cases of a more general framework [66]. Depending on the setting, $\mathbf{u}$ can then be interpreted as an environment index, a time segment, class label or indicate other kinds of auxiliary information and complex structures such as spatial dependencies [5, 62, 64–66, 78, 79]. In many cases $\mathbf{u}$ is assumed to be observed separately from $\mathbf{x}$, though [78] is an exception, and already in [65] the authors exploit dynamics of time series instead of a separate $\mathbf{u}$ variable.

Contrastive learning. In the setting described above, a constructive proof of identifiability can be attained by exploiting contrastive learning [66, 80]. This technique transforms a density ratio estimation problem into one of supervised function approximation. This idea, which we recapitulate in Appendix E.1, has a long history [81], and provides the foundation for many approaches to learning generative models [80, 82].

For nonlinear ICA with auxiliary variables, contrastive learning can be exploited by training a classifier to distinguish between a tuple sampled

22: In this case, $\mathbf{u} = \mathbf{s}(t-1)$—or rather, conditioning on obsereved variables, $\mathbf{u} = \mathbf{x}(t-1)$; note that conditioning on $\mathbf{x}(t-1)$ or $\mathbf{s}(t-1)$ is the same since the two are related by deterministic transformations. Note that autocorrelation and nonstationarity, which are classically considered as separate criteria in linear ICA, are in fact both special cases of the general auxiliary variables setting [66]. The idea of exploiting temporal structure for nonlinear blind source separation had already been discussed in [75, 76].

23: It has been argued that the i.i.d. assumption, which is ubiquitous in machine learning, is problematic, and that moving beyond such assumption will be required to solve many of the important problems in modern machine learning [77].

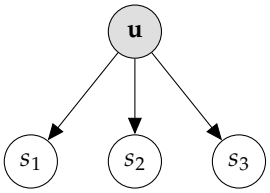


Figure 2.3: The auxiliary variable $\mathbf{u}$ can be seen as a parent of the s_i variables in a directed acyclic graph. The conditional independence in Equation 2.17 would also be consistent with the case where the arrows between $\mathbf{u}$ and $\mathbf{s}$ are undirected; however they cannot be in the inverse direction—otherwise conditioning on $\mathbf{u}$ (a collider) would create dependence among the s_i. Note that in this figure arrows indicate *probabilistic* relations, unlike the deterministic ones in Fig. 3.3b. Observations $\mathbf{x} = \mathbf{f}(\mathbf{s})$ are given by a deterministic mixing as in Fig. 3.3b.

from the joint distribution, which we denote as $(\mathbf{x}, \mathbf{u})$, and one where $\mathbf{u}^*$ is a sample generated from the marginal $p(\mathbf{u})$ *independently of* $\mathbf{x}$, $(\mathbf{x}, \mathbf{u}^*)$. For example, in the nonstationary setting, a tuple $(\mathbf{x}, \mathbf{u}^*)$ may be given by an observation at time t, $\mathbf{x}(t)$, together with a time segment index $\mathbf{u}(t')$, $t \neq t'$, which may be different from the true one $\mathbf{u}(t)$, and taken at random from the collection of time indices. The marginals of both distributions are equal, and the classifier must learn to distinguish between *positive samples* $(\mathbf{x}, \mathbf{u})$ and *negative samples* $(\mathbf{x}, \mathbf{u}^*)$. It turns out that, under suitable assumptions on the regression function employed by the classifier, this problem is solved optimally when the final layer features used for classification are the latent sources up to some tolerable ambiguities. We will see an example of this in Chapter 6, where we will present an original contribution of this thesis based on [83] where an identifiability result based on this proof technique is given.

Equivalence classes and assumptions on $p_{\mathbf{s}|\mathbf{u}}$. In order to prove identifiability in this setting, further technical assumptions on the effect of variations in $\mathbf{u}$ on $\mathbf{x}$ are required, leading to identifiability up to different equivalence classes or tolerable ambiguities. One such condition, introduced in [66], is termed *variability*. Intuitively, variability demands that $\mathbf{u}$ has a sufficiently diverse influence on $\mathbf{x}$. Technically, it requires linear independence of vectors of first and second derivatives of scalar functions $q_1, \ldots, q_n$, which equal the univariate conditional densities of the sources given $\mathbf{u}$ (see Appendix A.2 for the formal definition). Under this assumption, it is possible to achieve identifiability up to the equivalence class in Defn. 2.4.1 (see [66, Thm.1]). The variability assumption is however hard to interpret, since it is not immediately apparent whether a given density satisfies it or not; [66, Thm. 2] provides some intuition on when variability is satisfied.

A different approach, also proposed in [66, Thm. 3], is to assume that the conditional distribution in (2.17) is an exponential family. A different set of assumptions can then be adopted: this leads to a weaker form of identifiability, which can be interpreted in terms of sufficient statistics of the conditional exponential family and for which we refer the interested reader to [66, Thm.1] and [84, 85].

Auxiliary variables: supervised or unsupervised? We remark that the auxiliary variables setting might appear closer to a *supervised* than to an *unsupervised* learning scenario, particularly if $\mathbf{u}$ is interpreted as a (class) label. However, if we take blind source separation as the main objective, we can think of the auxiliary variables as providing a handle to perform feature extraction, rather than being targets of a regression or classification problem. Additionally, [5] studies supervised learning problems, but focusing on the final layer representations extracted by different neural network classifiers—in the spirit of the independent-listeners problem we outlined in § 1.2, and establishing an identifiability result in that context where the proof technique is also inspired by the auxiliary variables setting.

Contrastive learning: identification or estimation? It is worth remarking that, in the works we reviewed, contrastive learning provides both a

proof of identifiability and suggests how to perform estimation, thereby apparently mixing identification and estimation. In fact, most of the results of identifiability originally expressed through contrastive learning can be rephrased in terms of likelihood, and are more similar to (2.11), as noted in [84]. Moreover, since [66], many other estimation procedures have been proposed, including variational autoencoders [84] and energy based models [85].

For estimation, a computational advantage of contrastive learning estimation over naive maximum likelihood is the following. The model likelihood for the nonlinear ICA generative model (2.1) can be written as

$$\log p_{\mathbf{x}}(\mathbf{x}) = \sum_{i=1}^{n} \log p_{s_i}(g_i(\mathbf{x})) + \underbrace{\log \left|\det \mathbf{J}_{\mathbf{g}}(\mathbf{x})\right|}_{\text{"Jacobian Term"}} \tag{2.18}$$

Where $\mathbf{g} = \mathbf{f}^{-1}$ and the Jacobian matrix $\mathbf{J}_{\mathbf{g}}(\mathbf{x}) \in \mathbb{R}^{n \times n}$ is defined by $[\mathbf{J}\mathbf{g}(\mathbf{x})]_{ij} = \partial g_i(\mathbf{x})/\partial x^j$.

As we will review in Chapter 5, the "Jacobian Term" makes optimisation of this likelihood problematic. There, we will also present one of our original contributions in estimation based on [86]: A method to optimise the exact maximum likelihood objective (2.18) efficiently for a class of deep neural networks.

2.4.3 Summary

Independent component analysis provides a principled approach to the study of some central problems in representation learning. In particular, it provides a way to answer the motivating problems we introduced in § 1.1, § 1.2. The focus on model identifiability, and the methodological separation between identification and estimation, will recur in most of the contributions presented in this thesis.

Starting with [64], there has been a renaissance of results in identifiability for nonlinear ICA or closely inspired by it [62, 65, 66, 78, 79, 85, 87–91]. Recently, the problem of learning representations where the components are not statistically independent received significant attention [92]: for example, we may be interested in learning latent variables which are causally related [13, 93, 94]; or where latent variables should faithfully reproduce symmetries of the world [11, 95, 96]. The theory of nonlinear ICA presented in § 2.4.1 bears important implications even for these settings, and beyond the assumption of statistically independent components, for example in the form of impossibility results for fully unsupervised nonlinear representation learning, see [62, 72, 74].

[64]: Hyvärinen et al. (2016), 'Unsupervised feature extraction by time-contrastive learning and nonlinear ICA'

ICA is also important beyond the problems in § 1.1 and § 1.2. Estimation of independent components is central to unsupervised learning and probabilistic modeling even beyond identifiability, see [97–99]; moreover, estimation of models with non-independent components can be inspired by extensions of the ICA model [100, 101]. Finally, many methods for causal discovery are based on identifiable ICA models [102–105].

2.4.4 Concluding Remarks on the Cocktail-Party and the Independent-Listeners problems

In § 1.1 and § 1.2, we described the two metaphors of the cocktail-party and independent-listerners problems. At different points throughout this chapter we referred to one or the other metaphor to illustrate different aspects of identifiability. Some concluding comments might be necessary to relate the nontechnical metaphors in Chapter 1 to the technical content of this chapter and the rest of this manuscript.

We would like to stress that studying the similarity of representations learned by different intelligent systems, as in § 1.2, might be relevant beyond the question whether such representations faithfully reconstruct some ground-truth and invert the true data-generating process—which is the objective of blind source separation and § 1.1. Comparing representations extracted by neural networks trained "independently" on the same dataset is the objective of some works we referred, e.g., [5, 85],[24] where various notions of identifiability are discussed to this end without any explicit notion of a ground truth generating process.[25]

For example, in [85, Sec. 2], the authors are interested in the outputs (extracted features) of conditional energy-based models, for which they characterise uniqueness up to certain equivalence classes based on suitable assumptions: only in a later part of the paper [85, Sec. 3] they link this identifiability to a generative model—Independently Modulated Components Analysis (IMCA), an extension of ICA allowing for some correlation among the latent variables—and to reconstruction of some ground truth latent variables.

In conclusion, the question introduced in (§ 1.2) through the independent-listeners problem may be considered separately from that of identifiying a ground truth (§ 1.1), and the notion of identifiability may be relevant to investigate both.[26]

[5]: Roeder et al. (2021), 'On linear identifiability of learned representations'
[85]: Khemakhem et al. (2020), 'ICE-BeeM: Identifiable Conditional Energy-Based Deep Models Based on Nonlinear ICA'

24: Interestingly, here the term "independently" is not used in the usual statistical sense and might be closer in spirit to the some nonstatistical interpretations thereof used in the context of causal inference, see [106] and § 3.2.1.

25: It may sometimes be unclear what the ground truth latent variables are in the first place, or whether they can be unambiguously defined: for example, a given causal system may be equivalently described by different sets of variables [107] or at different levels of abstraction [108–110].

26: The question of how representations extracted by different neural network models relate to each other has also been the subject of extensive empirical investigation, see, e.g., [111, 112].

Structure and Contributions of this Manuscript

The structure and contributions of this thesis are as follows.

In the first part, we focus on identification and estimation in the fully unsupervised setting with independent and identically distributed (i.i.d.) observations. As mentioned in § 2.4.1, identifiability in this setting requires restrictions on the mixing function. In Chapter 3 we introduce a setting termed *independent mechanism analysis (IMA)*, where a restriction of the mixing function class is derived, inspired by the principle of independent causal mechanisms [106], and its benefits for identifiability are discussed. In Chapter 4 we then show that variational autoencoders (VAEs), a prominent approach to unsupervised learning, optimise the IMA objective in a certain regime and under mild assumptions. Finally, in Chapter 5, we discuss the role of the Jacobian term of (2.18) in maximum likelihood estimation, and present a way to efficiently optimise it (and the whole likelihood) for a class of neural networks, based on *relative gradients*.

In the second part, we instead present results on identification and estimation of models where multiple (possibly corrupted) views of the same latent souces are available. In Chapter 6, we present an identifiability result for nonlinear ICA with multiple views. In Chapter 7, a related (linear) model is applied to the statistical analysis of group studies in neuroimaging. Closely related to the multi-view setting is the so called *"weakly supervised"* setting [94, 113, 114]: taking inspiration from it, in Chapter 8, we present an identifiability result for self-supervised learning with data augmentations. This theory will not require statistically independent components, and will instead focus on a suitable notion of identifiability (block-identification) for a subspace encoding those components which remain invariant under data augmentation.

In the conclusion, we present concluding remarks on the work presented in this thesis and on identifiability in representation learning, causal inference and connections with current machine learning practice.

Detailed list of contributions. The results presented in this thesis are based on work published with multiple venues and with different collaborators. Each of the chapters from Chapter 3 to Chapter 8 are based on the original published works almost verbatim; minor changes have been made with respect to the original works.

Part I:

- Chapter 3 is based on:
  L. Gresele*, J. von Kügelgen*, V. Stimper, B. Schölkopf, and M. Besserve. 'Independent mechanisms analysis, a new concept?' In: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. *equal contribution. Curran Associates, Inc., Dec. 2021
  My contributions included suggesting that the ICM principle could inspire useful constraints for nonlinear ICA, and, together with my co-authors, conceptualising the project, developing the theory and devising the experiments. The background section in the original

paper provided the basis for § 2.3, and part of the conclusion was moved to Chapter 9.

- Chapter 4 is based on:
  P. Reizinger*, L. Gresele*, J. Brady*, J. von Kügelgen, D. Zietlow, B. Schölkopf, G. Martius, W. Brendel, and M. Besserve. 'Embrace the Gap: VAEs Perform Independent Mechanism Analysis'. In: *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. *equal first authorship. Curran Associates, Inc., Dec. 2022
  This project was lead by P.R., and all authors were involved in structuring the research question. My contributions included investigating the link between normalizing flows and VAEs, and conceiving and discussing the theory and experiments with the other authors.
- Chapter 5 is based on:
  L. Gresele*, G. Fissore*, A. Javaloy, B. Schölkopf, and A. Hyvärinen. 'Relative gradient optimization of the Jacobian term in unsupervised deep learning'. In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. *equal contribution. Curran Associates, Inc., Dec. 2020
  My contributions included conceptualising the project together with A.H. and the other authors, and working with them on the theoretical analysis and implementation of the method.

Part II:

- Chapter 6 is based on:
  L. Gresele*, P. K. Rubenstein*, A. Mehrjou, F. Locatello, and B. Schölkopf. 'The Incomplete Rosetta Stone problem: Identifiability results for Multi-view Nonlinear ICA'. In: *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*. Vol. 115. Proceedings of Machine Learning Research. *equal contribution. PMLR, July 2019
  My contributions included proposing that the multi-view setting in nonlinear ICA could be identifiable, structuring the research questions and working on identifiability theory together with P.K.R. and all other authors.
- Chapter 7 is based on:
  H. Richard*, L. Gresele*, A. Hyvärinen, B. Thirion, A. Gramfort, and P. Ablin. 'Modeling Shared responses in Neuroimaging Studies through MultiView ICA'. In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. *equal contribution. Red Hook, NY: Curran Associates, Inc., Dec. 2020
  This project was lead by H.R. and conceptualised together with all the authors. My contributions included proposing the application of multi-view ICA to the statistical analysis of group studies in neuroimaging, heuristically deriving the objective (7.4), and suggesting to P.A. that an estimation procedure based on his previous linear ICA work could be derived. All authors contributed to writing the paper.
- Chapter 8 is based on:
  J. von Kügelgen*, Y. Sharma*, L. Gresele*, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. 'Self-supervised learning with data augmentations provably isolates content from style'. In: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. *equal

contribution. Curran Associates, Inc., Dec. 2021
The project was lead by J.v.K. and conceptualised with F.L. and all co-authors. My contributions included suggesting to study the problem in connection to the multi-view setting, working with other authors on the causal interpretation of the problem and discussing the experiments and theory (particularly in connection with nonlinear ICA theory) with all co-authors.

Other works I published in journals or at conferences during my PhD and which are not included in this thesis:

- Luigi Gresele and Matteo Marsili. 'On maximum entropy and inference'. In: *Entropy* 19.12 (2017)
- G. Parascandolo*, A. Neitz*, A. Orvieto, L. Gresele, and B. Schölkopf. 'Learning explanations that are hard to vary'. In: *9th International Conference on Learning Representations (ICLR)*. *equal contribution. May 2021
- J. von Kügelgen*, L. Gresele*, and B. Schölkopf. 'Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects'. In: *IEEE Transactions on Artificial Intelligence* 2.1 (2021). *equal contribution
- L. Gresele*, J. von Kügelgen*, J. M. Kübler*, E. Kirschbaum, B. Schölkopf, and D. Janzing. 'Causal Inference Through the Structural Causal Marginal Problem'. In: *Proceedings of the 39th International Conference on Machine Learning (ICML)*. Vol. 162. *equal contribution. PMLR, July 2022

Independent Influences and Estimation of the Jacobian Term

Independent Mechanism Analysis, A New Concept? 3

Independent component analysis provides a principled framework for unsupervised representation learning, with solid theory on the identifiability of the latent code that generated the data, given only observations of mixtures thereof. While as reviewed in § 2.4.1 the model is provably nonidentifiable when the mixing is nonlinear, identifiability can be recovered in settings where auxiliary variables are included in the generative process (§ 2.4.2). In this chapter, we investigate an alternative path and consider instead including assumptions reflecting the principle of *independent causal mechanisms* exploited in the field of causality. Specifically, our approach is motivated by thinking of each source as independently influencing the mixing process. This gives rise to a framework which we term independent mechanism analysis. We provide theoretical and empirical evidence that our approach circumvents a number of nonidentifiability issues arising in nonlinear blind source separation.

3.1 Introduction

One of the goals of unsupervised learning is to uncover properties of the data generating process, such as latent structures giving rise to the observed data. Identifiability formalises this desideratum: under suitable assumptions, a model learnt from observations should match the ground truth, up to well-defined ambiguities. In order to achieve identifiability in nonlinear BSS, a growing body of research postulates additional supervision or structure in the data generating process, often in the form of auxiliary variables (§ 2.4.2) or multiple views (Chapter 6).

In this chapter, we investigate a different route to identifiability by drawing inspiration from the field of *causal inference* [106, 122] which has provided useful insights for a number of machine learning tasks, including semi-supervised [123, 124], transfer [125–135], reinforcement [136–143], and unsupervised [13, 117, 144–149] learning. To this end, we *interpret the ICA mixing as a causal process* and apply the principle of independent causal mechanisms (ICM) which postulates that the generative process consists of independent modules which do not share information [106, 123, 150].

In this context, "independent" does not refer to *statistical* independence of random variables, but rather to the notion that the distributions and functions composing the generative process are chosen independently by Nature [150, 151]. While a formalisation of ICM [150, 152] in terms of algorithmic (Kolmogorov) complexity [153] exists, it is not computable, and hence applying ICM in practice requires assessing such non-statistical independence with suitable domain specific criteria [154]. The goal of our work is thus to *constrain the nonlinear ICA problem, in particular the mixing function, via suitable ICM measures*, thereby ruling out common counterexamples to identifiability which intuitively violate the ICM principle.

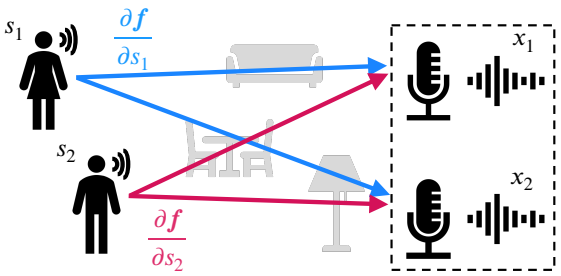


Figure 3.1: For the cocktail-party problem, the ICM principle *as traditionally understood* would say that the content of speech $p_{\mathbf{s}}$ is independent of the mixing or recording process $\mathbf{f}$ (microphone placement, room acoustics). IMA refines, or extends, this idea *at the level of the mixing function* by postulating that the contributions $\partial \mathbf{f}/\partial s_i$ of each source to $\mathbf{f}$, as captured by the speakers' positions relative to the recording process, should not be fine-tuned to each other.

Traditionally, ICM criteria have been developed for causal discovery, where *both cause and effect are observed* [155–158]. They enforce an independence between (i) the cause (source) distribution and (ii) the conditional or mechanism (mixing function) generating the effect (observations), and thus rely on the fact that the *observed* cause distribution is informative. As we will show, this renders them insufficient for nonlinear ICA, since the constraints they impose are satisfied by common counterexamples to identifiability. With this in mind, we introduce a new way to characterise or *refine* the ICM principle for unsupervised representation learning tasks such as nonlinear ICA.

Motivating example. To build intuition, we return to the cocktail-party problem of § 1.1, where a number of conversations are happening in parallel and the task is to recover the individual voices s_i from the recorded mixtures x_i. The mixing or recording process $\mathbf{f}$ is primarily determined by the room acoustics and the locations at which microphones are placed. Moreover, each speaker influences the recording through their positioning in the room, and we may think of this influence as $\partial \mathbf{f}/\partial s_i$, as illustrated in Fig. 3.1. Our independence postulate then amounts to stating that the speakers' positions are not fine-tuned to the room acoustics and microphone placement, or to each other, i.e., *the contributions $\partial \mathbf{f}/\partial s_i$ should be independent (in a non-statistical sense).*[1]

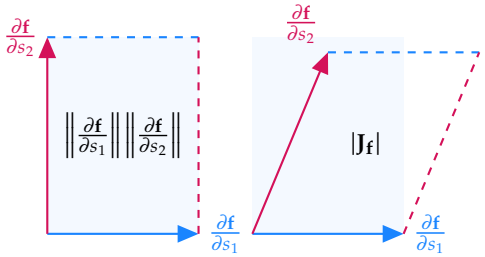


Figure 3.2: We formalise t independence between the $\partial \mathbf{f}/\partial s_i$, which are the columns of the Jacobian $\mathbf{J}_{\mathbf{f}}$, as an *orthogonality condition*: the absolute value of the determinant $|\mathbf{J}_{\mathbf{f}}|$, i.e., the volume of the parallelepiped spanned by $\partial \mathbf{f}/\partial s_i$, should decompose as the product of the norms of the $\partial \mathbf{f}/\partial s_i$.

Our approach. We formalise this notion of independence between the contributions $\partial \mathbf{f}/\partial s_i$ of each source to the mixing process (i.e., the columns of the Jacobian matrix $\mathbf{J}_{\mathbf{f}}$ of partial derivatives) as an orthogonality condition, see Fig. 3.2. Specifically, the absolute value of the determinant $|\mathbf{J}_{\mathbf{f}}|$, which describes the local change in infinitesimal volume induced by mixing the sources, should factorise or decompose as the product of the norms of its columns. This can be seen as a decoupling of the local influence of each partial derivative in the pushforward operation (mixing function) mapping the source distribution to the observed one, and gives rise to a novel framework which we term independent mechanism analysis (IMA). IMA can be understood as a refinement of the ICM principle that applies the idea of independence of mechanisms at the level of the mixing function.

Structure and contributions of this Chapter.

- We review existing ICM criteria (§ 3.2.1), and show that the latter do not sufficiently constrain nonlinear ICA (§ 3.3);

1: For additional intuition and possible violations in the context of the cocktail-party problem, see Appendix B.1.4.

- We propose a more suitable ICM criterion for unsupervised representation learning which gives rise to a new framework that we term independent mechanism analysis (IMA) (§ 3.4); we provide geometric and information-theoretic interpretations of IMA (§ 3.4.1), introduce an IMA contrast function which is invariant to the inherent ambiguities of nonlinear ICA (§ 3.4.2), and show that it rules out a large class of counterexamples and is consistent with existing identifiability results (§ 3.4.3);
- We experimentally validate our theoretical claims and propose a regularised maximum-likelihood learning approach based on the IMA constrast which outperforms the unregularised baseline (§ 3.5); additionally, we introduce a method to learn nonlinear ICA solutions with triangular Jacobian and a metric to assess BSS which can be of independent interest for the nonlinear ICA community.

3.2 Background and Preliminaries

Our work builds on and connects related literature from the fields of independent component analysis (§ 2.1) and causal inference. We start by assuming the generative model in (2.1). To recover identifiability, rather than relying only on additional assumptions on $\mathcal{P}$ (e.g., via auxiliary variables), we seek to further constrain (2.1) by also placing assumptions on the set $\mathcal{F}$ of mixing functions $\mathbf{f}$. To this end, we draw inspiration from the field of causal inference [106, 122]. Below we review concepts from causal inference relevant for our work.

[106]: Peters et al. (2017), *Elements of causal inference: foundations and learning algorithms*
[122]: Pearl (2009), *Causality*

3.2.1 Causal Inference and the Principle of Independent Causal Mechanisms (ICM)

Of central importance to our approach is the *Principle of Independent Causal Mechanisms* (ICM) [123, 150, 159].

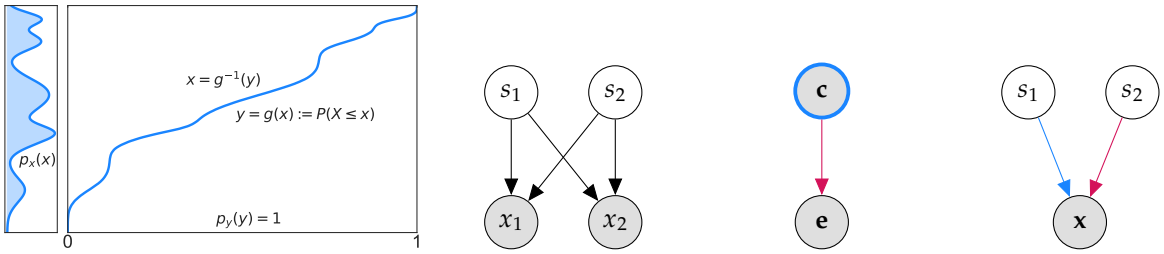
Principle 3.2.1 (ICM principle [106]) *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.*

These "modules" are typically thought of as the conditional distributions of each variable given its direct causes. Intuitively, the principle then states that these *causal conditionals* correspond to *independent mechanisms of nature* which do not share information. Crucially, here "independent" does not refer to *statistical* independence of random variables, but rather to independence of the underlying distributions as *algorithmic* objects. For a bivariate system comprising a cause $\mathbf{c}$ and an effect $\mathbf{e}$, this idea reduces to an independence of cause and mechanism, see Fig. 3.3c. One way to formalise ICM uses Kolmogorov complexity $K(\cdot)$ [153] as a measure of algorithmic information [150].

[150]: Janzing et al. (2010), 'Causal inference using the algorithmic Markov condition'

However, since Kolmogorov complexity is is not computable, using ICM in practice requires assessing Principle 3.2.1 with other suitable proxy criteria [106, 145, 157, 158, 160–167].[2] Allowing for deterministic relations between cause (sources) and effect (observations), the criterion which

2: *"This can be seen as an algorithmic analog of replacing the empirically undecidable question of statistical independence with practical independence tests that are based on assumptions on the underlying distribution"* [150].



(a) CDF transform **(b)** ICA **(c)** ICM **(d)** IMA

Figure 3.3: (a) Any observed density p_x can be mapped to a uniform p_y via the CDF transform $g(x) = \mathbb{P}(X \leq x)$; Darmois solutions $(\mathbf{f}^{\mathrm{D}}, p_{\mathbf{u}})$ constructed from (2.15) therefore automatically satisfy the independence postulated by IGCI (3.1). (b) ICA setting with $n = 2$ sources (shaded nodes are observed, white ones are unobserved). (c) Existing ICM criteria typically enforce independence between an observed input or cause distribution $p_{\mathbf{c}}$ and a mechanism $p_{\mathbf{e}|\mathbf{c}}$ (independent objects are highlighted in blue and red). (d) IMA enforces independence between the contributions of different sources s_i to the mixing function $\mathbf{f}$ as captured by $\partial \mathbf{f}/\partial s_i$.

is most closely related to the ICA setting in (2.1) is *information-geometric causal inference* (IGCI) [155, 156].[3] IGCI assumes a nonlinear relation $\mathbf{e} = \mathbf{f}(\mathbf{c})$ and formulates a notion of independence between the cause distribution $p_{\mathbf{c}}$ and the deterministic mechanism $\mathbf{f}$ (which we think of as a degenerate conditional $p_{\mathbf{e}|\mathbf{c}}$) via the following condition (in practice, assumed to hold approximately),

$$C_{\mathrm{IGCI}}(\mathbf{f}, p_{\mathbf{c}}) := \int \log |\mathbf{J}_{\mathbf{f}}(\mathbf{c})| \, p_{\mathbf{c}}(\mathbf{c}) d\mathbf{c} - \int \log |\mathbf{J}_{\mathbf{f}}(\mathbf{c})| \, d\mathbf{c} = 0 \,, \qquad (3.1)$$

where $(\mathbf{J}_{\mathbf{f}}(\mathbf{c}))_{ij} = \partial f_i / \partial c_j(\mathbf{c})$ is the Jacobian matrix and $|\cdot|$ the absolute value of the determinant. C_{IGCI} can be understood as the covariance between $p_{\mathbf{c}}$ and $\log |\mathbf{J}_{\mathbf{f}}|$ (viewed as r.v.s on the unit cube w.r.t. the Lebesgue measure), so that $C_{\mathrm{IGCI}} = 0$ rules out a form of fine-tuning between $p_{\mathbf{c}}$ and $|\mathbf{J}_{\mathbf{f}}|$. As its name suggests, IGCI can, from an information-geometric perspective, also be seen as an orthogonality condition between cause and mechanism in the space of probability distributions [156], see Appendix B.1.2, particularly Equation B.4 for further details.

3: For a similar criterion which assumes linearity [157, 158] and its relation to linear ICA, see Appendix B.1.1.

3.3 Existing ICM Measures are Insufficient for Nonlinear ICA

Our aim is to use the ICM Principle 3.2.1 to further constrain the space of models $\mathscr{M} \subseteq \mathscr{F} \times \mathscr{P}$ and rule out common counterexamples to identifiability such as those presented in § 2.4.1.

Intuitively, both the Darmois construction (2.15) and the rotated Gaussian MPA (2.16) give rise to "*non-generic*" solutions which should violate ICM: the former, $(\mathbf{f}^{\mathrm{D}}, p_{\mathbf{u}})$, due the triangular Jacobian of $\mathbf{f}^{\mathrm{D}}$ (see Remark 2.4.1), meaning that each observation $x_i = f_i^{\mathrm{D}}(\mathbf{y}_{1:i})$ only depends on a subset of the inferred independent components $\mathbf{y}_{1:i}$, and the latter, $(\mathbf{f} \circ \mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}}), p_{\mathbf{s}})$, due to the dependence of $\mathbf{f} \circ \mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}})$ on $p_{\mathbf{s}}$ (2.16).

However, the ICM criteria described in § 3.2.1 were developed for the task of cause-effect inference where *both variables are observed*. In contrast, in this work, we consider an unsupervised representation learning task where *only the effects* (mixtures $\mathbf{x}$) *are observed*, but the causes (sources $\mathbf{s}$) are not. It turns out that this renders existing ICM criteria insufficient

for BSS: they can easily be satisfied by spurious solutions which are not equivalent to the true one. We can show this for IGCI. Denote by $\mathcal{M}_{\text{IGCI}} = \{(\mathbf{f}, p_{\mathbf{s}}) \in \mathcal{F} \times \mathcal{P} : C_{\text{IGCI}}(\mathbf{f}, p_{\mathbf{s}}) = 0\} \subset \mathcal{F} \times \mathcal{P}$ the class of nonlinear ICA models satisfying IGCI (3.1). Then the following negative result holds.

Proposition 3.3.1 (IGCI is insufficient for $\sim_{\text{BSS}}$-identifiability) (2.1) *is not $\sim_{\text{BSS}}$-identifiable on $\mathcal{M}_{\text{IGCI}}$.*

Proof. IGCI (3.1) is satisfied when $p_{\mathbf{s}}$ is uniform. However, the Darmois construction (2.15) yields uniform sources, see Fig. 3.3a. This means that $(\mathbf{f}^{\text{D}} \circ \mathbf{a}^{\mathbf{R}}(p_{\mathbf{u}}), p_{\mathbf{u}}) \in \mathcal{M}_{\text{IGCI}}$, so IGCI can be satisfied by solutions which do not separate the sources in the sense of Defn. 2.4.1, see § 2.4.1 and [70]. □

As illustrated in Fig. 3.3c, condition (3.1) and other similar criteria enforce a notion of "genericity" or "decoupling" of the mechanism w.r.t. the *observed* input distribution.[4] They thus rely on the fact that the cause (source) distribution is informative, and are generally not invariant to reparametrisation of the cause variables. In the (nonlinear) ICA setting, on the other hand, the *learnt* source distribution may be fairly uninformative. This poses a challenge for existing ICM criteria since any mechanism is generic w.r.t. an uninformative (uniform) input distribution.

4: In fact, many ICM criteria can be phrased as special cases of a unifying group-invariance framework [145].

3.4 Independent Mechanism Analysis (IMA)

As argued in § 3.3, enforcing independence between the input distribution and the mechanism (Fig. 3.3c), as existing ICM criteria do, is insufficient for ruling out spurious solutions to nonlinear ICA. We therefore propose a new ICM-inspired framework which is more suitable for BSS and which we term *independent mechanism analysis* (IMA).[5] All proofs are provided in Appendix B.2.

5: The title of the paper in which the material presented in this chapter was originally published is thus a reverence to Pierre Comon's seminal 1994 paper [37].

3.4.1 Intuition Behind IMA

As motivated using the cocktail party example in § 1.1 and Fig. 3.1, our main idea is to enforce a notion of *independence between the contributions or influences of the different sources s_i on the observations* $\mathbf{x} = \mathbf{f}(\mathbf{s})$ as illustrated in Fig. 3.3d—as opposed to between the source distribution and mixing function, cf. Fig. 3.3c. These contributions or influences are captured by the vectors of partial derivatives $\partial \mathbf{f}/\partial s_i$. IMA can thus be understood as a *refinement of ICM at the level of the mixing* $\mathbf{f}$: in addition to *statistically independent components* s_i, we look for a mixing with *contributions* $\partial \mathbf{f}/\partial s_i$ *which are independent*, in a non-statistical sense which we formalise as follows.

Principle 3.4.1 (IMA) *The mechanisms by which each source s_i influences the observed distribution, as captured by the partial derivatives $\partial \mathbf{f}/\partial s_i$, are*

independent of each other in the sense that for all $\mathbf{s}$*:*

$$\log |\mathbf{J_f}(\mathbf{s})| = \sum_{i=1}^{n} \log \left\| \frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{s}) \right\| \tag{3.2}$$

Geometric interpretation. Geometrically, the IMA principle can be understood as an *orthogonality condition*, as illustrated for $n = 2$ in Fig. 3.2. First, the vectors of partial derivatives $\partial \mathbf{f}/\partial s_i$, for which the IMA principle postulates independence, are the *columns* of $\mathbf{J_f}$. $|\mathbf{J_f}|$ thus measures the volume of the $n-$dimensional parallelepiped spanned by these columns, as shown on the right. The product of their norms, on the other hand, corresponds to the volume of an n-dimensional box, or rectangular parallelepiped with side lengths $\|\partial \mathbf{f}/\partial s_i\|$, as shown on the left. The two volumes are equal if and only if all columns $\partial \mathbf{f}/\partial s_i$ of $\mathbf{J_f}$ are orthogonal. Note that (3.2) is trivially satisfied for $n = 1$, i.e., if there is no mixing, further highlighting its difference from ICM for causal discovery.

Independent influences and orthogonality. In a high dimensional setting (large n), this orthogonality can be intuitively interpreted from the ICM perspective as *Nature choosing the direction of the influence of each source component in the observation space independently and from an isotropic prior*. Indeed, it can be shown that the scalar product of two independent isotropic random vectors in $\mathbb{R}^n$ vanishes as the dimensionality n increases (equivalently: two high-dimensional isotropic vectors are typically orthogonal). This property was previously exploited in other linear ICM-based criteria (see [168, Lemma 5] and [157, Lemma 1 & Thm. 1]).[6] The principle in (3.2) can be seen as a constraint on the function space, enforcing such orthogonality between the columns of the Jacobian of $\mathbf{f}$ at all points in the source domain, thus approximating the high-dimensional behavior described above.[7]

6: This has also been used as a *"leading intuition"* [sic] to interpret IGCI in [156].

7: To provide additional intuition on how IMA differs from existing principles of independence of cause and mechanism, we give examples, both technical and pictorial, of violations of both in Appendix B.1.4.

Information-geometric interpretation and comparison to IGCI. The additive contribution of the sources' influences $\partial \mathbf{f}/\partial s_i$ in (3.2) suggests their local *decoupling at the level of the mechanism* $\mathbf{f}$. Note that IGCI (3.1), on the other hand, postulates a different type of decoupling: one between $\log |\mathbf{J_f}|$ and $p_\mathbf{s}$. There, dependence between cause and mechanism can be conceived as a fine tuning between the derivative of the mechanism and the input density. The IMA principle leads to a complementary, non-statistical measure of independence between the influences $\partial \mathbf{f}/\partial s_i$ of the individual sources on the vector of observations. Both the IGCI and IMA postulates have an information-geometric interpretation related to the influence of ("non-statistically") independent modules on the observations: both lead to an *additive decomposition of a KL-divergence between the effect distribution and a reference distribution.* For IGCI, independent modules correspond to the cause distribution and the mechanism mapping the cause to the effect (see (B.4) in Appendix B.1.2). For IMA, on the other hand, these are the influences of each source component on the observations in an interventional setting (under soft interventions on individual sources), as measured by the KL-divergences between the original and intervened distributions. See Appendix B.1.3, and especially (B.7), for a more detailed account.

We finally remark that while recent work based on the ICM principle has mostly used the term "mechanism" to refer to causal Markov kernels $p(X_i|PA_i)$ or structural equations [106], we employ it in line with the broader use of this concept in the philosophical literature.[8] To highlight just two examples, [170] states that *"Causal processes, causal interactions, and causal laws provide the mechanisms by which the world works; to understand why certain things happen, we need to see how they are produced by these mechanisms"*; and [171] states that *"Mechanisms are events that alter relations among some specified set of elements"*. Following this perspective, we argue that a causal mechanism can more generally denote any process that describes the way in which causes influence their effects: the partial derivative $\partial \mathbf{f}/\partial s_i$ thus reflects a causal mechanism in the sense that it describes the infinitesimal changes in the observations $\mathbf{x}$, when an infinitesimal perturbation is applied to s_i.

8: See Table 1 in [169] for a long list of definitions from the literature.

3.4.2 Definition and Useful Properties of the IMA Contrast

We now introduce a contrast function based on the IMA principle (3.2) and show that it possesses several desirable properties in the context of nonlinear ICA. First, we define a local contrast as the difference between the two integrands of (3.2) for a particular value of the sources $\mathbf{s}$.

Definition 3.4.1 (Local IMA contrast) *The local IMA contrast $c_{\text{IMA}}(\mathbf{f},)$ of $\mathbf{f}$ at a point $\mathbf{s}$ is given by*

$$c_{\text{IMA}}(\mathbf{f},) = \sum_{i=1}^{n} \log \left\| \frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{s}) \right\| - \log |\mathbf{J}_{\mathbf{f}}(\mathbf{s})| \ . \tag{3.3}$$

Remark 3.4.1 This corresponds to the left KL measure of diagonality [172] for $\sqrt{\mathbf{J}_{\mathbf{f}}(\mathbf{s})^\top \mathbf{J}_{\mathbf{f}}(\mathbf{s})}$.

The local IMA contrast $c_{\text{IMA}}(\mathbf{f},)$ quantifies the extent to which the IMA principle is violated at a given point $\mathbf{s}$. We summarise some of its properties in the following proposition.

Proposition 3.4.1 (Properties of $c_{\text{IMA}}(\mathbf{f},)$) *The local IMA contrast $c_{\text{IMA}}(\mathbf{f},)$ defined in (3.3) satisfies:*

- *(i)* $c_{\text{IMA}}(\mathbf{f},) \geq 0$, *with equality if and only if all columns $\partial \mathbf{f}/\partial s_i(\mathbf{s})$ of $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ are orthogonal.*
- *(ii)* $c_{\text{IMA}}(\mathbf{f},)$ *is invariant to left multiplication of $\mathbf{J}_{\mathbf{f}}(\mathbf{s})$ by an orthogonal matrix and to right multiplication by permutation and diagonal matrices.*

Property *(i)* formalises the geometric interpretation of IMA as an orthogonality condition on the columns of the Jacobian from § 3.4.1, and property *(ii)* intuitively states that changes of orthonormal basis and permutations or rescalings of the columns of $\mathbf{J}_{\mathbf{f}}$ do not affect their orthogonality. Next, we define a global IMA contrast w.r.t. a source distribution $p_{\mathbf{s}}$ as the expected local IMA contrast.

Definition 3.4.2 (Global IMA contrast) *The global IMA contrast* $C_{IMA}(\mathbf{f}, p_\mathbf{s})$ *of* $\mathbf{f}$ *w.r.t.* $p_\mathbf{s}$ *is given by*

$$C_{IMA}(\mathbf{f}, p_\mathbf{s}) = \mathbb{E}_{\mathbf{s}\sim p_\mathbf{s}}[c_{IMA}(\mathbf{f}, \mathbf{s})] = \int c_{IMA}(\mathbf{f}, \mathbf{s}) p_\mathbf{s}(\mathbf{s}) d\mathbf{s}\,. \tag{3.4}$$

The global IMA contrast $C_{\text{IMA}}(\mathbf{f}, p_\mathbf{s})$ thus quantifies the extent to which the IMA principle is violated for a particular solution $(\mathbf{f}, p_\mathbf{s})$ to the nonlinear ICA problem. We summarise its properties as follows.

Proposition 3.4.2 (Properties of $C_{\text{IMA}}(\mathbf{f}, p_\mathbf{s})$) *The global IMA contrast* $C_{IMA}(\mathbf{f}, p_\mathbf{s})$ *from* (3.4) *satisfies:*

(i) $C_{IMA}(\mathbf{f}, p_\mathbf{s}) \geq 0$, *with equality iff.* $\mathbf{J}_\mathbf{f}(\mathbf{s}) = \mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s})$ *almost surely w.r.t.* $p_\mathbf{s}$, *where* $\mathbf{O}(\mathbf{s}), \mathbf{D}(\mathbf{s}) \in \mathbb{R}^{n\times n}$ *are orthogonal and diagonal matrices, respectively;*

(ii) $C_{IMA}(\mathbf{f}, p_\mathbf{s}) = C_{IMA}(\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}})$ *for any* $\tilde{\mathbf{f}} = \mathbf{f} \circ \mathbf{h}^{-1} \circ \mathbf{P}^{-1}$ *and* $\tilde{\mathbf{s}} = \mathbf{P}\mathbf{h}(\mathbf{s})$, *where* $\mathbf{P} \in \mathbb{R}^{n\times n}$ *is a permutation and* $\mathbf{h}(\mathbf{s}) = (h_1(s_1), \ldots, h_n(s_n))$ *an invertible element-wise function.*

Property *(i)* is the distribution-level analogue to *(i)* of Proposition 3.4.1 and only allows for orthogonality violations on sets of measure zero w.r.t. $p_\mathbf{s}$. This means that C_{IMA} can only be zero if $\mathbf{f}$ is an *orthogonal coordinate transformation* almost everywhere [173–175], see Fig. 3.4 for an example. We particularly stress property *(ii)*, as it precisely matches the inherent indeterminacy of nonlinear ICA: C_{IMA} *is blind to reparametrisation of the sources by permutation and element wise transformation*.

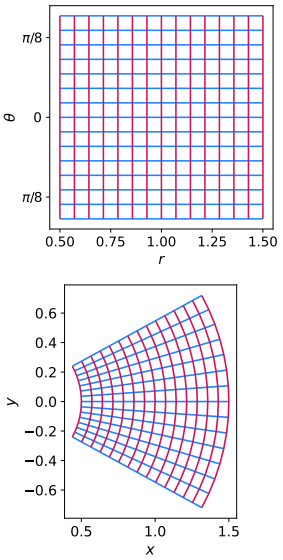


Figure 3.4: An example of a (non-conformal) orthogonal coordinate transformation from polar (left) to Cartesian (right) coordinates.

3.4.3 Theoretical Analysis and Justification of C_{IMA}

We now show that, under suitable assumptions on the generative model (2.1), a large class of spurious solutions—such as those based on the Darmois construction (2.15) or measure preserving automorphisms such as $\mathbf{a}^\mathbf{R}$ from (2.16) as described in § 2.4.1—exhibit nonzero IMA contrast. Denote the class of nonlinear ICA models satisfying (3.2) (IMA) by $\mathcal{M}_{\text{IMA}} = \{(\mathbf{f}, p_\mathbf{s}) \in \mathcal{F} \times \mathcal{P} : C_{\text{IMA}}(\mathbf{f}, p_\mathbf{s}) = 0\} \subset \mathcal{F} \times \mathcal{P}$. Our first main theoretical result is that, under mild assumptions on the observations, Darmois solutions will have strictly positive C_{IMA}, making them distinguishable from those in $\mathcal{M}_{\text{IMA}}$.

Theorem 3.4.3 *Assume the data generating process in* (2.1) *and assume that* $x_i \not\perp\!\!\!\perp x_j$ *for some* $i \neq j$. *Then any Darmois solution* $(\mathbf{f}^D, p_\mathbf{u})$ *based on* $\mathbf{g}^D$ *as defined in* (2.15) *satisfies* $C_{IMA}(\mathbf{f}^D, p_\mathbf{u}) > 0$. *Thus a solution satisfying* $C_{IMA}(\mathbf{f}, p_\mathbf{s}) = 0$ *can be distinguished from* $(\mathbf{f}^D, p_\mathbf{u})$ *based on the contrast* C_{IMA}.

The proof is based on the fact that the Jacobian of $\mathbf{g}^\text{D}$ is triangular (see Remark 2.4.1) and on the specific form of (2.15). A specific example of a mixing process satisfying the IMA assumption is the case where $\mathbf{f}$ is a conformal (angle-preserving) map.

Definition 3.4.3 (Conformal map) *A smooth map* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ *is conformal if* $\mathbf{J}_\mathbf{f}(\mathbf{s}) = \mathbf{O}(\mathbf{s})\lambda(\mathbf{s})\ \forall \mathbf{s}$, *where* $\lambda : \mathbb{R}^n \to \mathbb{R}$ *is a scalar field, and*

$\mathbf{O} \in O(n)$ *is an orthogonal matrix.*

Corollary 3.4.4 *Under assumptions of Thm. 3.4.3, if additionally* $\mathbf{f}$ *is a conformal map, then* $(\mathbf{f}, p_\mathbf{s}) \in \mathcal{M}_{\text{IMA}}$ *for any* $p_\mathbf{s} \in \mathcal{P}$ *due to Proposition 3.4.2 (i), see Defn. 3.4.3. Based on Thm. 3.4.3,* $(\mathbf{f}, p_\mathbf{s})$ *is thus distinguishable from Darmois solutions* $(\mathbf{f}^D, p_\mathbf{u})$.

This is consistent with a result that proves identifiability of conformal maps for $n = 2$ and conjectures it in general [70].[9] However, conformal maps are only a small subset of all maps for which $C_{\text{IMA}} = 0$, as is apparent from the more flexible condition of Proposition 3.4.2 *(i)*, compared to the stricter Defn. 3.4.3.

9: Note that Corollary 3.4.4 holds for any dimensionality n.

Example 3.4.1 (Polar to Cartesian coordinate transform) Consider the *non-conformal* transformation from polar to Cartesian coordinates (see Fig. 3.4), defined as $(x, y) = \mathbf{f}(r, \theta) := (r\cos(\theta), r\sin(\theta))$ with independent sources $=(r, \theta)$, with $r \sim U(0, R)$ and $\theta \sim U(0, 2\pi)$.[a] Then, $C_{\text{IMA}}(\mathbf{f}, p_\mathbf{s}) = 0$ and $C_{\text{IMA}}(\mathbf{f}^D, p_\mathbf{u}) > 0$ for any Darmois solution $(\mathbf{f}^D, p_\mathbf{u})$ —see Appendix B.3 for details.

[a] For different $p_\mathbf{s}$, (x, y) can be made to have independent Gaussian components ([73], II.B), and C_{IMA}-identifiability is lost; this shows that the assumption of Thm. 3.4.3 that $x_i \not\perp\!\!\!\perp x_j$ for some $i \neq j$ is crucial.

Finally, for the case in which the true mixing is linear, we obtain the following result.

Corollary 3.4.5 *Consider a linear ICA model,* $\mathbf{x} = \mathbf{A}\mathbf{s}$, *with* $\mathbb{E}[\mathbf{s}^\top \mathbf{s}] = \mathbf{I}$, *and* $\mathbf{A} \in O(n)$ *an orthogonal, non-trivial mixing matrix, i.e., not the product of a diagonal and a permutation matrix* $\mathbf{DP}$. *If at most one of the* s_i *is Gaussian, then* $C_{\text{IMA}}(\mathbf{A}, p_\mathbf{s}) = 0$ *and* $C_{\text{IMA}}(\mathbf{f}^D, p_\mathbf{u}) > 0$.

In a "blind" setting, we may not know a priori whether the true mixing is linear or not, and thus choose to learn a nonlinear unmixing. Corollary 3.4.5 shows that, in this case, Darmois solutions are still distinguishable from the true mixing via C_{IMA}. Note that unlike in Corollary 3.4.4, the assumption that $x_i \not\perp\!\!\!\perp x_j$ for some $i \neq j$ is not required for Corollary 3.4.5. In fact, due to Theorem 11 of [37], it follows from the assumed linear ICA model with non-Gaussian sources, and the fact that the mixing matrix is not the product of a diagonal and a permutation matrix (see also § 2.2.2).

Having shown that the IMA principle allows to distinguish a class of models (including, but not limited to conformal maps) from Darmois solutions, we next turn to a second well-known counterexample to identifiability: the "rotated-Gaussian" MPA $\mathbf{a}^\mathbf{R}(p_\mathbf{s})$ (2.16) from Defn. 2.4.3. Our second main theoretical result is that, under suitable assumptions, this class of MPAs can also be ruled out for "non-trivial" $\mathbf{R}$.

Theorem 3.4.6 *Let* $(\mathbf{f}, p_\mathbf{s}) \in \mathcal{M}_{\text{IMA}}$ *and assume that* $\mathbf{f}$ *is a conformal map. Given* $\mathbf{R} \in O(n)$, *assume additionally that* $\exists$ *at least one non-Gaussian* s_i *whose associated canonical basis vector* $\mathbf{e}_i$ *is not transformed by* $\mathbf{R}^{-1} = \mathbf{R}^\top$ *into another canonical basis vector* $\mathbf{e}_j$. *Then* $C_{\text{IMA}}(\mathbf{f} \circ \mathbf{a}^\mathbf{R}(p_\mathbf{s}), p_\mathbf{s}) > 0$.

Thm. 3.4.6 states that for conformal maps, applying the $\mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}})$ transformation at the level of the sources leads to an increase in C_{IMA} except for very specific rotations $\mathbf{R}$ that are "fine-tuned" to $p_{\mathbf{s}}$ in the sense that they permute all non-Gaussian sources s_i with another s_j. Interestingly, as for the linear case, non-Gaussianity again plays an important role in the proof of Thm. 3.4.6.

3.5 Experiments

Our theoretical results from § 3.4 suggest that C_{IMA} is a promising contrast function for nonlinear blind source separation. We test this empirically by evaluating the C_{IMA} of spurious nonlinear ICA solutions (§ 3.5.1), and using it as a learning objective to recover the true solution (§ 3.5.2).

We sample the ground truth sources from a uniform distribution in $[0, 1]^n$; the reconstructed sources are also mapped to the uniform hypercube as a reference measure via the CDF transform. Unless otherwise specified, the ground truth mixing $\mathbf{f}$ is a Möbius transformation [176] (i.e., a conformal map) with randomly sampled parameters, thereby satisfying Principle 3.4.1. In all of our experiments, we use JAX [177] and Distrax [178]. For additional technical details, equations and plots see Appendix B.4. The code to reproduce our experiments is available at https://github.com/lgresele/independent-mechanism-analysis.

3.5.1 Numerical Evaluation of the C_{IMA} Contrast for Spurious Nonlinear ICA Solutions

Learning the Darmois construction. To learn the Darmois construction from data, we use normalising flows, see [71, 179]. Since Darmois solutions have triangular Jacobian (Remark 2.4.1), we use an architecture based on residual flows [180] which we constrain such that the Jacobian of the full model is triangular. This yields an expressive model which we train effectively via maximum likelihood.

C_{IMA} of Darmois solutions. To check whether Darmois solutions (learnt from finite data) can be distinguished from the true one, as predicted by Thm. 3.4.3, we generate 1000 random mixing functions for $n = 2$, compute the C_{IMA} values of learnt solutions, and find that all values are indeed significantly larger than zero, see Fig. 3.5 **(a)**. The same holds for higher dimensions, see Fig. 3.5 **(b)** for results with 50 random mixings for $n \in \{2, 3, 5, 10\}$: with higher dimensionality, both the mean and variance of the C_{IMA} distribution for the learnt Darmois solutions generally attain higher values.[10] We confirmed these findings for mappings which are not conformal, while still satisfying (3.2), in Appendix B.4.5.

10: the latter possibly due to the increased difficulty of the learning task for larger n

C_{IMA} of MPAs. We also investigate the effect on C_{IMA} of applying an MPA $\mathbf{a}^{\mathbf{R}}(\cdot)$ from (2.16) to the true solution or a learnt Darmois solution. Results for $n = 2$ dim. for different rotation matrices $\mathbf{R}$ (parametrised by the angle θ) are shown in Fig. 3.5 **(c)**. As expected, the behavior is periodic in θ, and vanishes for the true solution (blue) at multiples of $\pi/2$,

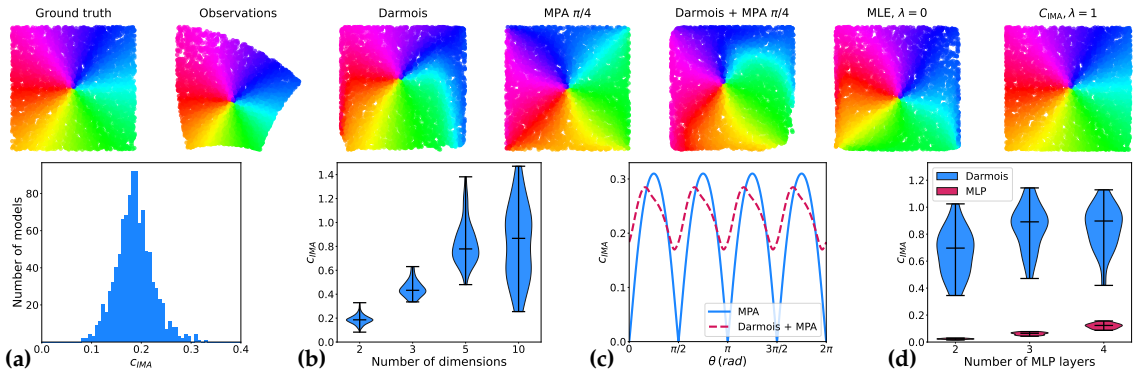


Figure 3.5: Top. Visual comparison of different nonlinear ICA solutions for $n = 2$: *(left to right)* true sources; observed mixtures; Darmois solution; true unmixing, composed with the measure preserving automorphism (MPA) from (2.16) (with rotation by $\pi/4$); Darmois solution composed with the same MPA; maximum likelihood ($\lambda = 0$); and C_{IMA}-regularised approach ($\lambda = 1$). **Bottom.** Quantitative comparison of C_{IMA} for different spurious solutions: learnt Darmois solutions for **(a)** $n = 2$, and **(b)** $n \in \{2, 3, 5, 10\}$ dimensions; **(c)** composition of the MPA (2.16) in $n = 2$ dim. with the true solution (blue) and a Darmois solution (red) for different angles. **(d)** C_{IMA} distribution for true MLP mixing (red) vs. Darmois solution (blue) for $n = 5$ dim., $L \in \{2, 3, 4\}$ layers.

i.e., when $\mathbf{R}$ is a permutation matrix, as predicted by Thm. 3.4.6. For the learnt Darmois solution (red, dashed) C_{IMA} remains larger than zero.

C_{IMA} values for random MLPs. Lastly, we study the behavior of spurious solutions based on the Darmois construction under deviations from our assumption of $C_{\text{IMA}} = 0$ for the true mixing function. To this end, we use invertible MLPs with orthogonal weight initalisation and `leaky_tanh` activations [86] as mixing functions; the more layers L are added to the mixing MLP, the larger a deviation from our assumptions is expected. We compare the true mixing and learnt Darmois solutions over 20 realisations for each $L \in \{2, 3, 4\}$, $n = 5$. Results are shown in figure Fig. 3.5 **(d)**: the C_{IMA} of the mixing MLPs grows with L; still, the one of the Darmois solution is typically higher.

Summary. We verify that spurious solutions can be distinguished from the true one based on C_{IMA}.

3.5.2 Learning Nonlinear ICA Solutions with C_{IMA}-Regularised Maximum Likelihood

Experimental setup. To use C_{IMA} as a learning signal, we consider a regularised maximum-likelihood approach, with the following objective:

$$\mathcal{L}(\mathbf{g}) = \mathbb{E}_{\mathbf{x}}[\log p_{\mathbf{g}}(\mathbf{x})] - \lambda\, C_{\text{IMA}}(\mathbf{g}^{-1}, p_{\mathbf{y}})\,, \tag{3.5}$$

where $\mathbf{g}$ denotes the learnt unmixing, $\mathbf{y} = \mathbf{g}(\mathbf{x})$ the reconstructed sources, and $\lambda \geq 0$ a Lagrange multiplier. For $\lambda = 0$, this corresponds to standard maximum likelihood estimation, whereas for $\lambda > 0$, $\mathcal{L}$ lower-bounds the likelihood, and recovers it exactly iff. $(\mathbf{g}^{-1}, p_{\mathbf{y}}) \in \mathcal{M}_{\text{IMA}}$. We train a residual flow $\mathbf{g}$ (with full Jacobian) to maximise $\mathcal{L}$. For evaluation, we compute (i) the KL divergence to the true data likelihood, as a measure of goodness of fit for the learnt flow model; and (ii) the mean correlation coefficient (MCC) between ground truth and reconstructed sources [62, 64]. We also

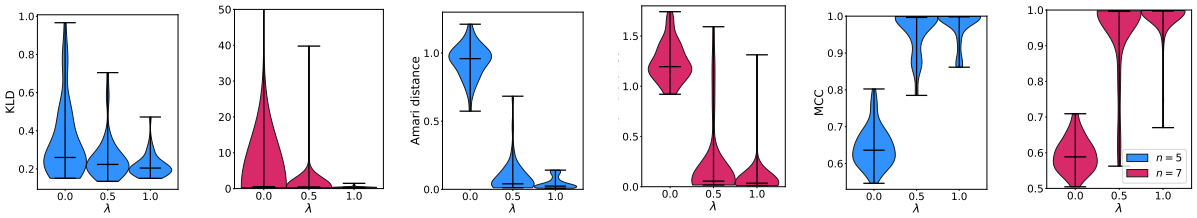


Figure 3.6: BSS via C_{IMA}-regularised MLE for, side by side, $n = 5$ (blue) and $n = 7$ (red) dim. with $\lambda \in \{0.0, 0.5, 1.0\}$. *(Left)* KL-divergence between ground truth likelihood and learnt model; *(center)* nonlinear Amari distance given true mixing and learnt unmixing; *(right)* MCC between true and reconstructed sources.

introduce (iii) a nonlinear extension of the Amari distance [181] between the true mixing and the learnt unmixing, which is larger than or equal to zero, with equality iff. the learnt model belongs to the BSS equivalence class (Defn. 2.4.1) of the true solution, see Appendix B.4.5 for details.

Results. In Fig. 3.5 *(Top)*, we show an example of the distortion induced by different *spurious* solutions for $n = 2$, and contrast it with a solution learnt using our proposed objective *(rightmost plot)*. Visually, we find that the C_{IMA}-regularised solution (with $\lambda = 1$) recovers the true sources most faithfully. Quantitative results for 50 learnt models for each $\lambda \in \{0.0, 0.5, 1.0\}$ and $n \in \{5, 7\}$ are summarised in Fig. 3.6 (see Appendix B.4 for additional plots). As indicated by the KL divergence values *(left)*, most trained models achieve a good fit to the data across all values of λ.[11] We observe that using C_{IMA} (i.e., $\lambda > 0$) is beneficial for BSS, both in terms of our nonlinear Amari distance *(center, lower is better)* and MCC *(right, higher is better)*, though we do not observe a substantial difference between $\lambda = 0.5$ and $\lambda = 1$.[12]

11: models with $n = 7$ have high outlier KL values, seemingly less pronounced for nonzero values of λ

12: In Appendix B.4.5, we also show that our method is superior to a linear ICA baseline, FastICA [182].

Summary: C_{IMA} can be a useful learning signal to recover the true solution.

3.6 Discussion

Assumptions on the mixing function. Instead of relying on weak supervision in the form of auxiliary variables or multiple views, our IMA approach places additional constraints on the functional form of the mixing process.

In a similar vein, the *minimal nonlinear distortion principle* [183] proposes to favor solutions that are as close to linear as possible. Another example is the *post-nonlinear model* [73, 103], which assumes an element-wise nonlinearity applied after a linear mixing. IMA is different in that it still allows for strongly nonlinear mixings (see, e.g., Fig. 3.4) provided that the columns of their Jacobians are (close to) orthogonal. In the related field of disentanglement [4, 72], a line of work that focuses on image generation with adversarial networks [82] similarly proposes to constrain the "generator" function via regularisation of its Jacobian [184] or Hessian [185], though mostly from an empirically-driven, rather than from an identifiability perspective as in the present work.

Towards identifiability with C_{IMA}. The IMA principle rules out a large class of spurious solutions to nonlinear ICA. While we do not present a full identifiability result, our experiments show that C_{IMA} can be used to recover the BSS equivalence class, suggesting that identifiability might indeed hold, possibly under additional assumptions—e.g., for conformal maps [70].

IMA and independence of cause and mechanism. While inspired by measures of independence of cause and mechanism as traditionally used for cause-effect inference [155–158], we view the IMA principle as addressing a different question, in the sense that they evaluate independence between different elements of the causal model. Any nonlinear ICA solution that satisfies the IMA Principle 3.4.1 can be turned into one with uniform reconstructed sources—thus satisfying IGCI as argued in § 3.3—through composition with an element-wise transformation which, according to Proposition 3.4.2 *(ii)*, leaves the C_{IMA} value unchanged. Both IGCI (3.1) and IMA (3.2) can therefore be fulfilled simultaneosly, while the former on its own is inconsequential for BSS as shown in Proposition 3.3.1.

BSS through algorithmic information. Algorithmic information theory has previously been proposed as a unifying framework for identifiable approaches to *linear* BSS [186, 187], in the sense that commonly-used contrast functions could, under suitable assumptions, be interpreted as proxies for the total complexity of the mixing and the reconstructed sources. However, to the best of our knowledge, the problem of specifying suitable proxies for the complexity of *nonlinear* mixing functions has not yet been addressed. We conjecture that our framework could be linked to this view, based on the additional assumption of algorithmic independence of causal mechanisms [150], thus potentially representing an approach to *nonlinear* BSS by minimisation of algorithmic complexity.

Conclusion. We introduced IMA, a path to nonlinear BSS inspired by concepts from causality. We postulate that the *influences* of different sources on the observed distribution should be approximately independent, and formalise this as an orthogonality condition on the columns of the Jacobian. We prove that this constraint is generally violated by well-known spurious nonlinear ICA solutions, and propose a regularised maximum likelihood approach which we empirically demonstrate to be effective in recovering the true solution. Our IMA principle holds exactly for orthogonal coordinate transformations, and is thus of potential interest for learning spatial representations [188], robot dynamics [189], or physics problems where orthogonal reference frames are common [175].

Embrace the Gap: VAEs Perform Independent Mechanism Analysis

4

Variational autoencoders (VAEs) are a popular framework for modeling complex data distributions; they can be efficiently trained via variational inference by maximizing the evidence lower bound (ELBO), at the expense of a gap to the exact (log-)marginal likelihood. While VAEs are commonly used for representation learning, it is unclear why ELBO maximization would yield useful representations, since unregularised maximum likelihood estimation cannot invert the data-generating process. Yet, VAEs often succeed at this task. We seek to elucidate this apparent paradox by studying nonlinear VAEs in the limit of near-deterministic decoders. We first prove that, in this regime, the optimal encoder approximately inverts the decoder—a commonly used but unproven conjecture—which we refer to as *self-consistency*. Leveraging self-consistency, we show that the ELBO converges to a regularized log-likelihood. This adds an inductive bias towards decoders with column-orthogonal Jacobians, and allows VAEs to perform independent mechanism analysis. The gap between ELBO and log-likelihood is therefore welcome, since it bears unanticipated benefits for nonlinear representation learning. In experiments on synthetic and image data, we show that VAEs uncover the true latent factors when the data generating process satisfies the IMA assumption.

4.1 Introduction

Latent Variable Models (LVMs) allow to effectively approximate a complex data distribution and to sample from it [67, 190]. Deep LVMs employ a neural network (the *decoder* or *generator*) to parameterize the conditional distribution of the observations given latent variables, which are typically assumed to be independent. However, Maximum Likelihood Estimation (MLE) of the model parameters is computationally intractable. In *VAEs* [191, 192], the exact log-likelihood is substituted with a tractable lower bound, the ELBO. This objective introduces an approximate posterior of the latents given the observations (the *encoder*) from a suitable variational distribution whose mean and covariance are parametrized by neural networks. The encoder is introduced to efficiently train a deep LVM: however, it is not explicitly designed to extract useful representations [68, 193].

[191]: Kingma et al. (2014), 'Auto-Encoding Variational Bayes'
[192]: Rezende et al. (2014), 'Stochastic Backpropagation and Approximate Inference in Deep Generative Models'

Nonetheless, VAEs and their variants are widely used in representation learning [194, 195], where they often recover semantically meaningful representations [196–199]. Our understanding of this empirical success is still incomplete, since the theory reviewed in § 2.4.1 shows that (deep) LVMs with independent latents are nonidentifiable from i.i.d. data: different models fitting the data equally well may yield arbitrarily different representations, thus rendering the recovery of a ground truth generative model impossible. While specific model constraints [70, 183, 200, 201] can help identifiability, and one such constraint was discussed in Chapter 3, the mechanism through which the ELBO may enforce a useful inductive bias remains unclear, despite recent efforts [199, 202–205].

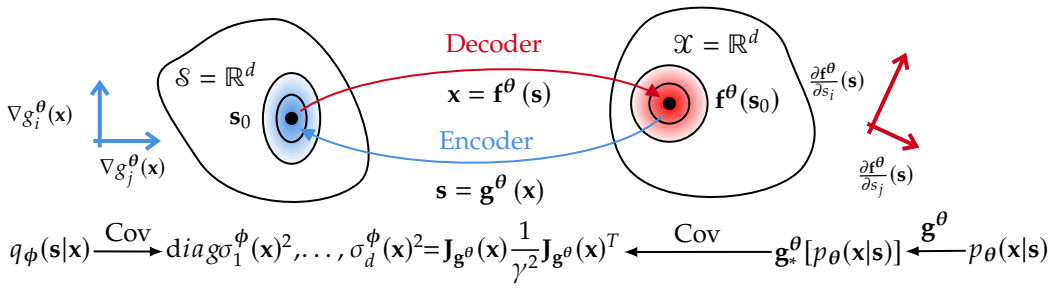


Figure 4.1: Modeling choices in VAEs promote Independent Mechanism Analysis. We assume a Gaussian VAE (4.3), and prove that in the near-deterministic regime the mean encoder approximatetely inverts the mean decoder, $\mathbf{g}^{\boldsymbol{\theta}} \approx \mathbf{f}^{\boldsymbol{\theta}-1}$ (*self-consistency*, Proposition 4.3.1). **Bottom:** Closing the gap requires matching the covariances of the variational (LHS, $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$) and the true posterior (RHS, approximated by $\mathbf{g}_*^{\boldsymbol{\theta}}[p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})]$, cf. § 4.3.2 for details). Under self-consistency, an encoder with diagonal covariance enforces a row-orthogonal encoder Jacobian $\mathbf{J}_{\mathbf{g}^{\boldsymbol{\theta}}}(\mathbf{x})$—or equivalently, a column-orthogonal decoder Jacobian $\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}}(\mathbf{s})$. Through this regularization, VAEs effectively performs independent mechanism analysis: the connection therefore elucidates unintended benefits of using the ELBO for representation learning.

In this work, we investigate the benefits of optimizing the ELBO for representation learning by analyzing VAEs in a *near-deterministic* limit for the conditional distribution parametrized by the nonlinear decoder. Our first result concerns the encoder's optimality in this regime. Previous works relied on the intuitive assumption that the encoder inverts the decoder in the optimum [203, 205, 206]; we formalize this *self-consistency* assumption and prove its validity for the optimal variational posterior in the near-deterministic nonlinear regime.

Using self-consistency, we show that the ELBO tends to a regularized log-likelihood—rather than to the exact one as conjectured in previous work [206]. The regularization term encourages column orthogonality of the decoder's Jacobian, and thus allows VAEs to perform IMA (Chapter 3). This generalizes previous findings based on linearizations or approximations of the ELBO [202, 203, 207], and allows us to characterize the gap with respect to the log-likelihood in the deterministic limit. Our results elucidate the gap between ELBO and exact log-likelihood as a possible mechanism through which the ELBO implements a useful inductive bias. We verify this by training VAEs in experiments on synthetic and image data, showing that they can recover the ground truth factors when the IMA assumptions are met.

[206]: Nielsen et al. (2020), 'SurVAE Flows: Surjections to Bridge the Gap between VAEs and Flows'

Structure and contributions of this Chapter.

- In (§ 4.3.1), we characterize and prove *self-consistency* of VAEs in the near-deterministic regime (i.e., when the decoder variance tends to zero), justifying its usage in previous works.
- In (§ 4.3.2), we show that under self-consistency, the ELBO converges to a regularized log-likelihood, and discuss its possible role as a useful inductive bias in representation learning;
- In (§ 4.4), we test the applicability of our theoretical results in experiments on synthetic and image data, and show that VAEs recover the true latent factors when the IMA assumptions are met.

4.2 Background

We will connect two unsupervised learning objectives: the ELBO in VAEs and the IMA-regularized log-likelihood. Both stem from LVMs with latent variables $\mathbf{s}$ distributed according to a *prior* $p_{\mathbf{s}}(\mathbf{s})$, and a mapping from $\mathbf{s}$ to observations $\mathbf{x}$ given by a conditional generative model $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})$.

Variational Autoencoders. Optimizing the data likelihood $p_{\boldsymbol{\theta}}(\mathbf{x})$ in deep LVMs—*i.e.*, finding decoder parameters $\boldsymbol{\theta}$ maximizing $\int p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})p_{\mathbf{s}}(\mathbf{s})d\mathbf{s}$—is intractable in general, so approximate objectives are required. Variational approximations [208] replace the true posterior $p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})$ by an approximate one, called the *variational posterior* $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$, which is a stochastic mapping $\mathbf{x} \mapsto \mathbf{s}$ with parameters $\boldsymbol{\phi}$. This allows to evaluate a tractable evidence lower bound (ELBO) [191, 192] of the model's log-likelihood that can be defined as

$$\text{ELBO}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})\right] - \text{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})||p_{\mathbf{s}}(\mathbf{s})\right]. \tag{4.1}$$

The two terms in (4.1) are sometimes interpreted as a reconstruction term measuring the sample quality of the decoder and a regularizer—the Kullback-Leibler Divergence (KL) between the prior and the encoder [209]. The variational approximation trades off computational efficiency with a difference with respect to the exact log-likelihood, which is expressed alternatively as (see [193, 209] and Appendix C.1)

$$\text{ELBO}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \log p_{\boldsymbol{\theta}}(\mathbf{x}) - \text{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})\right], \tag{4.2}$$

where the KL between variational and true posteriors characterizes the *gap*: if the variational family of $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$ does not include $p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})$, the ELBO will be strictly smaller than $\log p_{\boldsymbol{\theta}}(\mathbf{x})$.

VAEs [191] rely on the variational approximation in (4.1) to train deep LVMs where neural networks parametrize the *encoder* $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$ and the *decoder* $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})$. Common modeling choices include constraining the variational family of $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$ to be a factorized Gaussian with posterior means $\mu_k^{\boldsymbol{\phi}}(\mathbf{x})$ and variances $\sigma_k^{\boldsymbol{\phi}}(\mathbf{x})^2$ for each $s_k|\mathbf{x}$, and with a diagonal covariance $\boldsymbol{\Sigma}_{\mathbf{s}|\mathbf{x}}^{\boldsymbol{\phi}}$; and the decoder to a factorized Gaussian, conditional on $\mathbf{s}$, with mean $\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})$ and an isotropic covariance in n dimensions,

$$s_k|\mathbf{x} \sim \mathcal{N}(\mu_k^{\boldsymbol{\phi}}(\mathbf{x}), \sigma_k^{\boldsymbol{\phi}}(\mathbf{x})^2); \qquad \mathbf{x}|\mathbf{s} \sim \mathcal{N}\left(\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}), \gamma^{-2}\mathbf{I}_n\right). \tag{4.3}$$

The deterministic limit of VAEs. The stochasticity of VAEs makes it nontrivial to relate them to generative models with deterministic decoders such as Independent Component Analysis (see paragraph below), though postulating a deterministic regime (where the decoder precision γ^2 becomes infinite) is possible. Interestingly, [206] explored this deterministic limit and argued that *deterministic* VAEs optimize an exact log-likelihood, similar to normalizing flows [71, 210]. Normalizing flows model arbitrarily complex distributions with a simple base distribution $p_{\mathbf{s}}(\mathbf{s})$ and a juxtaposition of *deterministic and invertible* transformations $\mathbf{f}^{\boldsymbol{\theta}}$ through the change of variables[1]

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log p_{\mathbf{s}}(\mathbf{s}) - \log\left|\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}}(\mathbf{s})\right|. \tag{4.4}$$

1: note that the RHS of the change of variables is often written in terms of $\mathbf{x}$ and the inverse of $\mathbf{f}^{\boldsymbol{\theta}}$; cfr. (2.18).

The comparison is nontrivial, since VAEs contain an encoder and a decoder, whereas normalizing flows consist of a single architecture. [206] made this analogy by resorting to what we call a *self-consistency assumption*, stating that the VAE encoder inverts the decoder. We define self-consistency in the *near-deterministic* regime: as the decoder variance goes to zero, i.e. $\gamma \to +\infty$.

Definition 4.2.1 ((Near-deterministic) self-consistency) *For a fixed $\boldsymbol{\theta}$, assume that mean encoder $\mathbf{f}^{\boldsymbol{\theta}}$ is invertible with inverse $\mathbf{g}^{\boldsymbol{\theta}}$, and that a map associates each choice of decoder parameters and observation $(\boldsymbol{\theta}, \gamma, \mathbf{x})$ to an encoder parameter $(\boldsymbol{\theta}, \gamma, \mathbf{x}) \mapsto \widehat{\boldsymbol{\phi}}(\boldsymbol{\theta}, \gamma, \mathbf{x})$, we say the VAE is self-consistent whenever*

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x}) \to \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) \quad \text{and} \quad \boldsymbol{\sigma}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})^2 \to \mathbf{0} \text{ , as } \gamma \to +\infty \,. \tag{4.5}$$

The encoder parameter map reflects the choice of a particular encoder model for each $(\boldsymbol{\theta}, \gamma)$ pair:[2] in § 4.3.1, we study this problem by introducing and justifying a particular choice for $\widehat{\boldsymbol{\phi}}$ (see also § 4.5). This self-consistency assumption appears central to deterministic claims [203, 206], but has not yet been proven. In particular, [206] assume that taking the deterministic limit is well-behaved. However, VAEs' *near-deterministic* properties have not been investigated analytically.

2: both the ELBO and $\widehat{\boldsymbol{\phi}}$ depends on the decoder precision γ: we will omit this in the following for simplicity

4.3 Theory

Our theoretical analysis assumes that all the model's defining densities ($p_{\mathbf{s}}(\mathbf{s})$, $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$ and $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})$) are factorized. We also assume a Gaussian decoder, matching common modeling practice in VAEs.

Assumption 4.3.1 (Factorized VAE class with isotropic Gaussian decoder and log-concave prior) *We are given a fixed latent prior and three parameterized classes of $\mathbb{R}^n \to \mathbb{R}^n$ mappings: the mean decoder class $\boldsymbol{\theta} \mapsto \mathbf{f}^{\boldsymbol{\theta}}$, and the mean and standard deviation encoder classes, $\boldsymbol{\phi} \mapsto \boldsymbol{\mu}^{\boldsymbol{\phi}}$ and $\boldsymbol{\phi} \mapsto \boldsymbol{\sigma}^{\boldsymbol{\phi}}$ s.t.*

- *(i) $p_{\mathbf{s}}(\mathbf{s}) \sim \prod_k d(s_k)$, with d being smooth and fully supported on $\mathbb{R}$, having a bounded non-positive second-order, and bounded third-order logarithmic derivatives;*
- *(ii) the encoder and decoder are of the form in* (4.3)*, with isotropic decoder covariance $1/\gamma^2 \mathbf{I}_n$;*
- *(iii) the variational mean and variance encoder classes are universal approximators;*
- *(iv) for all $\boldsymbol{\theta}$, $\mathbf{f}^{\boldsymbol{\theta}} : \mathbb{R}^n \to \mathbb{R}^n$ is a bijection with inverse $\mathbf{g}^{\boldsymbol{\theta}}$, and both are C^2 with bounded first and second order derivatives.*

Crucially, *both the mean encoder and the mean decoder can be nonlinear*. Moreover, the family of log-concave priors contains the commonly-used Gaussian distribution as a special case. We study the *near-deterministic decoder* regime of such models, where $\gamma \to +\infty$. This regime is expected to model data generating processes with vanishing observation noise well—in line with the typical ICA setting—and is commonly considered in theoretical analyses of VAEs, *e.g.* in [206] (which additionally assumes

quasi-deterministic encoders), and in [203, 207]. Unlike [206], we consider a large but finite γ, not *at* the limit $\gamma = \infty$, where the decoder is fully deterministic. In fact, for any large but finite γ, the objective is well-behaved and amenable to theoretical analysis, while the KL-divergence is undefined in the deterministic setting. The requirement in assumption *(iv)* deviates from common practice in VAEs—where observations are typically higher-dimensional—but it allows to connect VAEs and exact likelihood methods such as normalizing flows [206] (see also § 4.5).

Due to considering $\gamma \to +\infty$, results are stated in the following "big-O" notation for an integer p:

$$f(\mathbf{x}, \gamma) = g(\mathbf{x}, \gamma) + O_{\gamma \to +\infty}(1/\gamma^p) \iff \gamma^p \|f(\mathbf{x}, \gamma) - g(\mathbf{x}, \gamma)\| \text{ is bounded as } \gamma \to +\infty\,.$$

4.3.1 Self-Consistency

In this section, we will prove a *self-consistency* result in the near-deterministic regime. This rests on characterizing optimal variational posteriors (i.e., those minimizing the ELBO gap with respect to the likelihood) for a *particular point* $\mathbf{x}$ and *fixed decoder parameters* $\boldsymbol{\theta}$. Based on (4.2), any associated optimal choice of encoder parameters satisfies

$$\widehat{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\theta}) \in \arg\max_{\boldsymbol{\phi}} \text{ELBO}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \arg\min_{\boldsymbol{\phi}} \text{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x}) || p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})\right]\,. \tag{4.6}$$

We call *self-consistent* ELBO the resulting achieved value, denoted

$$\text{ELBO}^*(\mathbf{x}; \boldsymbol{\theta}) = \text{ELBO}(\mathbf{x}; \boldsymbol{\theta}, \widehat{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\theta}))\,. \tag{4.7}$$

The expression in (4.6) corresponds to a problem of *information projection* [67, 211] of $p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})$ onto the set of factorized Gaussian distributions. This means that given a variational family, we search for the optimal $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$ to minimize the KL to $p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})$. While such information projection problems are well studied for closed convex sets where they yield a unique minimizer [212], the set projected onto in our case is not convex (convex combinations of arbitrary Gaussians are not Gaussian), making this problem of independent interest. After establishing upper and lower bounds on the KL divergence (exposed in Prop. C.3.1-C.3.2 in Appendix C.3.2), we obtain the following self-consistency result.

Proposition 4.3.1 *[Self-consistency of near-deterministic VAEs] Under Assumption 4.3.1, for all* $\mathbf{x}$, $\boldsymbol{\theta}$, *as* $\gamma \to +\infty$, *there exists at least one global minimum solution of (4.6). These solutions satisfy*

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x}) = \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) + O(1/\gamma) \quad \textit{and} \quad \sigma_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})^2 = O(1/\gamma^2)\,, \textit{ for all } k\,. \tag{4.8}$$

Proposition 4.3.1 states that minimizing the ELBO gap (equivalently, maximizing the ELBO) with respect to the encoder parameters $\boldsymbol{\phi}$ implies in the limit of large γ that the encoder's mean $\boldsymbol{\mu}^{\boldsymbol{\phi}}(\mathbf{x})$ tends to $\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})$, the image of $\mathbf{x}$ by the *inverse* decoder. We can interpret this as the decoder "inverting" the encoder. Additionally, the variances of the encoder will converge to zero, in line with empirical observations of practitioners. This is also stated as one property of the *polarized regime* (cf. Definition 1 in [202]).

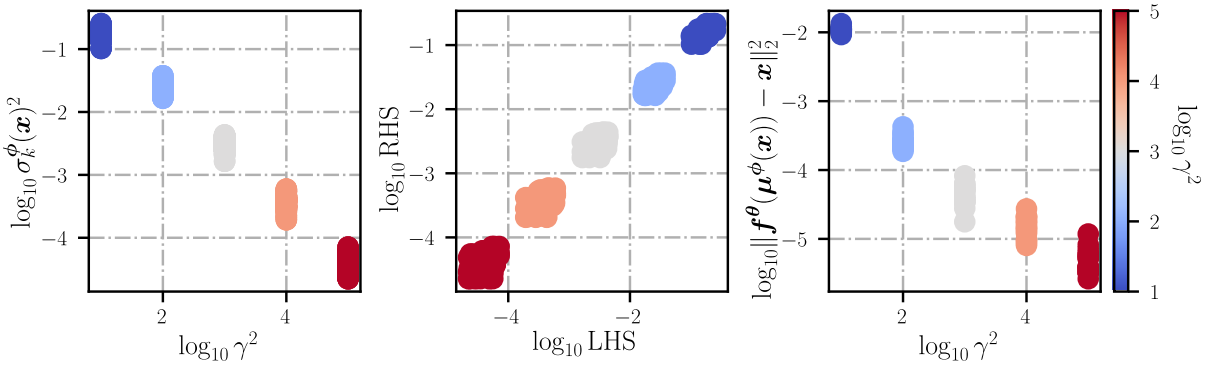


Figure 4.2: Self-consistency (Proposition 4.3.1) in VAE training, on a log-log plot, cf. 4.4.1 for details. **Left**: convergence of $\sigma_k^{\widehat{\phi}}(\mathbf{x})^2$ to 0; **Center:** connecting $\sigma_k^{\widehat{\phi}}(\mathbf{x})^2$, γ^2, and the column norms of the decoder Jacobian via LHS and RHS of (4.11); **Right:** convergence of $\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})$ to $\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})$

Let us now consider the relevance of this result for training VAEs, *i.e.*, maximizing the expectation of the ELBO for an observed distribution $p(\mathbf{x})$. While maximization *only* with respect to $\boldsymbol{\phi}$ in (4.6) does not match common practice—which is learning $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ *jointly*—it models this process in the limit of large-capacity encoders. Indeed, in this case, (4.6) can be solved for each $\mathbf{x}$ as a separate learning problem, which entails that the following inequality is satisfied for any parameter choice

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\mathrm{ELBO}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\phi})\right] &= \int p(\mathbf{x})\mathrm{ELBO}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\phi})d\mathbf{x} \\ &\leq \int p(\mathbf{x})\mathrm{ELBO}(\mathbf{x};\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}(\mathbf{x},\boldsymbol{\theta}))d\mathbf{x} =: \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\mathrm{ELBO}^*(\mathbf{x};\boldsymbol{\theta})\right] . \quad (4.9)\end{aligned}$$

The joint optimization of encoder and decoder parameters thus reduces to optimizing the subset of pairs $(\boldsymbol{\theta}, \widehat{\boldsymbol{\phi}}(\mathbf{x},\boldsymbol{\theta}))$, and is equivalent to optimizing the expected self-consistent ELBO, that is

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\mathrm{ELBO}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\phi})\right] \iff \min_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\mathrm{ELBO}^*(\mathbf{x};\boldsymbol{\theta})\right] \quad (4.10)$$

This problem reduction is aligned with the original purpose of the ELBO: building a tractable but optimal likelihood approximation. Namely, (i) ELBO* depends on the same parameters as the likelihood ($\mathbf{x}$, γ and $\boldsymbol{\theta}$), (ii) its gap KL $\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})\right]$ is minimal. Clearly, the problem reduction of (4.10) is more informative only at the critical points of the VAE loss, since it allows us to compare the optimality of different decoders and Proposition 4.3.1 allows us to study this question for near-deterministic decoders.

4.3.2 Self-Consistent ELBO, IMA-Regularized Log-Likelihood and Identifiability of VAEs

We want to investigate how the choice of $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$ and $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})$ implicitly regularizes the Jacobians of their means $\boldsymbol{\mu}^{\boldsymbol{\phi}}(\mathbf{x})$ and $\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})$ in the near-deterministic regime. Exploiting self-consistency, we are able to precisely characterize how this happens: we formalize this in Thm. 4.3.2.

Theorem 4.3.2 *[VAEs with a near-deterministic decoder approximate the* IMA *objective] Under Assumption 4.3.1, the variational posterior satisfies*

$$\sigma_k^{\widehat{\phi}}(\mathbf{x})^2 = \left(-\frac{d^2 \log p_0}{ds_k^2}(g_k^{\boldsymbol{\theta}}(\mathbf{x})) + \gamma^2 \left\| \left[\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}} \left(\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) \right) \right]_{:k} \right\|^2 \right)^{-1} + O(1/\gamma^3)\,, \tag{4.11}$$

and the self-consistent ELBO (4.7) *approximates the* IMA*-regularized log-likelihood* (3.5)*:*

$$\text{ELBO}^*(\mathbf{x};\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\mathbf{x}) - c_{\text{IMA}}(\mathbf{f}^{\boldsymbol{\theta}}, \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})) + O_{\gamma\to\infty}(1/\gamma^2)\,. \tag{4.12}$$

Proof is in Appendix C.2. Below, we provide a qualitative argument on the interplay between distributional assumptions in the VAE and implicit constraints on the decoder's Jacobian and its inverse.

Modeling assumptions implicitly regularize the mean decoder class $\mathbf{f}^{\boldsymbol{\theta}}$ under self-consistency. In the near deterministic regime, $p_{\boldsymbol{\theta}}(\mathbf{x})$ gets close to the pushforward distribution of the prior by the mean decoder $\mathbf{f}^{\boldsymbol{\theta}}_*[p_{\mathbf{s}}(\mathbf{s})]$, which can be used to show that the true posterior $p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})p_{\mathbf{s}}(\mathbf{s})/p_{\boldsymbol{\theta}}(\mathbf{x})$ is approximately the pushforward through the inverse mean decoder $\mathbf{g}^{\boldsymbol{\theta}}_*[p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})]$ (see Appendix C.1 for more details). If we select a given latent $\mathbf{s}_0$ and denote its image by $\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0)$, then we can locally linearize $\mathbf{g}^{\boldsymbol{\theta}}$ by its Jacobian $\mathbf{J}_{\mathbf{g}^\theta} = \mathbf{J}_{\mathbf{g}^\theta}(\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0))$, yielding a Gaussian for the pushforward distribution $\mathbf{g}^{\boldsymbol{\theta}}_*[p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})]$ with covariance $1/\gamma^2 \mathbf{J}_{\mathbf{g}^\theta}\mathbf{J}_{\mathbf{g}^\theta}^T$. As the sufficient statistics of a Gaussian are given by its mean and covariance, the structure of the posterior covariance $\boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\mathbf{s}|\mathbf{x}}$ (which is by design diagonal, cf. (4.3)) is crucial for minimizing the gap in (4.2). Practically, this implies that in the zero gap limit, the covariances of $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$ and $p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})$ should match, *i.e.*, $1/\gamma^2 \mathbf{J}_{\mathbf{g}^\theta}\mathbf{J}_{\mathbf{g}^\theta}^T$ will be diagonal with entries $\sigma_k^{\boldsymbol{\phi}}(\mathbf{x})^2$ and therefore $\mathbf{J}_{\mathbf{g}^\theta}$ has orthogonal rows. We can express the decoder Jacobian via the inverse function theorem as $\mathbf{J}_{\mathbf{f}^\theta}(\mathbf{s}_0) = \mathbf{J}_{\mathbf{g}^\theta}(\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0))^{-1}$. As the inverse of a row-orthogonal matrix has orthogonal columns, $\mathbf{f}^{\boldsymbol{\theta}}$ satisfies the IMA principle. Additionally, we can relate the variational posterior's variances to the column-norms of $\mathbf{J}_{\mathbf{f}^\theta}$ as $\sigma_k^{\boldsymbol{\phi}}(\mathbf{x})^2 = 1/\gamma^2 \| [\mathbf{J}_{\mathbf{f}^\theta}(\mathbf{s}_0)]_{:k} \|^{-2}$. The self-consistent ELBO therefore converges to the IMA-regularized log-likelihood (3.5).

Our argument indicates that minimizing the gap between the ELBO and the log-likelihood encourages column-orthogonality in $\mathbf{J}_{\mathbf{f}^\theta}$ by matching the covariances of $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$ and $\mathbf{g}^{\boldsymbol{\theta}}_*[p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})]$. When $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})$, the gap is closed; this is only possible if the decoder is in the IMA class, for which c_{IMA} vanishes and the ELBO *tends to an exact log-likelihood*. To the best of our knowledge, we are the first to prove this for nonlinear functions, extending related work for linear VAEs [207].

Implications for identifiability of VAEs. While previous works argued that the VAE objective favors decoders with a column-orthogonal Jacobian [202, 203], they did not exactly characterize how: our result shows that the self-consistent ELBO tends to a regularized log-likelihood, where the regularization term c_{IMA} explicitly enforces this (soft) constraint. Thus, it possibly explains why VAEs are successful in learning disentangled representations: namely, the IMA function class provably rules out certain spurious solutions for nonlinear Independent Component Analysis

(ICA) (§ 2.4.1), and the IMA-regularized log-likelihood was empirically shown to be beneficial in recovering the true latent factors when the data generating process satisfies the IMA principle (Chapter 3). Thus, we speak about *embracing the gap*, as its functional form equips VAEs with a useful inductive bias.

In the following, we empirically corroborate that VAEs: 1) recover the ground truth sources when the mixing satisfies IMA, and thereby 2) achieve unsupervised disentanglement.

4.4 Experiments

Our experiments serve three purposes: 1) demonstrating that self-consistency holds in practice (§ 4.4.1); 2) showing the relationship of the self-consistent ELBO*, the IMA-regularized and unregularized log-likelihood objectives (§ 4.4.2); and 3) providing empirical evidence that the connection to the IMA function class in VAEs can lead to success in learning disentangled representations (§ 4.4.3). More details are provided in Appendix C.6.

4.4.1 Self-Consistency in Practical Conditions

Experimental setup. We use a 3-layer MultiLayer Perceptron (MLP) with smooth Leaky ReLU nonlinearities [86] and orthogonal weight matrices—which intentionally does not belong to the IMA class, as our results are more general. The 60,000 source samples are drawn from a standard normal distribution and fed into a VAE composed of a 3-layer MLP encoder and decoder with a Gaussian prior. We use 20 seeds for each $\gamma^2 \in \{1e1; 1e2; 1e3; 1e4; 1e5\}$.

Results. Fig. 4.2 summarizes our results, featuring the *logarithms* on each axes. The **left** plot shows that the posterior variances $\sigma_k^{\phi}(\mathbf{x})^2$ converge to zero with a $1/\gamma^2$ rate, as predicted by (4.8). The **center** plot shows that the expression for $\sigma_k^{\phi}(\mathbf{x})^2$ corresponds to (4.11) in the optimum of the ELBO by comparing both sides of the equation. The **right** plot shows approximate convergence of the mean encodings $\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})$ to $\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})$ with a $1/\gamma$ rate (see § 4.5). As $\mathbf{f}^{\boldsymbol{\theta}}$ is not guaranteed to be invertible, we use instead the *optimal* encoder and decoder parameters to compare $\mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x}))$ to $\mathbf{x}$.

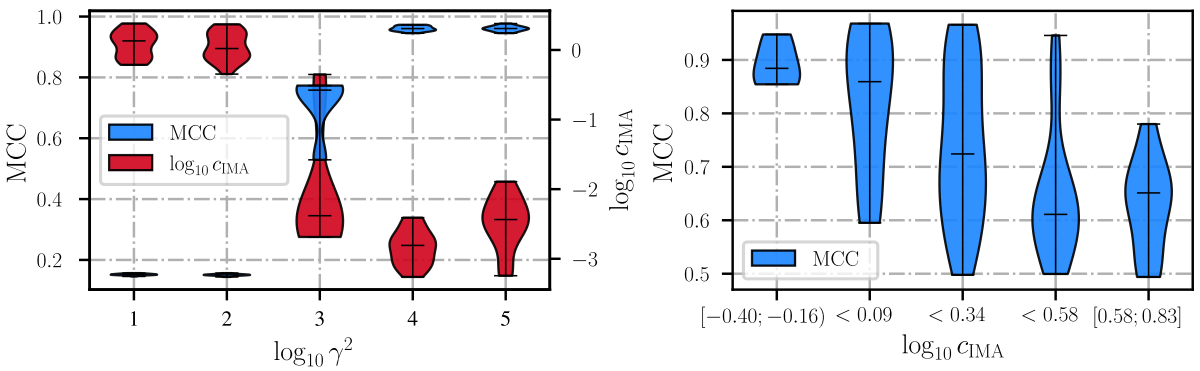


Figure 4.3: Left: c_{IMA} and Mean Correlation Coefficient (MCC) for 3-dimensional Möbius mixings **Right:** MCC depending on the *volume-preserving linear map's* c_{IMA} (γ^2 = 1e5)

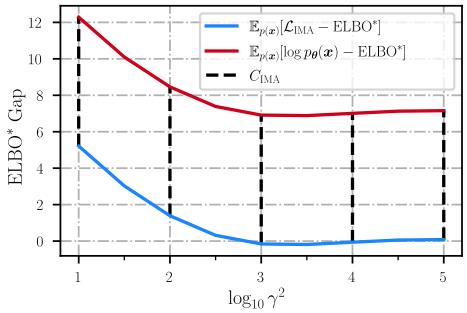


Figure 4.4: Comparison of the ELBO*, the IMA-regularized and unregularized log-likelihoods over different γ^2. Error bars are omitted as they are orders of magnitudes smaller

4.4.2 Relationship between ELBO*, IMA-Regularized, and Unregularized Log-Likelihoods

Experimental setup. We use an MLP $\mathbf{f}^{\boldsymbol{\theta}}$ with square upper-triangular weight matrices and invertible element-wise nonlinearities to construct a mixing not in the IMA class [63] and fix the VAE decoder to the ground truth such that (4.4) gives the true data log-likelihood. This way, we ensure that the unregularized and IMA-regularized log-likelihoods differ and make the claim of [206] comparable to ours. With a fixed decoder, the ELBO* depends only on $\boldsymbol{\phi}$, therefore we only train the encoder with γ^2 values from [1e1; 1e5] (5 seeds each).
Results. Fig. 4.4 compares the difference of the estimate of ELBO* and the unregularized/IMA-regularized log-likelihoods after convergence over the whole dataset. As the decoder and the data are fixed, $\log p_{\boldsymbol{\theta}}(\mathbf{x})$ and C_{IMA} will not change during training, only ELBO* does. The figure shows that as $\gamma \to +\infty$, ELBO* approaches $\mathcal{L}_{\text{IMA}}(\mathbf{f}^{\boldsymbol{\theta}}, \mathbf{s})$, as predicted by Thm. 4.3.2, and not $\log p_{\boldsymbol{\theta}}(\mathbf{x})$, as stated in [206]—the difference is C_{IMA}.

4.4.3 Connecting the IMA Principle, γ^2, and Disentanglement

Experimental setup (synthetic). We use 3-dimensional conformal mixings (*i.e.*, the Möbius transform [176]) from the IMA class with *uniform* ground-truth and prior distributions. Our results quantify the relationship of the decoder Jacobian's IMA-contrast and identifiability with MCC [64] and show how this translates to disentanglement—we note that MCC was already used to quantify disentanglement [88, 213]. To determine whether a mixing from the IMA class is beneficial for disentanglement, we apply a volume-preserving linear map after the Möbius transform (using 100 seeds) to make $c_{\text{IMA}} \neq 0$. Other parameters are the same as in § 4.4.1, with the exception of picking the best $\gamma^2 = 1\text{e}5$.
Results (synthetic). The **left** of Fig. 4.3 empirically demonstrates the benefits of optimizing the IMA-regularized log-likelihood. By increasing γ^2, MCC increases, while c_{IMA} decreases, suggesting that VAEs in the near-deterministic regime encourage disentanglement by enforcing the IMA principle. The **right** plot shows that when the mixing is outside the IMA class, MCC decreases, corroborating the benefits of such mixings for disentanglement.
Experimental setup (image). We train a VAE (not β-VAE) with a factorized Gaussian posterior and Beta prior on a Sprites image dataset generated using the spriteworld renderer [214] with a Beta ground truth

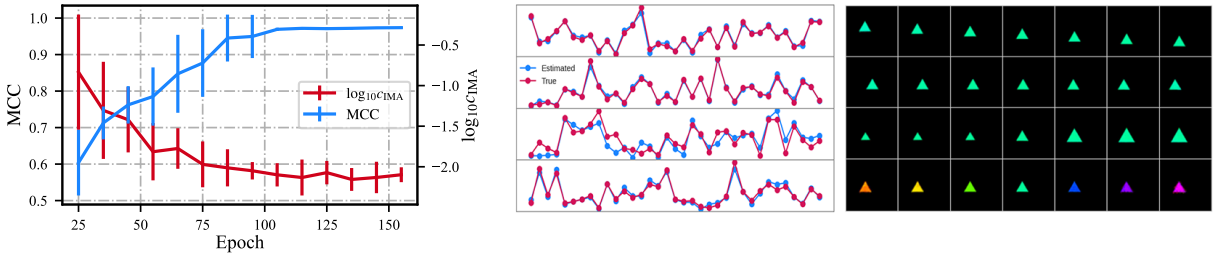


Figure 4.5: Left: c_{IMA} and MCC for Sprites [214] during training ($\gamma^2 = 1$); **Center:** true and estimated latent factors for the best trained VAE on Sprites; **Right:** the corresponding latent interpolations and MCC values (from top to bottom): y- (0.989), x-position (0.996), scale (0.933), and color (0.989)

distribution. Similar to [215], we use four latent factors, namely, *x- and y-position, color and size*, and omit factors that can be problematic, such as shape (as it is discrete) and rotation (due to symmetries) [202, 213]. Our choice is motivated by [201, 216] showing that the data-generating process presumably is in the IMA class.

Results (image). The **left** of Fig. 4.5 indicates that VAEs can learn the true latent factors and MCC is *anticorrelated* with c_{IMA}, reinforcing the hypothesis that the data-generating process belongs to the IMA class. The **center** plot compares estimated and true latent factors from the best model (scaling and permutation indeterminacies are removed), whereas the **right** plot shows the corresponding latent interpolations—thus, connecting identifiability (measured by MCC) to disentanglement.

4.5 Limitations

The near-deterministic regime. Our theory relies on $\gamma \to +\infty$; this is the regime where posterior collapse may be avoided [207], and where calculating the reconstruction loss may be possible even without sampling [203]. However, in practice it may be unclear when γ^2 is large enough. This seems to be problem-dependent [202, 207], and possibly tied to the covariance of the observations [217, 218]. Moreover, large values of γ^2 may be harder to optimize due to an exploding reconstruction term in (4.1). This may be one explanation for the slight deviation of Fig. 4.2, right from our theory's predictions: while convergence of $\boldsymbol{\mu}^{\boldsymbol{\phi}}(\mathbf{x})$ to $\mathbf{g}^{\boldsymbol{\theta}}$ matches the prediction in Proposition 4.3.1, its rate is not precisely the one predicted for the self-consistent ELBO (4.7). Another cause could be the encoder's finite capacity. Nonetheless, we have experimentally shown that for realistic hyperparameters, VAEs' behavior matches the predictions of our theory for the near-deterministic regime.

Dimensionality. The setting in § 4.3 requires same dimensionality of $\mathbf{s}$ and $\mathbf{x}$. This matches most work on normalizing flows [71] and connecting these to VAEs [206], as well as most of the work on nonlinear ICA [65, 66, 78] (but see, e.g., [62]). Nonetheless, we could empirically verify that the predictions of our theory also hold for dimensionality reduction (§ 4.4.3). Extending our theory to this setting could rely on, *e.g.*, [219, 220] and is left for future work.

The ELBO, the self-consistent ELBO, and amortized inference. There are in principle multiple ways to obtain self-consistency (Defn. 4.2.1). Notably, one could simply force the variational mean and variance encoder

maps to behave this way; unlike [203], we model the actual behavior of VAEs trained under ELBO maximization, and obtain self-consistency as a result. For this, we assume that the optimal encoder, which minimizes the gap between ELBO and log-likelihood, can be learnt. This is not guaranteed in general, since it requires universal approximation capability of the encoder. On the other hand, (4.7) requires *unamortized* inference to introduce ELBO*, which does not depend on $\boldsymbol{\phi}$. As in practice amortized inference may be used to efficiently estimate a single set of $\boldsymbol{\phi}$ for all $\mathbf{x}$ [221], it can lead to a suboptimal gap to the log-likelihood and discrepancies with our theoretical predictions.

4.6 Discussion

On disentanglement in unsupervised VAEs. It is widely believed that unsupervised VAEs cannot learn disentangled representations [62, 72], motivating work on models with, *e.g.*, conditional priors [62] or sparse decoding [32]. We show that under certain assumptions, ELBO optimization can implement useful inductive biases for representation learning, yielding disentangled representations in unsupervised VAEs. However, while our results are formulated for VAEs, some of the most successful models at disentanglement are modifications thereof—e.g., β-VAEs [194, 199], with an additional parameter β multiplying the KL in (4.1). Self-consistency is harder to prove for β-VAEs, as they deviate from the information projection setting considered in § 4.3.1. However, convergence of the loss in Thm. 4.3.2, mainly relies on self-consistency to hold. We thus conjecture that our results would be applicable to β-VAEs as well, as long as they are in a regime that satisfies self-consistency. Investigating this question is left to future work. Overall, we stress that we uncover *one* possible mechanism through which VAEs may achieve disentanglement. By connecting to IMA [63], we discuss implications on recovering the ground truth under suitable assumptions, extending uniqueness results presented in [203]. We speculate that our success in disentanglement is probably due to selecting data sets where the mixing is in the IMA class (cf. [201, 216]), which presumably was not the case in [72].

Characterizing the ELBO gap for nonlinear models. Thm. 4.3.2 characterizes the gap between ELBO and true log-likelihood for nonlinear VAEs, and extends the linear analysis of [207]; we also empirically characterize the gap in the deterministic limit in § 4.4.2. An unanticipated consequence of this result is that—consistent with [207]—VAEs optimize the IMA-regularized log-likelihood in the near-deterministic limit, and not the unregularized one, as stated in [206].

Extensions to related work. Several papers discuss the (near-)deterministic regime [202, 203, 206]. For example, [206] postulate a deterministic VAE with the encoder inverting the decoder. Also [203] work in that regime, but without justifying the relationship between the encoder and decoder. Although they show that the choice of $p_{\mathbf{s}}(\mathbf{s})$ and $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$ influences uniqueness (by, *e.g.*, ruling out rotations), this does not imply recovering the true latents. Our approach formalizes (Defn. 4.2.1), proves (Proposition 4.3.1), and demonstrates the practical feasibility of (§ 4.4) the near-deterministic regime. To the best of our knowledge, all previous work relied on the linear case [207] or a (linear) approximation and the

evaluation of the ELBO *around a point* to show the inductive bias on the decoder Jacobian. However, our main result (Thm. 4.3.2) yields a nonlinear equation where the decoder Jacobian can be evaluated at *any point* and is equipped with a convergence bound. Moreover, the consistency of VAE estimation for identifiable models [62] requires guarantees on $q_\phi(\mathbf{s}|\mathbf{x})$; our result helps proving these. We discuss extended connections to the literature in Appendix C.4 and Appendix C.5.

Covariance structure and IMA. We have shown that specific choices for encoder and decoder covariances regularize the decoder Jacobian. Assuming factorized $q_\phi(\mathbf{s}|\mathbf{x})$ and isotropic $p_\theta(\mathbf{x}|\mathbf{s})$, IMA holds only for the *decoder*; since in the other direction the pushforward of $q_\phi(\mathbf{s}|\mathbf{x})$ through $\mathbf{f}^\theta$ has covariance $\mathbf{J}_{\mathbf{f}^\theta}(\mathbf{s})\,\boldsymbol{\Sigma}^{\phi}_{\mathbf{s}|\mathbf{x}}\mathbf{J}_{\mathbf{f}^\theta}(\mathbf{s})^T$, which cannot be simplified in the general case. Additionally assuming an isotropic encoder would make IMA hold in both directions and would yield conformal Jacobians (as both $\mathbf{J}_{\mathbf{f}^\theta}(\mathbf{s})$ and $\mathbf{J}_{\mathbf{g}^\theta}(\mathbf{x})$ need to be column-orthogonal). On the other hand, if the observation model is non-isotropic, IMA would only hold for the *encoder* Jacobian. Covariance structure has an intuitive meaning: an isotropic $p_\theta(\mathbf{x}|\mathbf{s})$ covariance equals having i.i.d. noise for each x_i. For data recorded by the same device (*e.g.*, pixels) this may be reasonable, but multi-view settings with non-isotropic noise could require different modeling choices.

Conjecture for VAEs with isotropic encoder and decoder covariances. Following our intuition (Fig. 4.1), we suspect that an isotropic encoder entails a decoder with conformal Jacobian (*i.e.*, the product of a scalar field and an orthogonal matrix). This is an interesting constraint, especially as there is growing evidence that nonlinear ICA with conformal mixings may be identifiable: the two-dimensional case was proven by [70] and IMA was shown to rule out certain spurious solutions for conformal mixings [63]. Based on these, we conjecture:

Conjecture 4.6.1 (Unsupervised VAEs with isotropic $q_\phi(\mathbf{s}|\mathbf{x})$ and $p_\theta(\mathbf{x}|\mathbf{s})$ are identifiable) *When $q_\phi(\mathbf{s}|\mathbf{x})$ and $p_\theta(\mathbf{x}|\mathbf{s})$ have (possibly data-dependent) isotropic covariances, i.e., $\boldsymbol{\Sigma}_{\mathbf{s}|\mathbf{x}} = \alpha(\mathbf{x})\mathbf{I}_n$ and $\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{s}} = \beta(\mathbf{s})\mathbf{I}_n$, and self-consistency holds, then unsupervised VAEs are identifiable.*

Conclusion. We provide a theoretical justification for the widely-used self-consistency assumption in the near-deterministic regime of small decoder variance. Using this result, we show that the self-consistent ELBO converges to the IMA-regularized log-likelihood, and not to the unregularized one. Thus, we can characterize the gap between ELBO and true log-likelihood and reason about its possible role as an inductive bias for representation learning in nonlinear VAEs. We characterize a set of assumptions under which unsupervised VAEs can be expected to disentangle and we demonstrate this behavior in experiments on synthetic and image data.

Relative Gradient Optimization of the Jacobian Term in Unsupervised Deep Learning

5

Learning expressive probabilistic models correctly describing the data is a ubiquitous problem in machine learning. A popular approach for solving it is mapping the observations into a representation space with a simple joint distribution, which can typically be written as a product of its marginals — thus drawing a connection with nonlinear independent component analysis. Deep density models have been widely used for this task. Unfortunately, as we saw in § 2.2.3, their maximum likelihood based training requires estimating the log-determinant of the Jacobian and is computationally expensive, thus imposing a trade-off between computation and expressive power. In this work, we propose a new approach for exact training of such neural networks. Based on relative gradients, we exploit the matrix structure of neural network parameters to compute updates efficiently even in high-dimensional spaces; the computational cost of the training is quadratic in the input size, in contrast with the cubic scaling of naive approaches. This allows fast training with objective functions involving the log-determinant of the Jacobian, without imposing constraints on its structure, in stark contrast to autoregressive normalizing flows.

5.1 Introduction

Many problems of machine learning and statistics involve learning invertible transformations of complex, multimodal probability distributions into simple ones. One example is density estimation through latent variable models under a specified base distribution [222], which can also have applications in data generation [97, 223, 224] and variational inference [210]. Another example is nonlinear ICA, where we want to extract simple, disentangled features out of the observed data.

One approach to learn such transformations, introduced in [225] in the context of density estimation, is to represent them as a composition of simple maps, the sequential application of which enables high expressivity and a large class of representable transformations. Deep neural networks parameterize functions of multivariate variables as modular sequences of linear transformations and component-wise activation functions, thus providing a natural framework for implementing that idea, as already proposed in [226].

[225]: Tabak et al. (2013), 'A family of nonparametric density estimation algorithms'

Unfortunately, however, typical strategies employed in neural networks training do not scale well for objective functions like the aforementioned ones; in fact, through the change of variable formula, the logarithm of the absolute value of the determinant of the Jacobian appears in the objective. Its exact computation, let alone its optimization, quickly gets prohibitively computationally demanding as the data dimensionality grows.

A large part of the research on deep density estimation, generally referred to under the term *autoregressive normalizing flows*, has therefore been dedicated to considering a restricted class of transformations such that the

computation of the Jacobian term is trivial [97, 179, 210, 227–229], thus imposing a tradeoff between computation and expressive power. While such models can approximate arbitrary probability distributions, the extracted features are strongly restricted based on the imposed triangular structure, which prevents the system from learning a properly disentangled representation. Other strategies involve the optimization of an approximation of the exact objective [230], and continuous-time analogs of normalizing flows for which the likelihood (or some approximation thereof) can be computed using relatively cheap operations [224, 231].

In this chapter, we provide an efficient way to optimize the exact maximum likelihood objective for deep density estimation as well as for learning disentangled representations by latent variable models. We consider a nonlinear, invertible transformation from the observed to the latent space which is parameterized through fully connected neural networks. The weight matrices are merely constrained to be invertible. The starting point is that the parameters of the linear transformations are matrices; this allows us to exploit properties of the Riemannian geometry of matrix spaces to derive parameter updates in terms of the relative gradient, which was originally introduced as the natural gradient in the context of linear ICA [232, 233], and which can be feasibly computed. We show how this can be integrated with the usual backpropagation employed to compute gradients in neural network training, yielding an overall efficient way to optimize the Jacobian term in neural networks. This is a general optimization approach which is potentially useful for any objective involving such a Jacobian term, and is likely to find many applications in diverse areas of probabilistic modelling, for example in the context of Bayesian active learning for the computation of the information gain score [234], or for fitting the reverse Kullback-Leibler divergence in variational inference [235, 236].

The computational cost of our proposed optimization procedure is quadratic in the input size — essentially the same as ordinary backpropagation — which is in stark contrast with the cubic scaling of the naive way of optimizing via automatic differentiation. The joint asymptotic scaling of forward and backward pass as a function of the input size is therefore the same that aforementioned alternative methods achieve by imposing strong restrictions on the neural network structure [210] and thus on the class of functions they can represent. In contrast, our approach allows to efficiently optimize the exact objective for neural networks with arbitrary Jacobians.

Structure and contributions of this Chapter. In § 5.2 and § 5.3 we review maximum likelihood estimation for latent variable models, backpropagation and the Jacobian term for neural networks, and discuss the complexity of the naive approaches for optimizing the Jacobian term. Then in § 5.4 we discuss the relative gradient, and show how it can be integrated with backpropagation resulting in an efficient procedure. We verify empirically the computational speedup our method provides in § 5.5.

5.2 Background

5.2.1 Maximum Likelihood for Latent Variable Models

Consider a generative model of the form

$$\mathbf{x} = \mathbf{f}(\mathbf{s}) \tag{5.1}$$

where $\mathbf{s} \in \mathbb{R}^n$ is the latent variable, $\mathbf{x} \in \mathbb{R}^n$ represents the observed variable and $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ is a deterministic and invertible function, which we refer to as *forward* transformation. Under the model specified above, the log-likelihood of a single datapoint $\mathbf{x}$ can be written as

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \log p_{\mathbf{s}}(\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{x})) + \log |\det \mathbf{J}_{\mathbf{g}_\theta}(\mathbf{x})| \,, \tag{5.2}$$

where $\mathbf{g}_{\boldsymbol{\theta}}$ is some representation with parameters $\boldsymbol{\theta}$ of the *inverse* transformation[1] of $\mathbf{f}$; $\mathbf{J}_{\mathbf{g}_\theta}(\mathbf{x}) \in \mathbb{R}^{n \times n}$ its Jacobian computed at the point $\mathbf{x}$, whose elements are the partial derivatives $[\mathbf{J}_{\mathbf{g}_\theta}(\mathbf{x})]_{ij} = \partial g^i_{\boldsymbol{\theta}}(\mathbf{x}) / \partial x^j$; and $p_{\boldsymbol{\theta}}$ and $p_{\mathbf{s}}$ denote, respectively, the probability density functions of $\mathbf{x}$ and of the latent variable $\mathbf{s}$ under the specified model. In many cases, it is additionally assumed that the distribution of the latent variable is sufficiently simple; for example, that it factorizes in its components, thereby recovering an objective with the same functional form as (2.18),

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_i \log p_i(\mathbf{g}^i_{\boldsymbol{\theta}}(\mathbf{x})) + \log |\det \mathbf{J}_{\mathbf{g}_\theta}(\mathbf{x})| \,. \tag{5.3}$$

In this case, the problem can be interpreted as nonlinear independent component analysis (nonlinear ICA), and the components of $\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{x})$ are estimates of the original sources $\mathbf{s}$. In the identifiable auxiliary variable settings § 2.4.2, where the latent variables are not *unconditionally* independent, but rather *conditionally* independent given an additional, observed variable $\mathbf{u}$, the model likelihood can be written as

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u}) = \sum_i \log p_i(\mathbf{g}^i_{\boldsymbol{\theta}}(\mathbf{x})|\mathbf{u}) + \log |\det \mathbf{J}_{\mathbf{g}_\theta}(\mathbf{x})| \,. \tag{5.4}$$

Maximum likelihood estimation for the model parameters amounts to finding, through optimization, the parameters $\boldsymbol{\theta}^*$ such that the expectation of the likelihood given by the expression in equation (5.3) is maximized. For all practical purposes, the expectation will be substituted with the sample average. Specifically, for optimization purposes, we will be interested in the computation of a gradient of such term on mini-batches of one or few datapoints, such that stochastic gradient descent can be employed.

1: The forward transformation could also be parameterized, but here we only explicitly parameterize its inverse.

5.2.2 Neural Networks and Backpropagation

Neural networks provide a flexible parametric function class for representing $\mathbf{g}_{\boldsymbol{\theta}}$ through a sequential composition of transformations, $\mathbf{g}_{\boldsymbol{\theta}} = \mathbf{g}_L \circ \ldots \circ \mathbf{g}_2 \circ \mathbf{g}_1$, where L defines the number of layers of the

network. When an input pattern $\mathbf{x}$ is presented to the network, it produces a final output $\mathbf{z}_L$ and a series of intermediate outputs. By defining $\mathbf{z}_0 = \mathbf{x}$ and $\mathbf{z}_L = \mathbf{g}_{\boldsymbol{\theta}}(\mathbf{x})$, we can write the forward evaluation as

$$\mathbf{z}_k = \mathbf{g}_k(\mathbf{z}_{k-1}) \text{ for } k = 1, \dots, L\,. \tag{5.5}$$

Each module $\mathbf{g}_k$ of the network involves two transformations,

(a) a coupling layer $C_{\mathbf{W}_k}$, that couples the inputs to the layer with the parameters $\mathbf{W}_k$ to optimize;
(b) other arbitrary manipulations $\boldsymbol{\sigma}$ of inputs/outputs. Typically, these are element-wise nonlinear activation functions with fixed parameters; we can for simplicity think of them as operations of the form $\boldsymbol{\sigma}(\mathbf{x}) = (\sigma(x_1), \dots, \sigma(x_n))$ applied to vector variables.

The resulting transformation can thus be written as $\mathbf{g}_k(\mathbf{z}_{k-1}) = \boldsymbol{\sigma}(C_{\mathbf{W}_k}(\mathbf{z}_{k-1}))$.

We will focus on fully connected modules, where the coupling $C_{\mathbf{W}}$ is simply a matrix-vector multiplication between the weights $\mathbf{W}_k$ and the input to the k-th layer; overall, the transformation operated by such a module can be expressed as $\boldsymbol{\sigma}(\mathbf{W}_k\mathbf{z}_{k-1})$. Another kind of coupling layer is given by convolutional layers, typically used in convolutional neural networks [237].

The parameters of the network are randomly initialized and then learnt by gradient based optimization with an objective function $\mathcal{L}$, which is a scalar function of the final output of the network. At each learning step, updates for the weights are proportional to the partial derivative of the loss with respect to each weight.

The computation of these derivatives is typically performed by backpropagation [238], a specialized instance of automatic differentiation. Backpropagation involves a two-phase process. Firstly, during a *forward pass*, the intermediate and final outputs of the network $\mathbf{z}_1, \dots, \mathbf{z}_L$ are evaluated and a value for the loss is returned. Then, in a second phase termed *backward pass*, derivatives of the loss with respect to each individual parameter of the network are computed by application of the chain rule. The gradients are computed one layer at a time, from the last layer to the first one; in the process, the intermediate outputs of the forward pass are reused, employing dynamic programming to avoid redundant calculations of intermediate, repeated terms.[2]

2: Note that invertible neural networks provide the possibility to not save, but rather recompute the intermediate activations during the backward pass, thus providing a memory efficient approach to backpropagation [239].

In matrix notation, the updates for the weights of the k-th fully connected layer $\mathbf{W}_k$ can then be written as

$$\Delta\mathbf{W}_k \propto \mathbf{z}_{k-1}\boldsymbol{\delta}_k^\top\,, \tag{5.6}$$

where $\boldsymbol{\delta}_k$ is the cumulative result of the backward computation in the backpropagation step up to the k-th layer, also called backpropagated error. We report the full derivation in Appendix D.1. We adopt the convention of defining $\mathbf{x}$, $\mathbf{z}_k$ and $\boldsymbol{\delta}_k$ as column vectors.

5.2.3 Difficulty of Optimizing the Jacobian Term of Neural Networks

In the case of the objective function specified in Eq. (5.3), we have $\mathscr{L}(\mathbf{x}) = \log p_{\boldsymbol{\theta}}(\mathbf{x})$. By defining

$$\mathscr{L}_p(\mathbf{x}) = \sum_i \log p_i(\mathbf{g}^i_{\boldsymbol{\theta}}(\mathbf{x})); \quad \mathscr{L}_J(\mathbf{x}) = \log \left|\det \mathbf{J}_{\mathbf{g}_\theta}(\mathbf{x})\right| , \tag{5.7}$$

the objective can be rewritten as $\mathscr{L}(\mathbf{x}) = \mathscr{L}_p(\mathbf{x}) + \mathscr{L}_J(\mathbf{x})$. The evaluation of the gradient of the first term $\mathscr{L}_p$ can be performed easily if a simple form for the latent density is chosen, as it only requires simple operations on top of a single forward pass of the neural network. Given that the loss is a scalar, as backpropagation is an instance of reverse mode differentiation [240], backpropagating the error relative to it in order to evaluate the gradients does not increase the overall complexity with respect to the forward pass alone.

In contrast, the evaluation of the gradient of the second term, $\mathscr{L}_J$, is very problematic, and our main concern in this chapter. The key computational bottleneck is in fact given by the evaluation of the Jacobian during the forward pass. Since the Jacobian involves derivatives of the function $\mathbf{g}_{\boldsymbol{\theta}}$ with respect to its inputs $\mathbf{x}$, this evaluation can again be performed through automatic differentiation. Overall, it can be shown [240] that both forward and backward mode automatic differentiation for a L-layer, fully connected neural network scale as $\mathcal{O}(Ln^3)$, with L the number of layers. This is prohibitive in many practical applications with a large data dimension n.

Normalizing flows with simple Jacobians. An approach to alleviate the computational cost of this operation is to deploy special neural network architectures for which the evaluation of $\mathscr{L}_J$ is trivial. For example, in autoregressive normalizing flows [97, 179, 227, 228] the Jacobian of the transformation is constrained to be lower triangular. In this case, its determinant can be trivially computed with a linear cost in n. Notice however that the computational cost of the forward pass still scales quadratically in n; the overall complexity of forward plus backward pass is therefore still quadratic in the input size [210].

Most critically, such architectures imply a strong restriction on the class of transformations that can be learnt. While it can be shown, based on [70], that under certain conditions this class of functions has universal approximation capacity for *densities* [179], that is less general than other notions of universal approximation [241, 242]. In fact it is obvious that functions with such triangular Jacobians cannot be universal approximators of *functions*, since, for example, the first variable can only depend on the first variable. This is a severe problem in learning features for disentanglement, for example by nonlinear ICA, which would usually require unconstrained Jacobians. In other words, such restrictions might imply that the deployed networks are not general purpose: [230] showed that constrained designs typically used for density estimation can severely hurt discriminative performance. We further elaborate on this point in Appendix D.5. Note that fully connected modules have

elsewhere been termed *linear* flows [71], and are a strict generalization of autoregressive flows.[3]

3: Comprehensive reviews on normalizing flows can be found in [71, 98]. Other related methods are reviewed in Appendix D.2.

5.3 Log-Determinant of the Jacobian for Fully Connected Neural Networks

As a first step toward efficient optimization of the $\mathscr{L}_J$ term, we next provide the explicit form of the Jacobian for fully connected neural networks. As a starting point, notice that invertible and differentiable transformations are *composable*; given any two such transformations, their composition is also invertible and differentiable. Furthermore, the determinant of the Jacobian of a composition of functions is given by the product of the determinants of the Jacobians of each function,

$$\det \mathbf{J}_{\mathbf{g}_2 \circ \mathbf{g}_1}(\mathbf{x}) = \det \mathbf{J}_{\mathbf{g}_2}(\mathbf{g}_1(\mathbf{x})) \cdot \det \mathbf{J}_{\mathbf{g}_1}(\mathbf{x}) \,. \tag{5.8}$$

The log-determinant of the full Jacobian for a neural network therefore simply decomposes in a sum of the log-determinants of the Jacobians of each module, $\mathscr{L}_J(\mathbf{x}) = \sum_{k=1}^{L} \log |\det \mathbf{J}_{\mathbf{g}_k}(\mathbf{z}_{k-1})|$. We will focus on the Jacobian term relative to a single submodule k with respect to its input $\mathbf{z}_{k-1}$; with a slight abuse of notation, we will call it $\mathscr{L}_J(\mathbf{z}_{k-1})$. As we remarked, fully connected $\mathbf{g}_k$ are themselves compositions of a linear operation and an element-wise invertible nonlinearity; applying the same reasoning, we then have

$$\mathscr{L}_J(\mathbf{z}_{k-1}) = \sum_{i=1}^{n} \log \left|\sigma'(y_k^i)\right| + \log |\det \mathbf{W}_k| \;=: \mathscr{L}_J^1(\mathbf{y}_k) + \mathscr{L}_J^2(\mathbf{z}_{k-1}) \,. \tag{5.9}$$

where $\mathbf{y}_k = \mathbf{W}_k \mathbf{z}_{k-1}$. The first term $\mathscr{L}_J^1$ is a sum of univariate functions of single components of the output of the module, and it can be evaluated easily with few additional operations on top of intermediate outputs of a forward pass; gradients with respect to it can be simply computed via backpropagation, not unlike the $\mathscr{L}_p$ term introduced in § 5.2.3.

The second term $\mathscr{L}_J^2$ however involves a nonlinear function of the determinant of the weight matrix. From matrix calculus, we know that the derivative is equal to

$$\frac{\partial \log |\det \mathbf{W}_k|}{\partial \mathbf{W}_k} = \left(\mathbf{W}_k^\top\right)^{-1} \,. \tag{5.10}$$

Therefore, the computation of the gradient relative to such term involves a matrix inversion, with cubic scaling in the input size.[4] For a fully connected neural network of L layers, given that we have one such operation to perform for each of the layers, the gradient computation for these terms alone would have a complexity of $\mathcal{O}(Ln^3)$, thus matching the one which would be obtained if the Jacobian were to be computed via automatic differentiation as discussed in § 5.2.

4: Though slightly more favorable exponents can in principle be obtained, see Appendix D.3.

It can therefore be seen that these inverses of the weight matrices are the problematic element in the gradient computation. In the next section, we show how this problem can be solved using relative gradients.

5.4 Relative Gradient Descent for Neural Networks

We now derive the basic form of the relative gradient, following the approach in [232].[5] The starting point is that the parameters in a neural networks are matrices, in particular invertible in our case. Thus, we can make use of the geometric properties of invertible matrices, while they are usually completely neglected in gradient optimization in neural networks.

5: For linear blind source separation, this approach also corresponds to the natural gradient, which can be justified with an information-geometric approach [233]. See also § 7.2.2, where we used relative gradients for the MultiviewICA model of Chapter 7.

Relative gradient based on multiplicative perturbation. In a classical gradient approach for optimization, we add a small vector $\boldsymbol{\epsilon}$ to a point $\mathbf{x}$ in a Euclidean space. However, with matrices, we are actually perturbing a matrix with another, and this can be done in different ways. In the relative gradient approach, we make a *multiplicative* perturbation of the form

$$\mathbf{W}_k \rightarrow (\mathbf{I} + \boldsymbol{\epsilon})\mathbf{W}_k \tag{5.11}$$

where $\boldsymbol{\epsilon}$ is an infinitesimal matrix. If we consider the effect of such a perturbation on a scalar-valued function $f(\mathbf{W}_k)$, we have

$$f((\mathbf{I}+\boldsymbol{\epsilon})\mathbf{W}_k) - f(\mathbf{W}) = \langle \nabla f(\mathbf{W}_k), \boldsymbol{\epsilon}\mathbf{W}_k \rangle + o(\mathbf{W}_k) = \langle \nabla f(\mathbf{W}_k)\mathbf{W}_k^\top, \boldsymbol{\epsilon} \rangle + o(\mathbf{W}_k) \tag{5.12}$$

which shows that the direction of steepest descent in this case is given by making $\boldsymbol{\epsilon} = \mu \nabla f(\mathbf{W}_k)\mathbf{W}_k^\top$ where μ is an infinitesimal step size. Furthermore, when we combine this $\boldsymbol{\epsilon}$ with the definition of a multiplicative update, we find that the best perturbation to $\mathbf{W}$ is actually given as

$$\mathbf{W}_k \rightarrow \mathbf{W}_k + \mu \nabla f(\mathbf{W}_k)\mathbf{W}_k^\top \mathbf{W}_k \tag{5.13}$$

That is, the classical Euclidean gradient is replaced by $\nabla f(\mathbf{W}_k)\mathbf{W}_k^\top \mathbf{W}_k$, i.e. it is multiplied by $\mathbf{W}_k^\top \mathbf{W}_k$ from the right. This is the relative gradient.

A further alternative can be obtained by perturbing the weight matrices from the right, as $\mathbf{W}_k \rightarrow \mathbf{W}_k(\mathbf{I} + \boldsymbol{\epsilon})$. A similar derivation shows that in this case, the optimal $\boldsymbol{\epsilon}$ is given by $\mathbf{W}_k \mathbf{W}_k^\top \nabla f(\mathbf{W}_k)$; we refer to this as *transposed relative gradient*. In the context of linear ICA, the properties of the relative and transposed relative gradient were discussed in [243]. This version of the relative gradient might be useful in some cases; for example, the transposed relative gradient can be implemented more straightforwardly in neural network packages where the convention is that vectors are represented as rows.

The relative gradient belongs to the more general class of gradient descent algorithms on Riemannian manifolds [244]. Specifically, relative gradient descent is a first order optimization algorithm on the manifold of invertible $n \times n$ matrices. Almost sure convergence of the parameters to a critical point of the gradient of the cost function can be derived even for its stochastic counterpart, with decreasing step size and under suitable assumptions (see e.g. [245]).

Jacobian term optimization through the relative gradient. In § 5.3, we showed that the difficulty in computing the gradient of the log-

determinant is in the terms $\mathscr{L}_J^2$, whose gradient involves a matrix inversion. Now we show that by exploiting the relative gradient, this matrix inversion vanishes. In fact, when multiplying the right hand side of equation (5.10) by $\mathbf{W}_k^\top \mathbf{W}_k$ from the right we get

$$\left(\mathbf{W}_k^\top\right)^{-1} \mathbf{W}_k^\top \mathbf{W}_k = \mathbf{W}_k \,, \tag{5.14}$$

and similarly when multiplying by $\mathbf{W}_k \mathbf{W}_k^\top$ from the left. Most notably, we therefore have to perform *no additional operation* to get the relative gradient with respect to this term of the loss; it is, so to say, *implicitly* computed — as we know that the update for the parameters in $\mathbf{W}_k$ with respect to the error term $\mathscr{L}_J^2$ is proportional to $\mathbf{W}_k$ matrix itself.

As for the remaining terms of the loss, $\mathscr{L}_p$ and $\mathscr{L}_J^1$, simple backpropagation allows us to compute the weight updates given by the ordinary gradient in equation (5.6), which still need to be multiplied by $\mathbf{W}_k^\top \mathbf{W}_k$ to turn it into a relative gradient. We will next see that we can do this avoiding matrix-matrix multiplications, which would be computationally expensive. Note that backpropagation necessarily computes the $\boldsymbol{\delta}_k$ vector in equation (5.6) and for our model, by applying the relative gradient carefully, we can avoid matrix-matrix multiplication altogether by computing

$$(\Delta \mathbf{W}_k)\, \mathbf{W}_k^\top \mathbf{W}_k \propto \mathbf{z}_{k-1} \left(\left(\boldsymbol{\delta}_k^\top \mathbf{W}_k^\top \right) \mathbf{W}_k \right) \,. \tag{5.15}$$

Thus, we have a cheap method for computing the gradient of the log-determinant of the Jacobian, and of our original objective function. In Appendix D.4 we provide an explanation of how our procedure can be implemented with relative ease on top of existing deep learning packages.

While we so far only discussed update rules for the weight matrices of the neural network, our approach can be extended to include biases. Including bias terms in our multilayer network endows it with stronger approximation capacity. We detail how to do this in Appendix D.6.

Complexity. Note that the parentheses in equation (5.15) stress the point that the relative gradient updates only require matrix-vector or vector-vector multiplications, each of which scales as $\mathcal{O}(n^2)$, in a fixed number at each layer; that is, overall $\mathcal{O}(Ln^2)$ operations. They therefore do not increase the complexity of a normal forward pass. Furthermore, the overall complexity with respect to the input size is quadratic, resulting in an overall quadratic scaling with the input size as in normalizing flow methods [210], but without imposing strong restrictions on the Jacobian of the transformation.

Extension to convolutional layers. As we remarked in § 5.2.2, the formalism we introduced includes convolutional neural networks (CNNs) [237]. A natural question is therefore whether our approach can be extended to that case. The first natural question pertains the invertibility of convolutional neural networks; the convolution operation was shown [246] to be invertible under mild conditions (see Appendix D.7), and the standard pooling operation can be by replaced an invertible operation [31]. We therefore believe that the general formalism can be applied to CNNs;

this would require the derivation of the relative gradient for tensors. We believe that this should be possible but leave it for future work.

Invertibility and generation. Given that invertible and differentiable transformations are composable, as discussed in § 5.3, invertibility of our learnt transformation is guaranteed as long as the weight matrices and the element-wise nonlinearities are invertible. Square and randomly initialized (e.g. with uniform or normally distributed entries) weight matrices are known to be invertible with probability one; invertibility of the weight matrices throughout the training is guaranteed by the fact that the $\mathscr{L}_J^2$ terms would go to minus infinity for singular matrices (though high learning rates and numerical instabilities might compromise it in practice), as in estimation methods for linear ICA [46, 182, 232]. We additionally employ nonlinearities which are invertible by construction; we include more details about this in Appendix D.8. If we are interested in data generation, we also need to invert the learnt function. In practice, the cost of inverting each of the matrices is $\mathcal{O}(n^3)$, but the operation needs to be performed only once. As for the nonlinear transformation, the inversion is cheap since we only need to numerically invert a scalar function, for which often a closed form is available.

5.5 Experiments

In the following we experimentally verify the computational advantage of the relative gradient. The code used for our experiments can be found at `https://github.com/fissoreg/relative-gradient-jacobian`.

Computation of relative vs. ordinary gradient. As a first step, we empirically verify that our proposed procedure using the formulas in § 5.4 leads to a significant speed-up in computation of the gradient of the Jacobian term. We compare the relative gradient against an explicit computation of the ordinary gradient, as described in § 5.3, and with a computation based on automatic differentiation, as discussed in § 5.2.3, where the Jacobian is computed with the JAX package [177]. While the output and asymptotic computational complexity of the ordinary gradient and automatic differentiation methods should be the same, a discrepancy is to be expected at finite dimensionality due to differences in how the computation is implemented. In the experiment, we generate 100 random normally distributed datapoints and vary the dimensionality of the data from 10 to beyond 20,000. We then define a two-layer neural network and evaluate the gradient of the Jacobian. The main comparison is run on a Tesla P100 Nvidia GPU. For the main plots, we deactivated garbage collection. Plots with CPU and further details on garbage collection can be found in Appendix D.8.1. For each dimension we computed 10 iterations with a batch size of 100. Results are shown in figure 5.1, left. On the y-axis we report the average of the execution times of 100 successive gradient evaluations (forward plus backward pass in the automatic differentiation case). It can be clearly seen that *the relative gradient is much faster*, typically by two orders of magnitude. Autodiff computations could actually only be performed for the smallest dimension due to a memory problem. We report additional details on memory consumption in Appendix D.8.1.

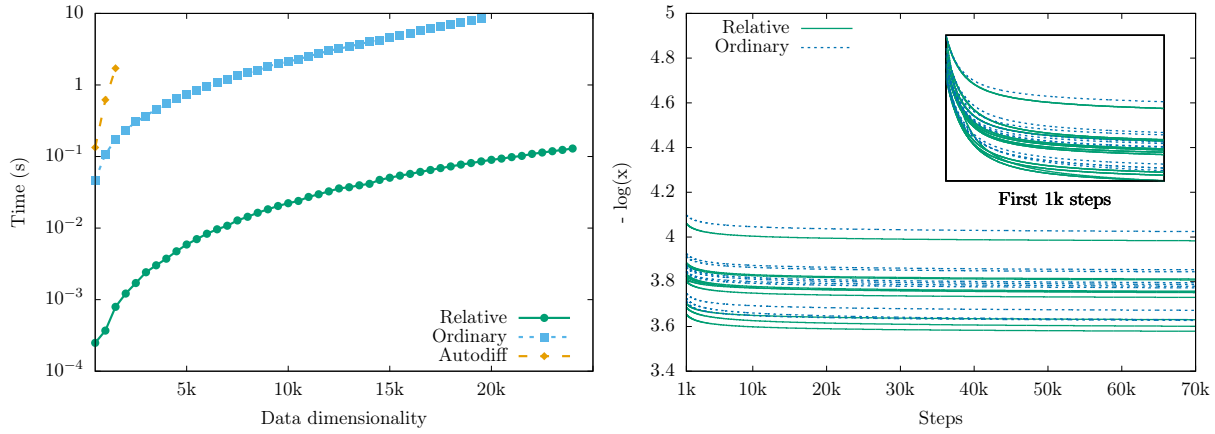


Figure 5.1: Left: Comparison of the average computation times of a single evaluation of the gradient of the log-likelihood; the standard error of the mean is not reported as it is orders of magnitude smaller then the scale of the plot. **Right:** Time-evolution of the negative log-likelihood for deterministic full-batch optimization for the two methods with the same initial points.

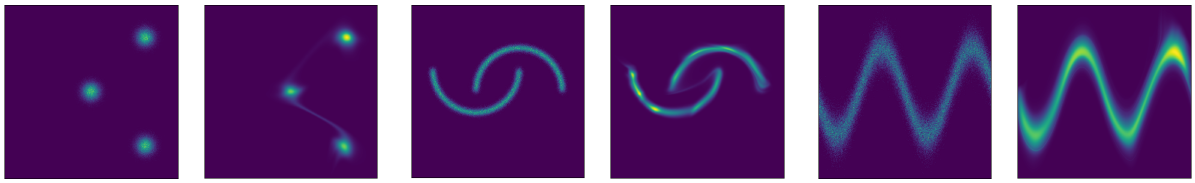


Figure 5.2: Illustrative examples of two-dimensional density estimation. Samples from the true distribution and predicted densities are shown, in this order, side by side.

Optimization by relative vs. ordinary gradient. Since the method presented in this chapter and the original publication [86] is, to the best of our knowledge, the first one proposing relative gradient optimization for neural networks (though other kinds of natural gradients have been studied [233]), we want to verify that the learning dynamics induced by the relative as opposed to the ordinary gradient do not bias the training procedure towards less optimal solutions or create other problems. We therefore perform a deterministic (full batch) gradient descent for both the relative and the ordinary gradient.[6] We employ 1,000 datapoints of dimensionality 2 and a two-layer neural network. We take 10 initial points and initialize both kinds of gradient descent at those same points. On the x-axis we plot the training epoch, while on the y-axis we plot the value of the loss. Figure 5.1, right shows the results: there is no big difference between the two gradient methods. There may actually be a slight advantage for the relative gradient, but that is immaterial since our main point here is merely to show that the *relative gradient does not need more iterations* to give the same performance.

6: Notice that there's no need to compare to autodiff in this case because the computed gradient should be exactly the same as the ordinary gradient with the formulas in § 5.3.

Combining these two results, we see that the proposed relative gradient approach leads to a *much faster optimization* than the ordinary gradient. Perhaps surprisingly, the results exhibit a rather constant speed-up factor of the order of 100 although the theory says it should be changing with the dimension n; in any case, the difference is very significant in practice.

Density estimation. Although our main contribution is the computational speed-up of the gradient computation demonstrated above, we further show some simple results on density estimation to highlight

the potential of the relative gradient used in conjuction with the unconstrained factorial approximation in § 5.2.1. We use a fairly simple feedforward neural network with a smooth version of leaky-ReLU as activation function. Our empirical results show that this system, despite having quite *minimal fine-tuning* (details in Appendix D.8.3), *achieves competitive results on all the considered datasets* compared with existing models—which are all tailored and fine-tuned for density estimation. First, we show in Figure 5.2 different toy examples that showcase the ability of our method to convincingly model arbitrarily complex densities. Second, in order to show the viability of our method in comparison with well-established methods we perform, as in [247], unconditional density estimation on four different UCI datasets [248] and a dataset of natural image patches (BSDS300) [249], as well as on MNIST [250]. The results are shown in Table 5.1. To achieve a fair comparison across models, the number of parameters was tuned so that the number of trainable parameters are as similar as possible. Note that, as we can perform every computation efficiently, all the experiments are suitable to run on usual hardware, thus avoiding the need of hardware accelerators such as GPUs. As a final remark, the reported results make no use of batch normalization, dropout, or learning-rate scheduling. Therefore, it is sensible to expect even better results by including them in future work.

Table 5.1: Test log-likelihoods (higher is better) on unconditional density estimation for different datasets and models (same as in Table 1 of [247]). Models use a similar number of parameters; results show mean and two standard deviations. Best performing models are in bold. More details in Appendix D.8.3

	POWER	GAS	HEPMASS	MINIBOONE	BSDS300	MNIST
Ours	0.065 ± 0.013	6.978 ± 0.020	−21.958 ± 0.019	−13.372 ± 0.450	151.12 ± 0.28	−1375.2 ± 1.4
MADE	−3.097 ± 0.030	3.306 ± 0.039	−21.804 ± 0.020	−15.635 ± 0.498	146.37 ± 0.28	−1380.8 ± 4.8
MADE MoG	**0.375 ± 0.013**	7.803 ± 0.022	**−18.368 ± 0.019**	−12.740 ± 0.439	150.84 ± 0.27	**−1038.5 ± 1.8**
Real NVP (10)	0.182 ± 0.014	**8.357 ± 0.019**	−18.938 ± 0.021	**−11.795 ± 0.453**	**153.28 ± 1.78**	−1370.7 ± 10.1
Real NVP (5)	−0.459 ± 0.010	6.656 ± 0.020	−20.037 ± 0.020	−12.418 ± 0.456	151.76 ± 0.27	−1323.2 ± 6.6
MAF (5)	−0.458 ± 0.016	7.042 ± 0.024	−19.400 ± 0.020	−11.816 ± 0.444	149.22 ± 0.28	−1300.5 ± 1.7
MAF (10)	−0.376 ± 0.017	7.549 ± 0.020	−25.701 ± 0.025	−11.892 ± 0.459	150.46 ± 0.28	−1313.1 ± 2.0
MAF MoG (5)	0.192 ± 0.014	7.183 ± 0.020	−22.747 ± 0.017	−11.995 ± 0.462	152.58 ± 0.66	−1100.3 ± 1.6

5.6 Conclusion

Using relative gradients, we proposed a new method for exact optimization of objective functions involving the log-determinant of the Jacobian of a neural network, as typically found in density estimation, nonlinear ICA, and related tasks. The relative gradient approach proposed here is quite simple, yet rather powerful. The importance of the optimization of the log-determinant of the Jacobian is well-known, but it has not been previously shown that there is a way around its difficulty for the class of models we considered. This allows for employing models which, unlike typical alternatives in the normalizing flows literature, have no strong limitation on the structure of the Jacobian. We use modules with fully connected layers, thus strictly generalizing normalizing flows with triangular Jacobians, while still supporting efficient combination of forward and backward pass. These neural network modules can represent a

larger function class than autoregressive ones, which can only represent transformations with triangular Jacobians. Our method can therefore provide an alternative in settings where more expressiveness is needed to learn a proper inverse transformation, such as in identifiable nonlinear ICA models.

Multiple Views and Data Augmentation

The Incomplete Rosetta Stone Problem: Identifiability Results for Multi-View Nonlinear ICA

6

We consider the problem of recovering a shared latent source vector with independent components from multiple views. This applies for example to settings in which a variable is measured with multiple experimental modalities, and where the goal is to synthesize the disparate measurements into a single unified representation; or to group studies where multiple subjects are exposed to the same experimental stimulus, and we are interested in their common or shared response to it. We consider the case that the observed views are a nonlinear mixing of component-wise corruptions of the shared sources. When the views are considered separately, this reduces to nonlinear ICA, for which it is provably impossible to undo the mixing—as shown in § 2.4.1. We present novel identifiability proofs that this is instead possible when the multiple views are considered jointly, showing that the mixing can be undone using function approximators such as deep neural networks. In contrast to known identifiability results for nonlinear ICA, we prove that independent latent sources with arbitrary mixing can be recovered as long as multiple, sufficiently different noisy views are available.

6.1 Introduction

Consider the setting described by the following generative model

$$\mathbf{x}_1 = \mathbf{f}_1(\mathbf{s}) \tag{6.1}$$

$$\mathbf{x}_2 = \mathbf{f}_2(\mathbf{s}) \tag{6.2}$$

$$p_{\mathbf{s}}(\mathbf{s}) = \prod_i p_{s_i}(s_i)\,, \tag{6.3}$$

where $\mathbf{x}_1, \mathbf{x}_2, \mathbf{s} \in \mathbb{R}^n$ and $\mathbf{f}_1, \mathbf{f}_2$ are arbitrary smooth and invertible transformations of the latent variable $\mathbf{s} = [s_1, \ldots, s_n]$ with mutually independent components. The goal is to recover $\mathbf{s}$, undoing the mixing induced by the $\mathbf{f}_i$, in the case where only observations of $\mathbf{x}_1$ and $\mathbf{x}_2$ are available. The two decoupled problems defined by considering pairs of Equations (6.1), (6.3) and (6.2), (6.3) separately are instances of nonlinear ICA.

As we reviewed in § 2.4.2, a breakthrough in the nonlinear ICA problem was to leverage contrastive learning, recasting the problem of unsupervised learning as a supervised one [64–66, 80]. This is a powerful proof technique, which additionally provides algorithms which can be practically implemented using modern deep learning frameworks. The setup with auxiliary variables makes strong assumptions on the data generating mechanism, but allows for arbitrary nonlinear mixing of the sources. However, the unconditional independence assumption on the source components ((6.3)) is replaced by a *conditional* independence (as in (2.17)).

In this chapter, we employ contrastive learning to address the setting specified by Equations 6.1–6.3. This corresponds to cases in which multiple recordings of the same process, acquired with different instruments and possibly different modalities, are available, and the goal is to find an

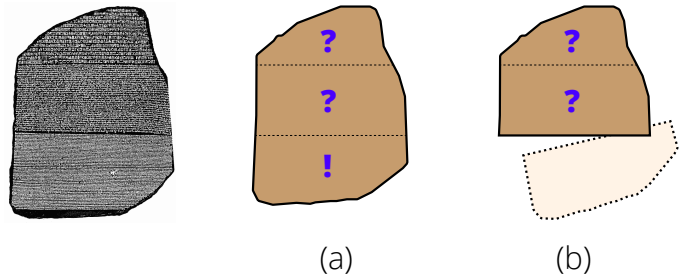


Figure 6.1: Left: The Rosetta Stone, a stele found in 1799, inscribed with three versions of a decree issued at Memphis, Egypt in 196 BC. The top and middle texts are in Ancient Egyptian using hieroglyphic script and Demotic script, respectively, while the bottom is in Ancient Greek. From [256]. **Right:** The Incomplete Rosetta Stone Metaphor: (a) The Rosetta Stone, with the known language denoted by an exclamation mark and the unknown ones denoted by question marks; (b) The Incomplete Rosetta Stone. Illustration courtesy of Alexander Neitz.

unambiguous representation of the latent state common to all. Multiview settings of this sort are common in large biomedical and neuroimaging datasets [251–254], motivating the need for reliable statistical tools enabling simultaneous handling of multiple sets of variables.

6.1.1 The Incomplete Rosetta Stone Metaphor

As a metaphor for such a setting, consider the story of the Rosetta Stone, a stele discovered during Napoleon's campaign in Egypt in 1799, inscribed with three versions of a decree issued at Memphis in 196 BC. The realisation that the stone reported the same text translated into three different languages led the French philologist Champollion to succeed in translating two unknown languages (Ancient Egyptian, in hieroglyphic script and Demotic script) by exploiting a known one (Ancient Greek) [255]. We instead consider the radically unsupervised task in which, given a Rosetta Stone with only two texts, both in unknown languages (Fig. 6.1), we want to learn an unambiguous common representation for both of them.

[255]: Champollion (1828), *Précis du systeme hiéroglyphique des anciens Egyptiens, ou Recherches sur les élémens premiers de cette écriture sacrée, sur leurs diverses combinaisons, et sur les rapports de ce systeme avec les autres méthodes graphiques égyptiennes avec un volume de planches*

The main contribution presented in this chapter is to show that jointly addressing multiple demixing problems allows for identifiability with assumptions which do not require abandoning the assumption of unconditional independence, nor restricting the class of mixing functions, but rather to the conditional probability distribution of one observation given the other. This provides identifiability results in a novel setting, with assumptions entailing a different interpretation - namely, that the views have to be sufficiently diverse.

Structure and contributions of this Chapter. In § 6.2 we present our main results, providing identifiability for different multi-view settings. In § 6.3 we discuss other related works in the literature. Finally, we summarise and discuss our results in § 6.4.

6.2 Nonlinear ICA with Multiple Views

We described how naively splitting Equations 6.1, 6.2 and 6.3 into two separate nonlinear ICA problems renders both problems non-identifiable, unless strong assumptions are made on the $\mathbf{f}_i$ or the distribution of $\mathbf{s}$.

In the Rosetta stone story, awareness that different texts reported on the stele were linked by a common topic helped solving the translation problem; similarly, in our setting, matched observations of the two views are linked through the shared latent variable $\mathbf{s}$. Thus the central question we investigate is whether these assumptions can be relaxed by exploiting the structure of the generative model; that is, whether jointly observing $\mathbf{x}_1$ and $\mathbf{x}_2$ provides sufficient constraints to the inverse problem, thus removing the ambiguities present in the vanilla nonlinear ICA setting. We consider a contrastive learning task in which a classifier is trained to distinguish between pairs $(\mathbf{x}_1, \mathbf{x}_2)$ corresponding to the same $\mathbf{s}$ and $(\mathbf{x}_1, \mathbf{x}_2^*)$ corresponding to different realisations of $\mathbf{s}$. In loose terms, the classifier will be forced to employ the information shared by the simultaneous views in order to distinguish the two classes. As we show, this ultimately results in recovering the sources up to the equivalence class in Defn. 2.4.1.

Component-wise corruptions of the sources. Our method requires some stochasticity in the relationship between $\mathbf{s}$ and at least one of the $\mathbf{x}_i$ for technical reasons discussed in Appendix E.1. A deviation from the basic setting described in (6.1)-(6.3) is required not only by our specific method and proof technique, based on contrastive learning: in fact, starting from that model and the true solution, we could simply apply a measure-preserving automorphism as those presented in § 2.4.1 to the true sources $\mathbf{s}$ and generate spurious solutions.[1]

1: This was first realised by Paul K. Rubenstein. Unfortunately, this was not included in the original publication.

We will therefore consider a component-wise independent corruption of our sources, i.e., $\mathbf{x}_1 = \mathbf{f}_1 \circ \mathbf{g}_1(\mathbf{s}, \mathbf{n}_1)$ with $g_{1i}(\mathbf{s}, \mathbf{n}_1) = g_{1i}(s_i, n_{1i})$, where the components of $\mathbf{n}_1$ are mutually independent, and similar for $\mathbf{x}_2$. The noise variables $\mathbf{n}_1$, $\mathbf{n}_2$ and the sources $\mathbf{s}$ are assumed to be mutually independent. Note that this only puts constraints on the way the signal is corrupted by the noise, namely $\mathbf{g}$, and not on the mixing $\mathbf{f}$. We will refer to such $\mathbf{g}$ as *component-wise corrupter* throughout, and to its output as *corruption*. In the the vanilla ICA setting, inverting the mixing and recovering the sources $\mathbf{s}$ are equivalent; in the setting that we consider, the inversion of the mixing $\mathbf{f}$ only implies recovering the sources up to the effect of the corrupter $\mathbf{g}$. This holds in general for noisy extensions of the ICA model: while the basic model in, e.g., (2.2) is noiseless and identifiability allows reconstruction of the random variable $\mathbf{s}$ up to scale and permutation, noisy models only allow reconstruction of the variables up to some uncertainty (see e.g. [62, Sec. 4.1]).

We will consider three instances of the general setting, providing identifiability results for each.

I. First we consider the case that only one of the observations, $\mathbf{x}_2$, is corrupted with noise. This corresponds, for instance, to a setting in which one accurate measurement device is supplemented with a second noisy device. We show that in this setting it is possible to fully reconstruct $\mathbf{s}$ using the noiseless variable (Section 6.2.1).
II. Next, we consider the case that both variables are corrupted with noise. In this setting, it is possible to recover $\mathbf{s}$ up to the corruptions. Furthermore, we show that $\mathbf{s}$ can be recovered with arbitrary precision in the limit that the corruptions go to zero (Section 6.2.2).

III. Finally, we consider the case of having N simultaneous views of the source $\mathbf{s}$ rather than just two. When considering the limit $N \to \infty$, we prove sufficient conditions under which it is possible to reconstruct $\mathbf{s}$ even if each observation is corrupted by noise (Section 6.2.3).

To the best of our knowledge, no result of identifiability of latent sources from multiple (corrupted) mixtures thereof, or *views*, had been given before the work on which this chapter is based.

6.2.1 One Noiseless View

Consider the generative model

$$\mathbf{x}_1 = \mathbf{f}_1(\mathbf{s}) \tag{6.4}$$
$$\mathbf{x}_2 = \mathbf{f}_2(\mathbf{g}(\mathbf{s}, \mathbf{n})) \tag{6.5}$$
$$p(\mathbf{s}) = \prod_i p_i(s_i)$$
$$p(\mathbf{n}) = \prod_i p_i(n_i) \tag{6.6}$$

where $\mathbf{f}_1$ and $\mathbf{f}_2$ are invertible, $\mathbf{g}$ is a component-wise corrupter, $\mathbf{n} \perp\!\!\!\perp \mathbf{s}$ and $\mathbf{x}_1$ and $\mathbf{x}_2$ are observed. This is represented in Fig. 6.2.

Subject to some assumptions, it is possible to recover $\mathbf{s}$ up to the equivalence class in Defn. 2.4.1.

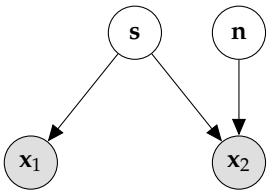


Figure 6.2: The setting considered in § 6.2.1. Two views of the sources are available, one of which, $\mathbf{x}_1$, is not corrupted by noise. In this and all other figures in this chapter, each node is a deterministic function of all its parents in the graph.

Theorem 6.2.1 *The difference of the log joint probability and log product of marginals of the observed variables in the generative model specified by Equations 6.4-6.6 admits the following factorisation:*

$$\begin{aligned}
&\log p(\mathbf{x}_1, \mathbf{x}_2) - \log p(\mathbf{x}_1)p(\mathbf{x}_2) \\
&= \log p(\mathbf{x}_2|\mathbf{x}_1) - \log p(\mathbf{x}_2) \\
&= \left(\sum_i \alpha_i(s_i, g_i(s_i, n_i)) + \log\det \mathbf{J}\mathbf{f}_2^{-1}(\mathbf{x}_2) \right) \\
&\quad - \left(\sum_i \delta_i(g_i(s_i, n_i)) + \log\det \mathbf{J}\mathbf{f}_2^{-1}(\mathbf{x}_2) \right) \\
&= \sum_i \alpha_i(s_i, g_i(s_i, n_i)) - \sum_i \delta_i(g_i(s_i, n_i))
\end{aligned} \tag{6.7}$$

where $s_i = f_{1i}^{-1}(\mathbf{x}_1)$, $g_i = f_{2i}^{-1}(\mathbf{x}_2)$, *and* $\mathbf{J}\mathbf{f}_2^{-1}(\mathbf{x}_2)$ *is the Jacobian of the transformation* $\mathbf{f}_2^{-1}$ *computed in* $\mathbf{x}_2$ *(note that the introduced Jacobians cancel). Suppose that*

1. $\boldsymbol{\alpha}$ *satisfies the* Sufficiently Distinct Views *assumption (see after this theorem).*
2. *We train a classifier to discriminate between*

$$(\mathbf{x}_1, \mathbf{x}_2) \text{ vs. } (\mathbf{x}_1, \mathbf{x}_2^*),$$

where $(\mathbf{x}_1, \mathbf{x}_2)$ *correspond to the same realisation of* $\mathbf{s}$ *and* $(\mathbf{x}_1, \mathbf{x}_2^*)$ *correspond to different realisations of* $\mathbf{s}$.

3. The classifier is constrained to use a regression function of the form

$$r(\mathbf{x}_1, \mathbf{x}_2) = \sum_i \psi_i(h_i(\mathbf{x}_1), \mathbf{x}_2)$$

where $\mathbf{h} = (h_1, \ldots, h_n)$ *are invertible, smooth and have smooth inverse.*

Then, in the limit of infinite data and with universal approximation capacity, $\mathbf{h}$ *inverts* $\mathbf{f}_1$ *in the sense that the* $h_i(\mathbf{x}_1)$ *recover the independent components of* $\mathbf{s}$ *up to component-wise invertible transformations and permutation.*

The proof can be found in Appendix E.3.1. The assumption of invertibility for $\mathbf{h}$ could be satisfied by, e.g., the use of normalizing flows [210, 231] or deep invertible networks [31].

We remark that at several points in this paper we consider the difference between two log-probabilities. In all of these cases, the Jacobians introduced by a change of variables cancel out as in Equation 6.7. For brevity we omit explanation of this fact in the rest of the results.

The *Sufficiently Distinct Views (SDV)* assumption specifies in a technical sense that the two views available are sufficiently different from one another, resulting in more information being available in totality than from each view individually. In the context of Theorem 6.2.1, it is an assumption about the log-probability of the *corruption* conditioned on the source. Informally, it demands that the probability distribution of the corruption should vary significantly as a result of conditioning on different values of the source.

Definition 6.2.1 (Sufficiently Distinct Views) *Let* $\alpha_i(y_i, t_i)$, $i = 1, \ldots, N$ *be functions of two arguments. Denote by* $\boldsymbol{\alpha}$ *the vector of functions and define*

$$\alpha_i'(y_i, t_i) = \partial \alpha_i(y_i, t_i)/\partial t, \tag{6.8}$$

$$\alpha_i''(y_i, t_i) = \partial^2 \alpha_i(y_i, t_i)/\partial t^2 \tag{6.9}$$

$$\boldsymbol{w}_{\boldsymbol{\alpha}}(\boldsymbol{y}, \boldsymbol{t}) = (\alpha_1'', \ldots, \alpha_D'', \alpha_1', \ldots, \alpha_D'). \tag{6.10}$$

We say that $\boldsymbol{\alpha}$ *satisfies the assumption of* Sufficiently Distinct Views (SDV) *if for any value of* $\boldsymbol{y}$*, there exist* $2D$ *distinct values* $\boldsymbol{t}_j$*,* $j = 1, \ldots, 2D$ *such that the vectors* $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_j)$ *are linearly independent.*

This is closely related to the Assumption of Variability in [66], see also Appendix A.2. We provide simple cases of conditional log-probability density functions satisfying and violating the SDV assumption in Appendix E.2.

Theorem 6.2.1 shows that by jointly considering the two views, it is possible to recover $\mathbf{s}$, in contrast to the single-view setting. This result can be extended to learn the inverse of $\mathbf{f}_2$ up to component-wise invertible functions and permutations.

Corollary 6.2.2 *Consider the setting of Theorem 6.2.1, and the alternative*

factorisation of the log joint probability given by

$$\begin{aligned}&\log p(\mathbf{x}_1, \mathbf{x}_2) - \log p(\mathbf{x}_1)p(\mathbf{x}_2)\\ &= \log p(\mathbf{x}_1|\mathbf{x}_2) - \log p(\mathbf{x}_1)\\ &= \sum_i \gamma_i(s_i, g_i(s_i, n_i)) - \sum_i \beta_i(s_i))\,.\end{aligned} \tag{6.11}$$

Suppose that $\boldsymbol{\gamma}$ satisfies the SDV assumption. Replacing the regression function with

$$r(\mathbf{x}_1, \mathbf{x}_2) = \sum_i \psi_i(\mathbf{x}_1, h_i(\mathbf{x}_2))$$

results in $\mathbf{h}$ inverting $\mathbf{f}_2$ in the sense that the $h_i(\mathbf{x}_2)$ recover the independent components of $\mathbf{g}(\mathbf{s}, \mathbf{n})$ up to component-wise invertible transformations.

The proof can be found in Appendix E.3.2. These two results together mean that it is possible to learn inverses $\mathbf{h}_1$ and $\mathbf{h}_2$ of $\mathbf{f}_1$ and $\mathbf{f}_2$, and therefore to recover $\mathbf{s}$ and $\mathbf{g}(\mathbf{s}, \mathbf{n})$, up to component-wise intertible functions. Note, however, that doing so requires running two separate algorithms. Furthermore, there is no guarantee that the learnt inverses $\mathbf{h}_1$ and $\mathbf{h}_2$ are 'aligned' in the sense that for each i the components $\mathbf{h}_{1i}(\mathbf{x}_1)$ and $\mathbf{h}_{2i}(\mathbf{x}_2)$ correspond to the same components of $\mathbf{s}$.

This problem of misalignment can be resolved by changing the form of the regression function.

Theorem 6.2.3 *Consider the settings of Theorem 6.2.1 and Corollary 6.2.2. Suppose that both $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ satisfy the SDV assumption. Replacing the regression function with*

$$r(\mathbf{x}_1, \mathbf{x}_2) = \sum_i \psi_i(h_{1,i}(\mathbf{x}_1), h_{2,i}(\mathbf{x}_2)) \tag{6.12}$$

results in $\mathbf{h}_1$, $\mathbf{h}_2$ inverting $\mathbf{f}_1$, $\mathbf{f}_2$ in the sense that the $h_{1,i}(\mathbf{x}_1)$ and $h_{2,i}(\mathbf{x}_2)$ recover the independent components of $\mathbf{s}$ and $\mathbf{g}(\mathbf{s}, \mathbf{n})$ up to two different component-wise invertible transformations. Furthermore, the two representations are aligned, i.e. for $i \neq j$,

$$h_{1,i}(\mathbf{x}_1) \perp\!\!\!\perp h_{2,j}(\mathbf{x}_2)$$

The proof can be found in Appendix E.4.

Note that Theorem 6.2.3 is *not* a generalisation of Theorem 6.2.1 or Corollary 6.2.2, since it makes stricter assumptions by imposing the SDV assumption on both $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. In contrast, Theorem 6.2.1 and Corollary 6.2.2 require that only one is valid for each.

For cases in which finding aligned representations for $\mathbf{s}$ and $\mathbf{g}(\mathbf{s}, \mathbf{n})$ are desired, Theorem 6.2.3 should be applied. If the only goal is recovery of $\mathbf{s}$, the assumptions of Theorem 6.2.1 are simpler to verify.

In practical applications, the multi-view scenario is useful in multimodal datasets where one of the two acquisition modalities has much higher signal to noise ratio than the other one (e.g., in neuroimaging, when simultaneous fMRI and Optical Imaging recordings are compared). In such cases, jointly exploiting the multiple modalities would help to discern a meaningful and identifiable latent representation which could not be attained through analysis of the more reliable modality alone.

Equivalence with Permutation Contrastive Learning for time-dependent sources. Note that the analysis of Theorem 6.2.1 covers the case of temporally dependent stationary sources analysed in [65]. Indeed, if it is further assumed that $\mathbf{s}$ and $\mathbf{g}(\mathbf{s}, \mathbf{n})$ are uniformly dependent [65, Definition 1], they can be seen as a pair of subsequent time points of an ergodic stationary stochastic process for which the analysis of Theorem 1 of [65] would hold. In other words, we can define a stochastic process as $p(\mathbf{s}_{t+1}|\mathbf{s}_t) := p(\mathbf{g}(\mathbf{s}, \mathbf{n})|\mathbf{s})$. Note that while the two formulations are theoretically equivalent, our view offers a wider applicability as it covers the asynchronous sensing of $\mathbf{s}$, provided that multiple measurements (i.e. $\mathbf{x}_1, \mathbf{x}_2$) are available; additionally, our *Sufficiently Distinct Views* assumption does not necessarily imply uniform dependency. Furthermore, while [65] considers a generative model of the form $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t))$, thus constraining the mixing function to be the same for any two data points $\mathbf{x}(t_1)$, $\mathbf{x}(t_2)$, in our setting we consider two different mixing functions, $\mathbf{f}_1$ and $\mathbf{f}_2$, for the two different views. Finally, we study this setting as an intermediate step for the following two sections, in which no deterministic function of the sources is observed, learning to invert any of the $\mathbf{f}_i$ can only recover $\mathbf{s}$ up to the corruption operated by $\mathbf{g}$.

[65]: Hyvärinen et al. (2017), 'Nonlinear ICA of Temporally Dependent Stationary Sources'

6.2.2 Two Noisy Views

We next consider the setting in which both variables are corrupted by noise. Consider the following generative model (represented in Fig. 6.3):

$$\begin{aligned} \mathbf{x}_1 &= \mathbf{f}_1(\mathbf{g}_1(\mathbf{s}, \mathbf{n}_1)) \\ \mathbf{x}_2 &= \mathbf{f}_2(\mathbf{g}_2(\mathbf{s}, \mathbf{n}_2)) \,, \end{aligned}$$

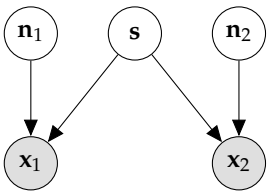


Figure 6.3: Setting with two views of the sources $\mathbf{s}$, both corrupted by noise.

where all variables take value in $\mathbb{R}^n$, and $\mathbf{f}_1$ and $\mathbf{f}_2$ are nonlinear, invertible, deterministic functions, $\mathbf{g}_1$ and $\mathbf{g}_2$ are component-wise corrupters, and $\mathbf{s}$ and the $\mathbf{n}_i$ are independent with independent components. This class of models generalises the setting of § 6.2.1 since by taking $\mathbf{g}_1(\mathbf{s}, \mathbf{n}_1) = \mathbf{s}$ we reduce to the case of one noiseless observation.

The difference $\log p(\mathbf{x}_1, \mathbf{x}_2) - \log p(\mathbf{x}_1)p(\mathbf{x}_2)$ admits similar factorisations to those given in Equations 6.7 and 6.11:

$$\begin{aligned} &\log p(\mathbf{x}_1, \mathbf{x}_2) - \log p(\mathbf{x}_1)p(\mathbf{x}_2) \\ &= \log p(\mathbf{x}_1|\mathbf{x}_2) - \log p(\mathbf{x}_1) \\ &= \sum_i \eta_i(g_{1i}(s_i, n_{i1}), g_{2i}(s_i, n_{2i})) - \sum_i \theta_i(g_{1i}(s_i, n_{1i}) \qquad (6.13) \\ &= \log p(\mathbf{x}_2|\mathbf{x}_1) - \log p(\mathbf{x}_2) \\ &= \sum_i \lambda_i(g_{2i}(s_i, n_{2i}), g_{1i}(s_i, n_{1i})) - \sum_i \mu_i(g_{2i}(s_i, n_{2i})) \qquad (6.14) \end{aligned}$$

Since we only have access to corrupted observations, exact recovery of $\mathbf{s}$ is not possible. Nonetheless, a generalisation of Theorem 6.2.3 holds showing that the $\mathbf{f}_i$ can be inverted and $\mathbf{s}$ recovered up to the corruptions induced by the $\mathbf{n}_i$ via $\mathbf{g}_i$.

Theorem 6.2.4 *Suppose that $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ satisfy the SDV assumption. The algorithm described in Theorem 6.2.1 with regression function specified in*

Equation 6.12 results in $\mathbf{h}_1$ *and* $\mathbf{h}_2$ *inverting* $\mathbf{f}_1$ *and* $\mathbf{f}_2$ *in the sense that the* $h_{1,i}(\mathbf{x}_1)$ *and* $h_{2,i}(\mathbf{x}_2)$ *recover the independent components of* $\mathbf{g}_1(\mathbf{s}, \mathbf{n}_1)$ *and* $\mathbf{g}_2(\mathbf{s}, \mathbf{n}_2)$ *up to two different component-wise invertible transformations. Furthermore, the two representations are aligned, i.e., for* $i \neq j$,

$$h_{1,i}(\mathbf{x}_1) \perp\!\!\!\perp h_{2,j}(\mathbf{x}_2)$$

The proof can be found in Appendix E.4.

We can thus recover the common source $\mathbf{s}$ up to the corruptions $\mathbf{g}_i(\mathbf{s}, \mathbf{n}_i)$. In the limit of the magnitude of one of the noise variables going to zero, the reconstruction of the sources $\mathbf{s}$ attained through the corresponding view is exact up to the component-wise invertible functions, as stated in the following corollary.

Corollary 6.2.5 *Let* $\mathbf{n}_1^{(k)} = \frac{1}{k} \cdot \tilde{\mathbf{n}}$ *for* $k \in \mathbb{N}$, *where* $\tilde{\mathbf{n}} \in \mathbb{R}^D$ *is a fixed random variable, and* $\mathbf{n}_2$ *be a random variable that does not depend on* k. *Let* $\mathbf{h}_1^{(k)}, \mathbf{h}_2^{(k)}$ *be the output of the algorithm specified by Theorem 6.2.4 with noise variables* $\mathbf{n}_1^{(k)}$ *and* $\mathbf{n}_2$.

Suppose that the corrupters $\mathbf{g}_i$ *satisfy the following two criteria:*

i) $\exists \mathbf{a} \in \mathbb{R}^D_{>0}$ *s.t.* $\left|\frac{\partial \mathbf{g}_1(\mathbf{s},\mathbf{n})}{\partial \mathbf{n}}\right|_{\mathbf{n}=0} \leq \mathbf{a}$ *for all* $\mathbf{s}$

ii) $\exists \mathbf{b} \in \mathbb{R}^D_{>0}$ *s.t.* $0 < \frac{\partial \mathbf{g}_1(\mathbf{s},0)}{\partial \mathbf{s}} \leq \mathbf{b}$

Then, denoting by $\mathbf{E}$ *the set of all scalar, invertible functions, we have that*

$$\lim_{k\to\infty} \inf_{\mathbf{e}\in\mathbf{E}} \left\| \mathbf{s} - \mathbf{e}(\mathbf{h}_1^{(k)}(\mathbf{x}_1)) \right\| = 0$$

The proof can be found in Appendix E.5.

Corollary 6.2.5 implies that in the limit of small noise, the sources $\mathbf{s}$ can be recovered exactly. Condition *i)* upper bounds the influence of $\mathbf{n}$ on the corruption: we can not hope to retrieve $\mathbf{s}$ if $\mathbf{g}(\mathbf{s}, \mathbf{n})$ contains too little signal. Condition *ii)* ensures that the function $\mathbf{g}$ is invertible with respect to $\mathbf{s}$ when $\mathbf{n}$ is equal to zero. If this were not satisfied, some information about $\mathbf{s}$ would be washed out by $\mathbf{g}$ even in absence of noise. This would make the recovery of $\mathbf{s}$ trivially impossible.

6.2.3 Multiple Noisy Views

The results of § 6.2.2 state that in the two noisy view setting, $\mathbf{s}$ can be recovered up to the corruptions. In the limit that the magnitude of the noises goes to zero, the uncorrupted $\mathbf{s}$ can be recovered. The intuition is that the less noise there is, the more information each observation provides about $\mathbf{s}$.

In this section we consider the multi-view setting, where N distinct noisy views of $\mathbf{s}$ are available,

$$\mathbf{x}_i = \mathbf{f}_i(\mathbf{g}_i(\mathbf{s}, \mathbf{n}_i)) \quad , i = 1, \ldots, N \,, \tag{6.15}$$

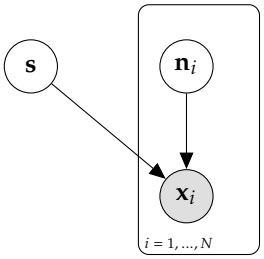


Figure 6.4: Setting with N corrupted views of the sources.

and the noise variables $\mathbf{n}_i$ are mutually independent, as represented in Fig. 6.4. Since each view provides additional information about $\mathbf{s}$, we ask: in the limit as $N \to \infty$, is it possible to reconstruct $\mathbf{s}$ exactly?

By applying Theorem 6.2.4 to the pair $(\mathbf{x}_1, \mathbf{x}_i)$ it is possible to recover $(\mathbf{g}_1(\mathbf{s}, \mathbf{n}_1), \mathbf{g}_i(\mathbf{s}, \mathbf{n}_i))$ such that the components are aligned, but up to different component-wise invertible functions $\mathbf{k}_1$ and $\mathbf{k}_i$. Running the algorithm on a different pair $(\mathbf{x}_1, \mathbf{x}_j)$ will result in recovery up to different component-wise invertible functions $\mathbf{k}'_1$ and $\mathbf{k}'_j$.

Note that these will *not* necessarily result in $\mathbf{k}_i \circ \mathbf{g}_i(\mathbf{s}, \mathbf{n}_i)$ and $\mathbf{k}'_j \circ \mathbf{g}_j(\mathbf{s}, \mathbf{n}_j)$ being aligned with each other. However, the components of $\mathbf{k}_1 \circ \mathbf{g}_1(\mathbf{s}, \mathbf{n}_1)$ and $\mathbf{k}'_1 \circ \mathbf{g}_1(\mathbf{s}, \mathbf{n}_1)$ are the same, up to permutation and component-wise invertible functions. This permutation can therefore be undone by performing independence testing between each pair of components. Components that are 'different' will be independent; those that are the same will be deterministically related. Therefore, they can be used as a reference to permute the components of $\mathbf{k}'_j$ and make it aligned with $\mathbf{k}_i$.

The problem is then how to combine the information from each aligned $\mathbf{k}_i \circ \mathbf{g}_i(\mathbf{s}, \mathbf{n}_i)$ to more precisely identify $\mathbf{s}$. The fact that the components are recovered up to *different* scalar invertible functions makes combining information from different views non-trivial.

As a first step in this direction, we consider the special case that each $\mathbf{g}_i$ acts additively and each $\mathbf{n}_i$ is zero mean and each of $\mathbf{s}$ and the $\mathbf{n}_i$ are independent with independent components.

$$\left.\begin{array}{l} \mathbf{x}_i = \mathbf{f}_i(\mathbf{s} + \mathbf{n}_i) \\ \mathbb{E}\mathbf{n}_i = 0 \end{array}\right\} \quad i \in \mathbb{N} \tag{6.16}$$

Suppose to begin with that we are able to recover each $\mathbf{s} + \mathbf{n}_i$ *without* the usual component-wise invertible functions. Then, writing $\mathbf{n}$ to denote all of the $\mathbf{n}_i$, it is possible to estimate $\mathbf{s}$ as

$$\mathbf{s} \approx \Omega^N(\mathbf{s}, \mathbf{n}) = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{s} + \mathbf{n}_i) \,.$$

Subject to mild conditions on the rate of growth of the variances $\mathrm{Var}(\mathbf{n}_i)$ as $i \to \infty$, Kolmogorov's strong law implies that $\Omega^N(\mathbf{s}, \mathbf{n})$ is a good approximation to $\mathbf{s}$ as $N \to \infty$ in the sense that $\Omega^N(\mathbf{s}, \mathbf{n}) \xrightarrow{a.s.} \mathbf{s}$. This implies moreover that it is possible to reconstruct the $\mathbf{n}_i$ by considering the residue $R_i^N(\mathbf{s}, \mathbf{n}) = (\mathbf{s} + \mathbf{n}_i) - \Omega^N(\mathbf{s}, \mathbf{n}) \xrightarrow{a.s.} \mathbf{n}_i$.

In the presence of the unknown functions $\mathbf{k}_i$, we would be able to reconstruct $\mathbf{s}$ and the $\mathbf{n}_i$ if we were able to identify the inverses $\mathbf{e}_i = \mathbf{k}_i^{-1}$ for each i. For any component-wise invertible functions $\mathbf{e}_i$, define

$$\begin{aligned} \Omega_{\mathbf{e}}^N(\mathbf{s}, \mathbf{n}) &= \frac{1}{N} \sum_{i=1}^{N} \mathbf{e}_i \circ \mathbf{k}_i(\mathbf{s} + \mathbf{n}_i) \\ R_{\mathbf{e},i}^N(\mathbf{s}, \mathbf{n}) &= \mathbf{e}_i \circ \mathbf{k}_i(\mathbf{s} + \mathbf{n}_i) - \Omega_{\mathbf{e}}^N(\mathbf{s}, \mathbf{n}). \end{aligned}$$

$\mathbf{e}_i$ is something we can choose and $\mathbf{k}_i(\mathbf{s}+\mathbf{n}_i) = \mathbf{h}_i(\mathbf{x}_i)$ is the output of the algorithm, and hence $\Omega^N_{\mathbf{e}}(\mathbf{s},\mathbf{n})$ and $R^N_{\mathbf{e},i}(\mathbf{s},\mathbf{n})$ are random variables with known distributions. Subject to mild conditions, the dependence of these quantities on most or all of the $\mathbf{n}_i$ becomes increasingly small as N grows and disappears in the limit $N \to \infty$.

Lemma 6.2.6 *Suppose that the sequence* $\mathbb{E}_{\mathbf{n}}[\Omega^N_{\mathbf{e}}(\mathbf{s},\mathbf{n})] = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbf{n}_i}[\mathbf{e}_i \circ \mathbf{k}_i(\mathbf{s}+\mathbf{n}_i)]$ *converges as* $N \to \infty$ *for almost all* $\mathbf{s}$*, and write*

$$\Omega_{\mathbf{e}}(\mathbf{s}) = \lim_{N\to\infty} \mathbb{E}_{\mathbf{n}}[\Omega^N_{\mathbf{e}}(\mathbf{s},\mathbf{n})].$$

Suppose further that there exists K *such that* $V_{\mathbf{e}_i} = \mathrm{Var}\left(\mathbf{e}_i \circ \mathbf{g}_i(\mathbf{s}+\mathbf{n}_i)\right) \leq K$ *for all* i*. Then*

$$\Omega^N_{\mathbf{e}}(\mathbf{s},\mathbf{n}) \xrightarrow{a.s.} \Omega_{\mathbf{e}}(\mathbf{s})$$
$$R^N_{\mathbf{e},i}(\mathbf{s},\mathbf{n}) \xrightarrow{a.s.} R_{\mathbf{e},i}(\mathbf{s},\mathbf{n}_i) = \mathbf{e}_i \circ \mathbf{k}_i(\mathbf{s}+\mathbf{n}_i) - \Omega_{\mathbf{e}}(\mathbf{s})$$

The proof can be found in Appendix E.6. Given some choice of $\mathbf{e}$, we can think of $\Omega_{\mathbf{e}}(\mathbf{s})$ and $R_{\mathbf{e},i}(\mathbf{s},\mathbf{n}_i)$ as our putative candidates for $\mathbf{s}$ and $\mathbf{n}_i$ respectively. As discussed earlier, if we could identify $\mathbf{e}_i = \mathbf{k}_i^{-1}$, then we would have $\Omega_{\mathbf{e}}(\mathbf{s}) = \mathbf{s}$ and $R_{\mathbf{e},i}(\mathbf{s},\mathbf{n}_i) = \mathbf{n}_i$, and thus $\Omega_{\mathbf{e}}$ and $R_{\mathbf{e},i}$ would satisfy the same independences and other statistical properties as $\mathbf{s}$ and $\mathbf{n}_i$ respectively. Can we use these properties as criteria to identify good choices of $\mathbf{e}_i$?

The following theorem gives a set of sufficient conditions under which each $\mathbf{e}_i$ inverts $\mathbf{k}_i$ up to some affine ambiguity which is the same for every i.

Theorem 6.2.7 *Suppose there exists* $C > 0$ *such that* $Var(\mathbf{n}_i) \leq C$ *for all* i *and let* $\mathcal{G}_K = \left\{\{\mathbf{e}_i\} \text{ s.t.}\right.$

$$V_{\mathbf{e}_i} \leq K \;\forall i \tag{6.17}$$
$$\Omega_{\mathbf{e}}(\mathbf{s}) < \infty \text{ for almost all } \mathbf{s} \tag{6.18}$$
$$R_{\mathbf{e},i} \perp\!\!\!\perp R_{\mathbf{e},j} \;\forall i \neq j, \tag{6.19}$$
$$\mathbb{E}R_{\mathbf{e},i} = 0 \;\forall i \tag{6.20}$$
$$\left. R_{\mathbf{e},i}(\mathbf{s},\mathbf{n}_i) = R_{\mathbf{e},i}(\mathbf{n}_i) \;\forall i \right\} \tag{6.21}$$

Then,

$$\mathcal{G}_K \subseteq \left\{\{\boldsymbol{\alpha}\mathbf{k}_i^{-1} + \boldsymbol{\beta}\} \;:\; \boldsymbol{\alpha} \in \mathbb{R}^D_{\neq 0},\, \boldsymbol{\beta} \in \mathbb{R}^D\right\}$$

where $\boldsymbol{\alpha}\mathbf{k}_i^{-1}$ *denotes the element-wise product with the scalar elements of* $\boldsymbol{\alpha}$*. If* $K \geq Var(\mathbf{s}) + C$*, then* $\{\mathbf{k}_i^{-1}\} \in \mathcal{G}_K$*, and so* $\mathcal{G}_K$ *is non-empty for* K *sufficiently large.*

The proof can be found in Appendix E.7. It follows that it is possible recover $\mathbf{s}$ and $\mathbf{n}_i$ up to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ via $\Omega_{\mathbf{e}}(\mathbf{s}) = \boldsymbol{\alpha}\mathbf{s} + \boldsymbol{\beta}$ and $R_{\mathbf{e},i}(\mathbf{n}_i) = \boldsymbol{\alpha}\mathbf{n}_i$.

We remark that each of the conditions 6.17–6.20 can be verified from known information. We conjecture that condition 6.21 can be relaxed to assuming the verifiable condition of independence between $\Omega_{\mathbf{e}}(\mathbf{s})$ and

$R_{\mathbf{e},i}(\mathbf{s}, \mathbf{n}_i)$ for all i along with additional regularity assumptions on the functional form of $R_{\mathbf{e},i}$ (e.g. smoothness).

To conclude, Theorem 8 provides sufficient conditions under which it is possible to fully reconstruct $\mathbf{s}$ with corrupted views. In contrast to previous results in Sections 6.2.1 and 6.2.2, this result leverages infinitely many corrupted views rather than vanishingly small corruption of finitely many views.

6.3 Related Work

A central concept in our work is that of multiple simultaneous views and joint extraction of features from them. We briefly review some related work considering similar settings.

6.3.1 Canonical Correlation Analysis

Given two (or more) random variables, the goal of Canonical Correlation Analysis (CCA) [257] is to find a corresponding pair of linear subspaces that have high cross-correlation, so that each component within one of the subspaces is correlated with a single component from the other subspace [190]. In dealing with correlation instead of independence, CCA is more closely related to Principal Component Analysis (PCA) than to ICA.

[257]: Hotelling (1936), 'Relations between two sets of variates.'

CCA can be interpretated probabilistically [258] and is equivalent to maximum likelihood estimation in a graphical model which is a special case of that depicted in Fig. 6.3. The differences compared to our setting are (i) the latent components retrieved in CCA are forced to be uncorrelated, whereas our method is retrieves independent components; (ii) in CCA, mappings between the sources $\mathbf{s}$ and $\mathbf{x}$ are linear, whereas our method allows for nonlinear mappings.

[258]: Bach et al. (2005), 'A probabilistic interpretation of canonical correlation analysis'

At a high level, the model we consider in § 6.2.2 is to CCA as nonlinear ICA is to PCA. Nonlinear extensions of the basic CCA framework have been proposed [259–262], and we review some additional work on unsupervised learning from multiple views in Appendix E.8. However, identifiability results in the sense we consider in this work are lacking.

6.3.2 Half-Sibling Regression

Half-sibling regression [263] is a method to reconstruct a source from noisy observations by exploiting other sources that are affected by the same noise process—but otherwise independent from it.

[263]: Schölkopf et al. (2016), 'Modeling confounding by half-sibling regression'

Suppose that a latent variable of interest Q is not directly available, and that we can only observe corrupted versions of it, denoted as Y, where the corruption is due to a noise N. Without knowledge of N, it is impossible to reconstruct Q. However, if one or more additional variables X, also influenced by N, are observed, we can exploit them to model the effect of N on Y by regressing Y on X.

Subtracting this from the observed Y recovers the latent variable Q up to a constant offset, provided that (1) the additivity assumption

$$Y = Q + f(N)$$

holds, and (2) that Y contains sufficient information about $f(N)$. Analogous to our aim of recovering $\mathbf{s}$, the goal of half-sibling regression is not to infer only the distribution of Q, but rather the random variable itself (almost surely).[2] Half-sibling regression provides an identifiability result for an overcomplete latent variable model, and serves as inspiration for the results presented in this chapter.

2: Cfr. also Fig. 6.2 and Fig. 1 in [263].

6.4 Discussion and Conclusion

We presented identifiability results in a novel setting by extending the formalism of nonlinear ICA. We have investigated different scenarios of multi-view latent variable models and provided theoretical proofs on the possibility of inverting the mixing function and recovering the sources in each case. Our results thus extend the scarce literature on identifiability for nonlinear ICA models.

In the classical noiseless ICA setting, the deterministic relationship between the sources and observations means that inverting the mixing function and recovering the sources are equivalent. In contrast, we consider views of corrupted versions of the common sources, resulting in the decoupling of the demixing and retrieval of the sources. Remarkably, Theorem 6.2.7 points towards the possibility of simultaneously solving the two problems in the limit of infinitely many views.

Classical nonlinear ICA is provably non-identifiable because a single view is not sufficiently informative to resolve non-trivial ambiguities when recovering the sources. In this chapter we considered exploiting additional views to constrain the problem. Intuitively, if a second view is identical to the first, then nothing is gained by its observation. Hence, in order for the second view to assist in resolving ambiguity, it must be sufficiently different from the first. This is the intuition behind the technical assumption of *sufficiently distinct views*.

Typically, noise is a nuisance variable that would be preferably non-existent. In our setting, however, the noise variables acting on the sources are a crucial component, without which the contrastive learning approach could not be applied. Furthermore, the assumption of sufficiently distinct views is ultimately an assumption about the complexity of the joint distribution of the (corrupted) sources corresponding to each view. Without the noise variables the sufficiently distinct views assumption could not hold.

Our setting is relevant in a number of practical real-world applications, namely in all datasets that include multiple distinct measurements of related phenomena. In practice, it may be better to think of the noise variables rather as intrinsic sources of variability specific to each view. An exemplary application of our method can be found in the field of neuroimaging. Consider a study involving a cohort of subjects (perceivers), measuring their response to the presentation of the same

stimulus. One of the key problems in the field is how to extract a shared response from all subjects despite high inter-subject variability and complex nonlinear mappings between latent source and observation [264, 265]. Our results provide principled ways to extract and decompose the components of the shared response. In particular, the setting described in our model is suited to account for the high variability of the responses throughout the cohort, since the measurement corresponding to each subject is given by a combination of individual variability and shared response. In fact, in Chapter 7, we will present an application of the multiview ICA setting to neuroimaging.

We note that Theorem 6.2.7 builds on the setting of Theorem 6.2.4 which only makes use of pairwise information from the observations. A natural extension of this work should investigate algorithms that explicitly make use of $N > 2$ views, which we conjecture would allow relaxation of the additivity assumption on the corruptions. Furthermore, Theorem 6.2.7 provides results that only hold for the asymptotic limit as the number of views becomes large. Other extensions to this result could include analysis of the case of finitely many views.

Modeling Shared Responses in Neuroimaging Studies through MultiView ICA

7

Group studies involving large cohorts of subjects are important to draw general conclusions about brain functional organization. However, the aggregation of data coming from multiple subjects is challenging, since it requires accounting for large variability in anatomy, functional topography and stimulus response across individuals. Data modeling is especially hard for ecologically relevant conditions such as movie watching, where the experimental setup does not imply well-defined cognitive operations. We propose a novel MultiView ICA model for group studies, where data from each subject are modeled as a linear combination of shared independent sources plus noise. Contrary to most group-ICA procedures, the likelihood of the model is available in closed form. We develop an alternate quasi-Newton method for maximizing the likelihood, which is robust and converges quickly. We demonstrate the usefulness of our approach first on fMRI data, where our model demonstrates improved sensitivity in identifying common sources among subjects. Moreover, the sources recovered by our model exhibit lower between-session variability than other methods. On magnetoencephalography (MEG) data, our method yields more accurate source localization on phantom data. Applied on 200 subjects from the Cam-CAN dataset it reveals a clear sequence of evoked activity in sensor and source space.

7.1 Introduction

The past decade has seen the emergence of two trends in neuroimaging: the collection of massive neuroimaging datasets, containing data from hundreds of participants [253, 266, 267], and the use of naturalistic stimuli to move closer to a real life experience with dynamic and multimodal stimuli [268]. Large scale datasets provide an unprecedented opportunity to assess the generality and validity of neuroscientific findings across subjects, with the potential of offering novel insights on human brain function and useful medical biomarkers. However, when using ecological conditions, such as movie watching or simulated driving, stimulations are difficult to quantify. Consequently the statistical analysis of the data using supervised regression-based approaches is difficult. This has motivated the use of unsupervised learning methods that leverage the availability of data from multiple subjects performing the same experiment; analysis on such large groups boosts statistical power.

Independent component analysis is a widely used unsupervised method for neuroimaging studies. It is routinely applied on individual subject electroencephalography (EEG) [269], magnetoencephalography (MEG) [270] or functional MRI (fMRI) [54] data. The identifiability theory of ICA states that having non-Gaussian independent sources is a strong enough condition to recover the model parameters (§ 2.2.2). ICA therefore does not make assumptions about what triggers brain activations in the stimuli, unlike confirmatory approaches like the general linear model [271, 272]. This explains why, in fMRI processing, it is a model of choice when analysing resting state data [273] or when subjects are exposed to

natural [274, 275] or complex stimuli such as simulated driving [276]. In M/EEG processing, it is widely used to isolate acquisitions artifacts from neural signal [277], and to identify brain sources of interest [278, 279].

However, unlike with univariate methods, statistical inference about multiple subjects using ICA is not straightforward: so-called group-ICA is the topic of various studies [280]. Several works assume that the subjects share a common mixing matrix [281, 282]. Instead, we focus on a model where the subjects share common sources, but have *different* mixing matrices. When the subjects are exposed to the same stimuli, the common sources corresponds to the group *shared responses*. Most methods proposed in this framework proceed in two steps [283, 284]. First, the data of individual subjects are aggregated into a single dataset, often resorting to dimension reduction techniques like Principal Component Analysis (PCA). Then, off-the-shelf ICA is applied on the aggregated dataset. This popular method has the advantage of being simple and straightforward to implement since it resorts to customary single-subject ICA method. However, it is not grounded in a principled probabilistic model of the problem, and does not have strong statistical guarantees like asymptotic efficiency.

We propose a novel group ICA method called *MultiView ICA*. It models each subject's dataset as a linear combination of a common sources matrix with additive Gaussian noise. Importantly, and inspired by the models discussed in Chapter 6, we consider that the noise is on the sources and not on the sensors. This greatly simplifies the likelihood of the model which can even be written in closed-form. Despite its simplicity, our model allows for an expressive representation of inter-subject variability through subject-specific functional topographies (mixing matrices) and variability in the individual response (with noise in the source domain). To the best of our knowledge, this is the first time that such a tractable likelihood is proposed for multi-subject ICA. The likelihood formulation shares similarities with the usual ICA likelihood, which allows us to develop a fast and robust alternate quasi-Newton method for its maximization.

Structure and contributions of this Chapter. In § 7.2, we introduce the MultiView ICA model and characterise its identifiability. We then write its likelihood in closed form, and maximize it using an alternate quasi-Newton method. We also provide a sensitivity analysis for MultiView ICA, and show that the choice of the noise parameter in the algorithm has little influence on the output. In § 7.3, we compare our approach to other group ICA methods. Finally, in § 7.4, we empirically verify through extensive experiments on fMRI and MEG data that it improves source identification with respect to competing methods, suggesting that the expressiveness and robustness of our model make it a useful tool for multivariate neural signal analysis.

7.2 Multiview ICA for Shared Response Modeling

7.2.1 Model, Likelihood and Approximation

Given m subjects, we model the data $\mathbf{x}^i \in \mathbb{R}^n$ of subject i as

$$\mathbf{x}^i = \mathbf{A}^i(\mathbf{s} + \mathbf{n}^i), \quad i = 1, \dots, m \tag{7.1}$$

where $\mathbf{s} = [s_1, \dots, s_n]^\top \in \mathbb{R}^n$ are the shared independent sources, $\mathbf{n}^i \in \mathbb{R}^n$ is individual noise, $\mathbf{A}^i \in \mathbb{R}^{n \times n}$ are the individual mixing matrices, assumed to be full-rank. We assume that samples are observed i.i.d. For simplicity, we assume that the sources share the same density, $p_{s_i} = d \ \forall i$, so that the distribution of the source vector can be written as $p_{\mathbf{s}}(\mathbf{s}) = \prod_{j=1}^n d(s_j)$. Finally, we assume that the noise is Gaussian decorrelated of variance σ^2, $\mathbf{n}^i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$, and that the noise is independent across subjects and independent from the sources. The assumption of additive white noise on the sources models individual deviations from the shared sources $\mathbf{s}$. It is equivalent to having noise on the sensors with covariance $\sigma^2 \mathbf{A}^i \left(\mathbf{A}^i\right)^\top$, i.e., a scaled version of the noiseless data covariance.[1]

1: The model in (7.1) is a special case of the one in (6.15), where the mixing functions are linear—i.e., for $\mathbf{z} \in \mathbb{R}^n$, $\mathbf{f}_i(\mathbf{z}) = \mathbf{A}^i \mathbf{z}$, and the corrupter function corrseponds to a sum—i.e., $\mathbf{g}(\mathbf{s}, \mathbf{n}^i) = \mathbf{s} + \mathbf{n}_i$.

Since the sources are shared by the subjects, there are many more observed variables than sources in the model: there are n sources, while there are $n \times m$ observations. Therefore, model (7.1) bears some similarities with the setting of *undercomplete* ICA. The goal of multiview ICA is to recover the mixing matrices $\mathbf{A}^i$ from observations of the $\mathbf{x}^i$. The following proposition extends the standard idenfitiability theory of ICA reviewed in § 2.2.2 to multiview ICA, and shows that recovering the sources/mixing matrices is a well-posed problem up to scale and permutation, as in Defn. 2.3.3.[2]

2: All proofs are deferred to Appendix F.3

Proposition 7.2.1 (Identifiability of MultiView ICA) *Consider* $\mathbf{x}^i$, $i = 1 \dots m$, *generated from* (7.1). *Assume that* $\mathbf{x}^i = \mathbf{A}'^i(\mathbf{s}' + \mathbf{n}'^i)$ *for some invertible matrices* $\mathbf{A}'^i \in \mathbb{R}^{n \times n}$, *independent non-Gaussian sources* $\mathbf{s}' \in \mathbb{R}^n$ *and Gaussian noise* $\mathbf{n}'^i$. *Then, there exists a scale and permutation matrix* $\mathbf{P} \in \mathbb{R}^{n \times n}$ *such that for all* i, $\mathbf{A}'^i = \mathbf{A}^i \mathbf{P}$.

We propose a maximum-likelihood approach to estimate the mixing matrices. We denote by $\mathbf{W}^i = (\mathbf{A}^i)^{-1}$ the unmixing matrices, and view the likelihood as a function of $\mathbf{W}^i$ rather than $\mathbf{A}^i$. As shown in Appendix F.1.1, the negative log-likelihood can be written by integrating over the sources[3]

$$\mathcal{L}(\mathbf{W}^1, \dots, \mathbf{W}^m) = -\sum_{i=1}^m \log |\mathbf{W}^i| + \\ - \log \left(\int_{\mathbf{s}} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^m \|\mathbf{W}^i \mathbf{x}^i - \mathbf{s}\|^2 \right) p(\mathbf{s}) d\mathbf{s} \right), \tag{7.2}$$

3: Unlike in Equation 2.6, here we write the likelihood in terms of the mixing matrices $\mathbf{W}^1, \dots, \mathbf{W}^m$ to make the dependence on those parameters which are optimised explicit, and denote it with the symbol $\mathcal{L}$ to improve readability.

up to additive constants. Since this integral factorizes (i.e., the integrand is a product of functions of s_j) we can perform the integration as shown

in Appendix F.1.2. We define a smoothened version of the logarithm of the source density d by convolution with a Gaussian kernel as

$$f(s) = \log\left(\int \exp(-\frac{m}{2\sigma^2}z^2)d(s-z)dz\right), \tag{7.3}$$

and $\tilde{\mathbf{s}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{W}^i\mathbf{x}^i$ the source estimate. The negative log-likelihood becomes

$$\mathcal{L}(W^1,\dots,W^m) = -\sum_{i=1}^{m}\log|\mathbf{W}^i| + \frac{1}{2\sigma^2}\sum_{i=1}^{m}\|\mathbf{W}^i\mathbf{x}^i - \tilde{\mathbf{s}}\|^2 + f(\tilde{\mathbf{s}}). \tag{7.4}$$

Multiview ICA is then performed by minimizing $\mathcal{L}$, and the estimated shared sources are $\tilde{\mathbf{s}}$. For the reasons discussed in § 2.2.3—namely, that for the purpose of source separation slight misspecifications of the true source densities are inconsequential—we can in practice fix f to be some nongaussian log-pdf.

The negative log-likelihood $\mathcal{L}$ is quite simple, and importantly, can be computed easily given the parameters of the model and the data; it does not involve any intractable integral.

For one subject ($m = 1$), $\mathcal{L}(\mathbf{W}^1)$ simplifies to the (negative) log-likelihood of ICA Equation 2.6, and we recover Infomax [46, 47], where the source log-pdf is replaced with the smoothened f.

7.2.2 Alternate Quasi-Newton Method for MultiView ICA

The parameters of the model are estimated by minimizing $\mathcal{L}$. We propose a combination of quasi-Newton method and alternate minimization for this task. First, $\mathcal{L}$ is non-convex: it is only defined when the $\mathbf{W}^i$ are invertible, which is a non-convex set. Therefore, we only look for local minima as usual in ICA. We propose an alternate minimization scheme, where $\mathcal{L}$ is alternatively diminished with respect to each $\mathbf{W}^i$. When all matrices $\mathbf{W}^1,\dots,\mathbf{W}^m$ are fixed but one, $\mathbf{W}^i$, $\mathcal{L}$ can be rewritten, up to an additive constant, as

$$\begin{aligned}\mathcal{L}^i(\mathbf{W}^i) = &-\log|\mathbf{W}^i| + \frac{1-1/m}{2\sigma^2}\|\mathbf{W}^i\mathbf{x}^i - \frac{m}{m-1}\tilde{\mathbf{s}}^{-i}\|^2 + \\ &+ f\left(1/m\,\mathbf{W}^i\mathbf{x}^i + \tilde{\mathbf{s}}^{-i}\right),\end{aligned} \tag{7.5}$$

with $\tilde{\mathbf{s}}^{-i} = 1/m\sum_{j\neq i}\mathbf{W}^j\mathbf{x}^j$. This function has the same structure as the usual maximum-likelihood ICA cost function: it is written $\mathcal{L}^i(\mathbf{W}^i) = -\log|\mathbf{W}^i| + g(\mathbf{W}^i\mathbf{x}^i)$, where $g(\mathbf{y}) = \sum_{j=1}^{k} f(\frac{y_j}{m} + \tilde{\mathbf{s}}_j^{-i}) + \frac{1-1/m}{2\sigma^2}(y_j - \frac{m}{m-1}\tilde{\mathbf{s}}_j^{-i})^2$. Fast quasi-Newton algorithms [48, 285] have been proposed for minimizing such functions. We employ a similar technique as [285], which we now describe.

Quasi-Newton methods are based on approximations of the Hessian of $\mathcal{L}^i$. As we mentioned in Chapter 5, the *relative gradient* (resp. Hessian) of $\mathcal{L}^i$ is defined as the matrix $\mathbf{G}^i \in \mathbb{R}^{k\times k}$ (resp. tensor $\mathcal{H}^i \in \mathbb{R}^{k\times k\times k\times k}$) such that for an infinitesimal matrix $\boldsymbol{\epsilon} \in \mathbb{R}^{k\times k}$, we have

$$\mathcal{L}^i((\mathbf{I}_k + \boldsymbol{\epsilon})\mathbf{W}^i) \simeq \mathcal{L}^i(\mathbf{W}^i) + \langle \mathbf{G}^i, \mathbf{W}^i\rangle + \frac{1}{2}\langle \boldsymbol{\epsilon}, \mathcal{H}^i\boldsymbol{\epsilon}\rangle\,. \tag{7.6}$$

This kind of multiplicative perturbation of the unmixing matrices lies at the core of some classic approaches to linear ICA [181, 232]. For the gradient and Hessian, the expression in (7.6) yields[4]

$$\mathbf{G}^i = \frac{1}{m} f'(\tilde{\mathbf{s}})(\mathbf{y}^i)^\top + \frac{1 - 1/m}{\sigma^2}(\mathbf{y}^i - \frac{m}{m-1}\tilde{\mathbf{s}}^{-i})(\mathbf{y}^i)^\top - I_k, \text{ where } \mathbf{y}^i = \mathbf{W}^i\mathbf{x}^i \tag{7.7}$$

$$\mathscr{H}^i_{abcd} = \delta_{ad}\delta_{bc} + \delta_{ac}\left(\frac{1}{m^2} f''(\tilde{\mathbf{s}}_a) + \frac{1 - 1/m}{\sigma^2}\right)\mathbf{y}^i_b\mathbf{y}^i_d, \text{ for } a, b, c, d = 1 \dots k \tag{7.8}$$

Newton's direction is then $-\left(\mathscr{H}^i\right)^{-1}\mathbf{G}^i$. However, this Hessian is costly to compute (since it has $\simeq k^3$ non-zero coefficients) and invert (it can be seen as a big $k^2 \times k^2$ matrix). Furthermore, to enforce that Newton's direction is a descent direction, the Hessian matrix should be regularised in order to eliminate its negative eigenvalues [287], and $\mathscr{H}^i$ is not guaranteed to be positive definite. These obstacles render the computation of Newton's direction impractical.

Luckily, if we assume that the signals in $\mathbf{y}^i$ are independent, severall coefficients cancel, and the Hessian simplifies to the approximation

$$H^i_{abcd} = \delta_{ad}\delta_{bc} + \delta_{ac}\delta_{bd}\Gamma^i_{ab} \text{ with } \Gamma^i_{ab} = \left(\frac{1}{m^2} f''(\tilde{\mathbf{s}}_a) + \frac{1 - 1/m}{\sigma^2}\right)\left(\mathbf{y}^i_b\right)^2. \tag{7.9}$$

This approximation is sparse: it only has $k(2k-1)$ non-zero coefficients. In order to better understand the structure of the approximation, we can compute the matrix $\left(H^i\mathbf{M}\right)$ for $\mathbf{M} \in \mathbb{R}^{k\times k}$. We find $\left(H^i\mathbf{M}\right)_{ab} = \Gamma^i_{ab}M_{ab} + M_{ba}$: i.e., $(H^i\mathbf{M})_{ab}$ only depends on M_{ab} and M_{ba}, indicating a simple block diagonal structure of H^i. The tensor H^i is therefore easily regularised and inverted: $\left((H^i)^{-1}\mathbf{M}\right)_{ab} = \frac{\Gamma^i_{ba}M_{ab} - M_{ba}}{\Gamma^i_{ab}\Gamma^i_{ba} - 1}$. Finally, since this approximation is obtained by assuming that the $\mathbf{y}^i$ are independent, the direction $-\left(H^i\right)^{-1}\mathbf{G}^i$ is close to Newton's direction when the $\mathbf{y}^i$ are close to independence, leading to fast convergence. Algorithm 1 alternates one step of the quasi-Newton method for each subject until convergence. A backtracking line-search is used to ensure that each iteration leads to a decrease of $\mathscr{L}^i$. The algorithm is stopped when maximum norm of the gradients over one pass on each subject is below some tolerance level, indicating that the algorithm is close to a stationary point.

[181]: Amari et al. (1996), 'A new learning algorithm for blind signal separation'
[232]: Cardoso et al. (1996), 'Equivariant adaptive source separation'

4: See, e.g., [286].

Algorithm 1: Alternate quasi-Newton method for MultiView ICA

Input: Dataset $(\mathbf{x}^i)_{i=1}^m$, initial unmixing matrices $\mathbf{W}^i$, noise parameter σ, function f, tolerance ε

Set tol= $+\infty$, $\tilde{\mathbf{s}} = 1/m \sum_{i=1}^k \mathbf{W}^i \mathbf{x}^i$
while *tol*$> \varepsilon$ **do**
  tol = 0
  for $i = 1 \dots m$ **do**
    Compute $\mathbf{y}^i = \mathbf{W}^i \mathbf{x}^i$, $\tilde{\mathbf{s}}^{-i} = \tilde{\mathbf{s}} - \frac{1}{m}\mathbf{y}^i$, gradient $\mathbf{G}^i$ (eq. (7.7)) and Hessian H^i (eq. (7.9))
    Compute the search direction $\mathbf{S} = -\left(H^i\right)^{-1} \mathbf{G}^i$
    Find a step size ρ such that $\mathscr{L}^i((\mathbf{I}_k + \rho\mathbf{S})\mathbf{W}^i) < \mathscr{L}^i(\mathbf{W}^i)$ with line search
    Update $\tilde{\mathbf{s}} = \tilde{\mathbf{s}} + \frac{\rho}{m}\mathbf{S}\mathbf{W}^i\mathbf{x}^i$, $\mathbf{W}^i = (\mathbf{I}_k + \rho\mathbf{S})\mathbf{W}^i$, tol= max(tol, $\|\mathbf{G}^i\|$)
  end
end
return *Estimated unmixing matrices* $\mathbf{W}^i$, *estimated shared sources* $\tilde{\mathbf{s}}$

7.2.3 Robustness to Model Misspecification

Algorithm 1 has two hyperparameters: σ and the function f. The latter is usual for an ICA algorithm, but the former is not. We study the impact of these parameters on the separation capacity of the algorithm, when these parameters do not correspond to those of the generative model (7.1).

Proposition 7.2.2 *We consider the cost function $\mathscr{L}$ in eq. (7.4) with noise parameters σ and function f. Assume sub-linear growth on f': $|f'(x)| \leq c|x|^\alpha + d$ for some $c, d > 0$ and $0 < \alpha < 1$. Assume that $\mathbf{x}^i$ is generated following model (7.1), with noise parameter σ' and density of the source d' which need not be related to σ and f. Then, there exists a diagonal matrix $\mathbf{D}$ such that $(\mathbf{D}(\mathbf{A}^1)^{-1}, \dots, \mathbf{D}(\mathbf{A}^m)^{-1})$ is a stationary point of $\mathscr{L}$, that is $\mathbf{G}^1, \dots, \mathbf{G}^m = 0$ at this point.*

The sub-linear growth of f' is a customary hypothesis in ICA which implies that d has heavier-tails than a Gaussian, and in Appendix F.3.2 we provide other conditions for the result to hold. In this setting, the shared sources estimated by the algorithm are $\tilde{\mathbf{s}} = \mathbf{D}(\mathbf{s} + \frac{1}{m}\sum_{i=1}^m \mathbf{n}^i)$, which is a scaled version of the best estimate of the shared sources under the Gaussian noise hypothesis.

This proposition shows that, up to scale, the true unmixing matrices are a stationary point for Algorithm 1: if the algorithm starts at this point it will not move. The question of stability is also interesting: if the algorithm is initialised *close* to the true unmixing matrices, will it converge to the true unmixing matrix? In Appendix F.3.3, we provide an analysis similar to [45], and derive sufficient numerical conditions for the unmixing matrices to be local minima of $\mathscr{L}$. We also study the practical impact of changing the hyperparameter σ on the accuracy of a machine learning pipeline based on MultiviewICA on real fMRI data in the appendix Sec. F.5.5. As expected from the theoretical study, the performance of the algorithm is barely affected by σ.

7.2.4 Dimensionality Reduction

So far, we have assumed that the dimensionality of each view (subject) and that of the sources is the same. This reflects the standard practice in ICA of having equal number of observations and sources which we also discussed in § 1.3.3. In practice, however, we might want to estimate fewer sources than there are observations per view; the original dimensionality of the data might in practice not be computationally tractable. The problem of how to perform subject-wise dimensionality reduction in group studies is an interesting one *per se*, and out of the main scope of this work. For our purposes, it can be considered as a preprocessing step for which well-known various solutions can be applied. We discuss this further in § 7.3 and in Appendix F.6.

7.3 Related Work

Many methods for data-driven multivariate analysis of neuroimaging group studies have been proposed. We summarize the characteristics of some of the most commonly used ones. A more thorough description of these methods can be found in Appendix F.6. For completeness, we start by describing PCA. For a zero-mean data matrix $\mathbf{X}$ of size $p \times n$ with $p \leq n$, we denote $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ the singular value decomposition of $\mathbf{X}$ where $\mathbf{U} \in \mathbb{R}^{p \times p}$, $\mathbf{V} \in \mathbb{R}^{n \times p}$ are orthogonal and $\mathbf{D}$ the diagonal matrix of singular values ordered in decreasing order. The PCA of $\mathbf{X}$ with k components is $\mathbf{Y} \in \mathbb{R}^{k \times n}$ containing the first k rows of $\mathbf{D}\mathbf{V}^\top$, and it does not hold in general that $\mathbf{Y}\mathbf{Y}^\top = \mathbf{I}_k$: for the rest of the paper, what we call PCA does not include whitening of the signals.

Group ICA. When datasets are high-dimensional, a three steps procedure is often used: first dimensionality reduction is performed on data of each subject separately; then the reduced data are merged into a common representation; finally, an ICA algorithm is applied for shared source extraction. The merging of the reduced data is often done by PCA [288] or multi set CCA [289]. This is a popular method for fMRI [283] and EEG [290] group studies. These methods directly recover only group level, shared sources; when individual sources are needed, additional steps are required (back-projection [288] or dual-regression [291]). In contrast, MultiView ICA finds individual and shared independent components in a single step. Finally, in contrast to the methods described above, our method maximizes a likelihood, which brings statistical guarantees like consistency or asymptotic efficiency. The SR-ICA approach of [292] performs dimension reduction, merging and independent component estimation. It is therefore similar to our method. However, they propose to modify the FastICA algorithm [182] in a rather heuristic way, without specifying an optimization problem, let alone maximizing a likelihood. In the experiments on fMRI data in Appendix F.5.4, we obtain better performance with MultiView ICA than the reported performance of SR-ICA.

Likelihood-based models. One can consider the more general model $\mathbf{x}^i = \mathbf{A}^i\mathbf{s}^i + \mathbf{n}^i$, where the noise covariance can be learnt from the data [293]. The likelihood for this model involves an intractable high dimensional integral that is cumbersome to evaluate, and is then optimised with the Expectation-Maximization (EM) algorithm, which is known to converge slowly and unreliably [294, 295]. Having the simpler model (7.1) leads to a closed-form likelihood, that can then be optimised by more efficient means than the EM algorithm. In model (7.1), the noise can be interpreted as individual variability rather than sensor noise. In Appendix F.9, we generate data following model $\mathbf{x}^i = \mathbf{A}^i\mathbf{s}^i + \mathbf{n}^i$ and report the reconstruction error. The difference in performance between algorithms is small.

Structured mixing matrices. One strength of our model is that we only assume that the mixing matrices are invertible and still enjoy identifiability whereas some other approaches impose additional constraints. For instance tensorial methods [296] assume that the mixing matrices are the same up to diagonal scaling. Other methods impose a common mixing matrix [297–300]. Like PCA, the Shared Response Model [264] (SRM) assumes orthogonality of the mixing matrices. While the model defines a simple likelihood and provides an efficient way to reduce dimension, the SRM model is not identifiable as shown in Appendix F.4, and the orthogonal constraint may not be plausible.

Matching sources a posteriori. A different path to multi-subject ICA is to extract independent components with individual ICA in each subject and align them. We propose a simple baseline approach to do so called *PermICA*. Inspired by the heuristic of the hyperalignment method [265] we choose a reference subject and first match the sources of all other subjects to the sources of the reference subject. The process is then repeated multiple times, using the average of previously aligned sources as a reference. Finally, group sources are given by the average of all aligned sources. We use the Hungarian algorithm to align pairs of mixing matrices [301]. Alternative approaches involving clustering have also been developed [302, 303].

Deep Learning. Deep Learning methods, such as convolutional autoencoders (CAE), can also be used to find the subject specific unmixing [304]. While these nonlinear extensions of the aforementioned methods are interesting, these models are hard to train and interpret. In the experiments on fMRI data in Appendix F.5.4, we obtain better accuracy with MultiView ICA than that of CAE reported in [304].

Correlated component analysis. Other methods can be used to recover the shared neural responses such as the correlated component approach of Dmochowski [305]. We benchmark our method against its probabilistic version [306] called BCorrCA in Fig. 7.3. Our method yields much better results.

Autocorrelation. Another way to perform ICA is to leverage spectral diversity of the sources rather than non-Gaussianity. These methods are popular alternative to non-Gaussian ICA in the single-subject setting [42, 307, 308] and they output significantly different sources than non-Gaussian ICA [279]. Extensions to multiview problems have been proposed [309, 310].

7.4 Experiments

All code for the experiments is written in Python. We use Matplotlib for plotting [311] , scikit-learn for machine-learning pipelines [312], MNE for MEG processing [313], Nilearn for fMRI processing and for its CanICA implementation [314], Brainiak [315] for its SRM implementation. In the following, the noise parameter in MultiviewICA is always fixed to $\sigma = 1$. We use the function $f(\cdot) = \log\cosh(\cdot)$, giving the non-linearity $f'(\cdot) = \tanh(\cdot)$. We use the Infomax cost function [46] with the same non-linearity to perform standard ICA, with the Picard algorithm [48] for fast and robust minimization of the cost function. Picard is applied with the default hyper-parameters.[5]

5: The code for MultiViewICA is available online at `https://github.com/hugorichard/multiviewica`.

We compare the following methods to obtain k components: *GroupPCA* is PCA on spatially concatenated data. It corresponds to a transposed version of [316]. *PermICA* is described in the previous section. *SRM* is the algorithm of [264]. *GroupICA* is ICA applied after GroupPCA. *PCA+GroupICA* corresponds to GroupICA applied on subject data that have been first individually reduced by PCA with k components. These two approaches correspond to transposed versions of [299], and are similar to [290]. *CanICA* corresponds to PCA+GroupICA where the merging is done using multi set CCA rather than PCA. The dimension reduction in MultiView ICA and PermICA is performed with SRM in fMRI experiments and subject-specific PCA in MEG experiments. Initialization is discussed in Appendix F.2. A summary of our quantitative results on real data is available in Appendix F.10.

Synthetic experiment. We validate our method on synthetic data generated according to the model in equation (7.1). The sources are generated i.i.d. from a Laplace density $d(x) = \frac{1}{2}\exp(-|x|)$. The mixing matrices $\mathbf{A}^1, \ldots, \mathbf{A}^m$ are generated with i.i.d. entries following a normal law. Each compared algorithm returns a sequence of estimated unmixing matrices $\mathbf{W}^1, \ldots, \mathbf{W}^m$. The performance of an algorithm is measured by the reconstruction error between the estimated sources and the true sources. We use $m = 10$ datasets, $k = 15$ sources and $n = 1000$ samples. Each experiment is repeated with 100 random seeds. We vary the noise level in the data generation from 10^{-2} to 10.

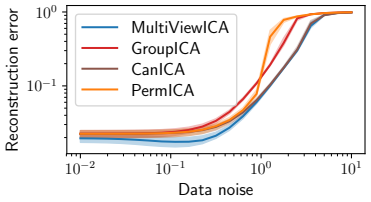


Figure 7.1: Synthetic experiment: reconstruction error of the algorithms on data following model (7.1).

Multiview ICA has uniformly better performance than the other algorithms, which illustrates the strength of maximum-likelihood based methods. In accordance with results of § 7.2, it is able to separate the sources even with misspecified noise parameter and source density.

fMRI data and preprocessing. We evaluate the performance of our approach on four different fMRI datasets. The *sherlock* dataset [317] contains recordings of 16 subjects watching an episode of the BBC TV show "Sherlock" (50 mins). The *forrest* dataset [318] was collected while 19 subjects were listening to an auditory version of the film "Forrest Gump" (110 mins). The *clips* dataset [319] was collected while 12 participants were exposed to short video clips (130 mins). The *raiders* dataset [319] was collected while 11 participants were watching the movie "Raiders of the Lost Ark" (110 mins). The *raiders-full* dataset [319] is an extension of the *raiders* dataset where the first two scenes of the movie are shown twice (130 mins). Like [292], we used full brain data. The rest of the preprocessing is identical to [317]. See F.5.1 for a detailed description of the datasets and preprocessing steps. Unless stated otherwise we use spatially unsmoothed data, except for the *sherlock* dataset, for which the available data are already preprocessed with a 6 mm spatial smoothing. All datasets are built from successive acquisitions called *runs* that typically last 10 minutes each. We define the chance level as the performance of an algorithm that computes unmixing matrices and projections to lower dimensional space by sampling random numbers from a standard normal distribution.

Reconstructing the BOLD signal of missing subjects. We want to show that once unmixing matrices have been learnt, they can be used to predict evoked responses across subjects, which can be useful to perform transfer learning [320]. We split the data into three groups. First, we randomly choose 80% of all runs from all subjects to form the training set. Then, we randomly choose 80% of subjects and take the remaining 20% runs as testing set. The left-out runs of the remaining 20% subjects form the validation set. The compared algorithms are run on the training set and evaluated using the testing and validation sets. After an algorithm is run on training data, it defines for each subject a *forward operator* that maps individual data to the source space and a *backward operator* that maps the source space to individual data. For instance in ICA the forward operator is the product of the dimensionality reduction projection and unmixing matrix. We estimate the shared responses on the testing set by applying the forward operators on the testing data and averaging. Finally, we reconstruct the individual data from subjects in the validation set by applying the backward operators to the shared responses. We measure the difference between the true signal and the reconstructed one using voxel-wise R^2 score. The R^2 score between two series $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$ is defined as $R^2(\mathbf{x}, \mathbf{y}) = 1 - \frac{1}{n \operatorname{Var}(\mathbf{y})} \sum_{t=1}^{n} (x_t - y_t)^2$, where $\operatorname{Var}(\mathbf{y}) = \frac{1}{n} \sum_{t=1}^{n} (y_t - \frac{1}{n} \sum_{t'=1}^{n} y_{t'})^2$ is the empirical variance of $\mathbf{y}$. The R^2 score is always smaller than 1, and equals 1 when $\mathbf{x} = \mathbf{y}$. The experiment is repeated 25 times with random splits to obtain error bars.

In this experiment we apply a 6 mm spatial smoothing to all datasets. The R^2 score per voxel depends heavily on which voxel is considered. For example voxels in the visual cortex are better reconstructed in the *sherlock* dataset than in the *forrest* dataset (see Fig. F.1 in Appendix F.5.2). In Fig. 7.2 (top) we plot the mean R^2 score inside a region of interest (ROI) in order to leave out regions where there is no useful information. ROIs are chosen based on the performance of GroupICA (more details in Appendix F.5.2). MultiView ICA has similar or better performance

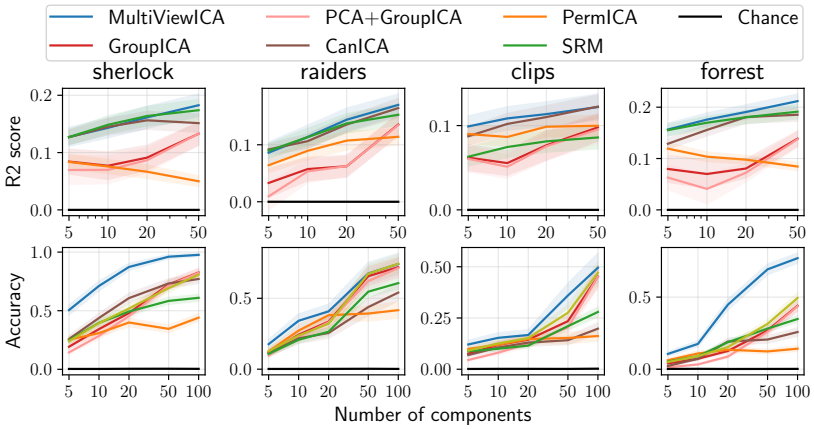


Figure 7.2: *Top:* **Reconstructing the BOLD signal of missing subjects**. Mean R^2 score between reconstructed data and true data (higher is better). *Bottom:* **Between subjects time-segment matching**. Mean classification accuracy. Error bars represent a 95 % confidence interval over cross validation splits.

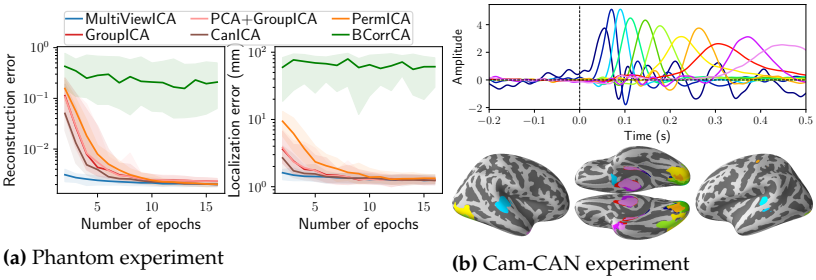
than the other methods on all datasets. This demonstrates its ability to capture inter-subject variability, making it a candidate of choice to handle missing data or perform transfer learning.

Between subjects time-segment matching. We reproduce the time-segment matching experiment of [264]. We split the runs into a train and test set. After fitting the model on the training set, we apply the forward operator of each subject on the test set yielding individual sources matrices. We estimate the shared responses by averaging the individual sources of each subjects but one. We select a target time-segment (9 consecutive timeframes) in the shared responses and try to localize the corresponding time segment in the sources of the left-out subject using a maximum-correlation classifier. This is a standard evaluation of SRM-like methods also used in [264], [321], [322] or [292]. The time-segment is said to be correctly classified if the correlation between the sample and target time-segment is higher than with any other time-segment (partially overlapping time windows are excluded). We use 5-Fold cross-validation across runs: the training set contains 80% of the runs and the test set 20%, and repeat the experiment using all possible choices for left-out subjects. The mean accuracy is reported in Fig. 7.2 (bottom). MultiView ICA yields a consistent and substantial improvement in accuracy compared to other methods on the four datasets. We see a marked improvement on the datasets sherlock and forrest. A possible explanation lies in the preprocessing pipeline. Sherlock data undergo a 6 mm spatial smoothing and Forrest data are acquired at a higher resolution (7T vs 3T for other data). This affects the signal to noise ratio. In Appendix F.5.5, we compute the accuracy of MultiviewICA on the sherlock dataset with 10 components when the noise parameter varies. MultiviewICA performs consistently well for a wide range of noise parameter values, and only breaks at very high values. It supports the theoretical claim of Prop 7.2.2 that the noise parameter is of little importance.

In Appendix F.5.3, we present a variation of this experiment. We measure the ability of each algorithm to extract meaningful shared sources that correlate more when they correspond to the same stimulus than when they correspond to distinct stimuli and show the improved performance of MultiView ICA. In Appendix F.8, we plot the average forward operator across subjects of MultiView ICA and GroupICA with 5 components on the forrest, sherlock, raiders and clips datasets.

Phantom MEG data. We demonstrate the usefulness of our approach on MEG data. The first experiment uses data collected with a realistic head phantom, which is a plastic device mimicking real electrical brain sources. Eight current dipoles positioned at different locations can be switched on or off. We view each dipole as a subject and therefore have $m = 8$. We only consider the 102 magnetometers. An epoch corresponds to 3 s of MEG signals where a dipole is switched on for 0.4 s with an oscillation at 20 Hz and a peak-to-peak amplitude of 200 nAm. This yields a matrix of size $p \times n$ where $p = 102$ is the number of sensors, and n is the number of time samples. We have access to 100 epochs per dipole. For each dipole, we chose $N_e = 2, \dots, 16$ epochs at random among our set of 100 epochs and concatenate them in the temporal dimension. We then apply algorithms on these data to extract $k = 20$ shared sources. As we know the true source (the timecourse of the dipole), we can compute the reconstruction error of each source as the squared norm of the difference between the estimated source and the true source, after normalization to unit variance and fixing the sign. We only retain the source of minimal error. We also estimate for each forward operator the localization of the source by performing dipole fitting using its column corresponding to the source of minimal error. We then compute the distance of the estimated dipole to the true dipole. These metrics are reported in Fig. 7.3 when the number of epochs considered N_e varies. We also compare our method to the Bayesian Canonical Correlation Analysis (BCorrCA) of [306]. On this task, BCorrCA is outperformed by ICA methods. MultiView ICA requires fewer epochs to correctly reconstruct and localize the true source.

Experiment on Cam-CAN dataset. Finally, we apply MultiView ICA on the Cam-CAN dataset [266]. We use the magnetometer data from the MEG of 200 subjects. Each subject is repeatedly presented an audio-visual stimulus. The MEG signal corresponding to these trials are then time-averaged to isolate the evoked response, yielding individual data. The MultiView ICA algorithm is then applied to extract 20 shared sources. 9 sources were found to correspond to noise by visual inspection, and the 11 remaining are displayed in Fig. 7.3. We observe that MultiView ICA recovers a very clean sequence of evoked potentials with sharp peaks for early components and slower responses for late components. In order to visualize their localization, we perform source localization for each subject by solving the inverse problem using sLORETA [323], providing a source estimate for each source. Then, we register each source estimate to a common reference brain. Finally, the source estimates are averaged, and thresholded maps are displayed in Fig. 7.3. Individual maps corresponding to each source are displayed in Appendix F.7. The figure highlights both early auditory and visual cortices, also suggesting a propagation of the activity towards the ventral regions and higher level visual areas.



(a) Phantom experiment

(b) Cam-CAN experiment

Figure 7.3: *Left:* **Experiment on MEG Phantom data**. Reconstruction error is the norm of the difference between the estimated and true source. Localization error is the distance between the estimated and true dipole. *Right:* **Experiment on 200 subjects from the CAM-can dataset** *Top:* Time course of 11 shared sources (one color per source). We recover clean evoked potentials. *Bottom:* Associated brain maps, obtained by averaging source estimates registered to a common reference.

7.5 Conclusion

In this chapter, we described an unsupervised algorithm that reveals latent sources observed through different views. The model is similar to the one introduced in Chapter 6, but its identifiability can be proved under much milder assumption due to linearity of the mixing. In contrast to previous approaches, the proposed model leads to a closed-form likelihood, which we then optimize efficiently using a dedicated alternate quasi-Newton approach. Our approach enjoys the statistical guarantees of maximum-likelihood theory, while still being tractable. We demonstrated the usefulness of MultiView ICA for neuroimaging group studies both on fMRI and MEG data, where it outperforms other methods. In the experiments on fMRI data, we used temporal ICA in order to make use of the fact that subjects were exposed to the same stimuli. However, applying MultiViewICA on transposed data would carry out spatial ICA. Therefore MultiViewICA can be readily used to analyse different kind of neuroimaging data such as resting state data. Our method is not specific to neuroimaging data and could be relevant to other observational sciences like genomics or astrophysics where ICA is already widely used.

Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style

8

Self-supervised representation learning has shown remarkable success in a number of domains. A common practice is to perform data augmentation via hand-crafted transformations intended to leave the semantics of the data invariant. We seek to understand the empirical success of this approach from a theoretical perspective. We formulate the augmentation process as a latent variable model by postulating a partition of the latent representation into a *content* component, which is assumed invariant to augmentation, and a *style* component, which is allowed to change. Unlike prior work on disentanglement and independent component analysis, we allow for both nontrivial statistical and causal dependencies in the latent space. We study the identifiability of the latent representation based on pairs of views of the observations and prove sufficient conditions that allow us to identify the invariant content partition up to an invertible mapping in both generative and discriminative settings. We find numerical simulations with dependent latent variables are consistent with our theory. Lastly, we introduce *Causal3DIdent*, a dataset of high-dimensional, visually complex images with rich causal dependencies, which we use to study the effect of data augmentations performed in practice.

8.1 Introduction

Over the last decade, *self-supervised learning* (SSL) has emerged as one of the dominant paradigms for learning good representations of high-dimensional observations [18, 324–335]. The main idea behind SSL is to extract a supervisory signal from unlabelled observations by leveraging known structure of the data, which allows for the application of supervised learning techniques—in a similar spirit to what we discussed in § 2.4.2.

A common approach is to directly predict some part of the observation from another part (e.g., future from past, or original from corruption), thus forcing the model to learn a meaningful representation in the process. While this technique has shown remarkable success in natural language processing [336–343] and speech recognition [344–347], where a finite dictionary allows one to output a distribution over the missing part, such *predictive* SSL methods are not easily applied to continuous or high-dimensional domains such as vision. Here, a common approach is to learn a *joint embedding* of similar observations or *views* such that their representation is close [17, 348–350]. This multiview setting is related to the ones discussed in Chapter 6 and Chapter 7: different views can come, for example, from different modalities (text & speech; video & audio) or time points. However, still images lack such multi-modality or temporal structure: recent advances in representation learning have therefore relied on generating similar views by means of *data augmentation*.

In order to be useful, data augmentation is thought to require the transformations applied to generate additional views to be generally chosen to *preserve the semantic characteristics* of an observation, while

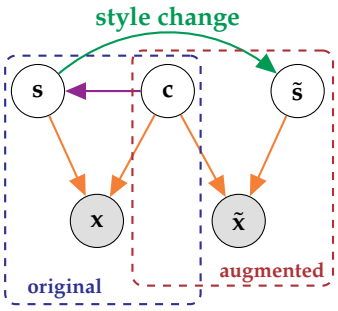


Figure 8.1: Overview of our problem formulation. We partition the latent variable $\mathbf{z}$ into content $\mathbf{c}$ and style $\mathbf{s}$, and allow for statistical and causal dependence of style on content. We assume that only style changes between the original view $\mathbf{x}$ and the augmented view $\tilde{\mathbf{x}}$, i.e., they are obtained by applying the same deterministic function $\mathbf{f}$ to $\mathbf{z} = (\mathbf{c}, \mathbf{s})$ and $\tilde{\mathbf{z}} = (\mathbf{c}, \tilde{\mathbf{s}})$.

changing other "nuisance" aspects. While this intuitively makes sense and has shown remarkable empirical results, the success of data augmentation techniques in practice is still not very well understood from a theoretical perspective—despite some efforts [351–353]. In the present chapter, we seek to better understand the empirical success of SSL with data augmentation by formulating the generative process as a latent variable model (LVM) and studying identifiability of the representation.

Related work and its relation to this Chapter. Prior work on unsupervised representation learning from an LVM perspective often postulates *mutually independent latent factors*: this independence assumption is, for example, at the heart of independent ICA, as we reviewed in Chapter 2. In works using auxiliary variables (§ 2.4.2) or multiple views (Chapter 6) to identify the individual independent latent factors, it is typically assumed that there is a chance that *each factor changes* across views, environments, or time points.

Our contribution—being directly motivated by common practices in SSL with data augmentation—differs from these works in the following two key aspects (see Fig. 8.1 for an overview). First, we do not assume independence and instead *allow for both nontrivial statistical and causal relations between latent variables*. This is in line with a recently proposed [354] shift towards causal representation learning [13, 63, 146, 148, 149, 355–358], motivated by the fact that many underlying variables of interest may not be independent but causally related to each other.[1] Second, instead of a scenario wherein all latent factors may change as a result of augmentation, we assume a *partition of the latent space* into two blocks: a *content* block which is shared or *invariant* across different augmented views, and a *style* block that *may change*. This is aligned with the notion that augmentations leave certain semantic aspects (i.e., content) intact and only affect style, and is thus a more appropriate assumption for studying SSL. In line with earlier work [62, 64, 66, 70, 72, 83, 88, 114, 359], we focus on the setting of continuous ground-truth latents, though we believe our results to hold more broadly.

[354]: Schölkopf (2019), 'Causality for machine learning'

1: E.g., [359], Fig. 11 where dependence between latents was demonstrated for multiple natural video data sets.

Structure and contributions of this Chapter. Following a review of SSL with data augmentation (§ 8.2), we formalise the process of data generation and augmentation as an LVM with content and style variables (§ 8.3). We then establish identifiability results of the invariant content partition (§ 8.4), validate our theoretical insights experimentally (§ 8.5), and discuss our findings and their limitations in the broader context of SSL with data augmentation (§ 8.6). We highlight the following contributions:

- we prove that SSL with data augmentations identifies the invariant content partition of the representation in generative (Thm. 8.4.1) and discriminative learning with invertible (Thm. 8.4.2) and non-invertible encoders with entropy regularisation (Thm. 8.4.3); in particular, Thm. 8.4.3 provides a theoretical justification for the empirically observed effectiveness of contrastive SSL methods that use data augmentation and InfoNCE [325] as an objective, such as `SimCLR` [329];
- we show that our theory is consistent with results in simulating statistical dependencies within blocks of content and style variables, as well as with style causally dependent on content (§ 8.5.1);
- we introduce *Causal3DIdent*, a dataset of 3D objects which allows for the study of identifiability in a causal representation learning setting, and use it to perform a systematic study of data augmentations used in practice, yielding novel insights on what particular data augmentations are truly isolating as invariant content and discarding as varying style when applied (§ 8.5.2).

8.2 Preliminaries and Background

Self-supervised representation learning with data augmentation. Given an unlabelled dataset of observations (e.g., images) $\mathbf{x}$, data augmentation techniques proceed as follows. First, a set of observation-level transformations $\mathbf{t} \in \mathcal{T}$ are specified together with a distribution $p_{\mathbf{t}}$ over $\mathcal{T}$. Both $\mathcal{T}$ and $p_{\mathbf{t}}$ are typically designed using human intelligence and domain knowledge with the intention of *not changing the semantic characteristics* of the data (which arguably constitutes a form of weak supervision).[2] For images, for example, a common choice for $\mathcal{T}$ are combinations of random crops [362], horizontal or vertical flips, blurring, colour distortion [362, 363], or cutouts [364]; and $p_{\mathbf{t}}$ is a distribution over the parameterisation of these transformations, e.g., the centre and size of a crop [329, 364]. For each observation $\mathbf{x}$, a pair of transformations $\mathbf{t}, \mathbf{t}' \sim p_{\mathbf{t}}$ is sampled and applied separately to $\mathbf{x}$ to generate a pair of augmented views $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') = (\mathbf{t}(\mathbf{x}), \mathbf{t}'(\mathbf{x}))$.

2: Note that recent work has investigated automatically discovering good augmentations [360, 361].

The joint-embedding approach to SSL then uses a pair of encoder functions $(\mathbf{g}, \mathbf{g}')$, i.e. deep nets, to map the pair $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$ to a typically lower-dimensional representation $(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}') = (\mathbf{g}(\tilde{\mathbf{x}}), \mathbf{g}'(\tilde{\mathbf{x}}'))$. Often, the two encoders are either identical, $\mathbf{g} = \mathbf{g}'$, or directly related (e.g., via shared parameters or asynchronous updates). Then, the encoder(s) $(\mathbf{g}, \mathbf{g}')$ are trained such that the representations $(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}')$ are "close", i.e., such that $\mathrm{sim}(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}')$ is large for some similarity metric $\mathrm{sim}(\cdot)$, e.g., the cosine similarity [88, 329], or negative L2 norm [88]. The advantage of directly optimising for similarity in representation space over generative alternatives is that reconstruction can be very challenging for high-dimensional data. The disadvantage is the problem of *collapsed representations*.[3] To avoid collapsed representations and force the encoder(s) to learn a meaningful representation, two main families of approaches have been used: (i) *contrastive learning* (CL) [324–329]; and (ii) *regularisation-based* SSL [330, 331, 365].

3: If the only goal is to make representations of augmented views similar, a degenerate solution which simply maps any observation to the origin trivially achieves this goal.

The idea behind CL is to not only learn similar representations for augmented views $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}'_i)$ of the same $\mathbf{x}_i$, or *positive pairs*, but to also use

other observations $\mathbf{x}_j$ $(j \neq i)$ to contrast with, i.e., to enforce a dissimilar representation across *negative pairs* $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}'_j)$. In other words, CL pulls representations of positive pairs together, and pushes those of negative pairs apart. Since both aims cannot be achieved simultaneously with a constant representation, collapse is avoided. A popular CL objective function (used, e.g., in `SimCLR` [329]) is InfoNCE [325] (based on noise-contrastive estimation [80, 366]):

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{g}; \tau, K) = \mathbb{E}_{\{\mathbf{x}_i\}_{i=1}^{K} \sim p_{\mathbf{x}}} \left[-\sum_{i=1}^{K} \log \frac{\exp\{\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_i)/\tau\}}{\sum_{j=1}^{K} \exp\{\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_j)/\tau\}} \right] \tag{8.1}$$

where $\tilde{\mathbf{z}} = \mathbb{E}_{\mathbf{t} \sim p_{\mathbf{t}}}[\mathbf{g}(\mathbf{t}(\mathbf{x}))]$, τ is a temperature, and $K - 1$ is the number of negative pairs. InfoNCE (8.1) has an interpretation as multi-class logistic regression, and lower bounds the mutual information across similar views $(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}')$—a common representation learning objective [46, 47, 367–373]. Moreover, (8.1) can be interpreted as *alignment* (numerator) and *uniformity* (denominator) terms, the latter constituting a nonparametric entropy estimator of the representation as $K \to \infty$ [374]. CL with InfoNCE can thus be seen as alignment of positive pairs with (approximate) entropy regularisation.

Instead of using negative pairs, as in CL, a set of recent SSL methods only optimise for alignment and avoid collapsed representations through different forms of regularisation. For example, `BYOL` [330] and `SimSiam` [331] rely on "architectural regularisation" in the form of moving-average updates for a separate "target" net $\mathbf{g}'$ (`BYOL` only) or a stop-gradient operation (both). `BarlowTwins` [365], on the other hand, optimises the cross correlation between $(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}')$ to be close to the identity matrix, thus enforcing redundancy reduction (zero off-diagonals) in addition to alignment (ones on the diagonal).

8.3 Problem Formulation

We specify our problem setting by formalising the processes of data generation and augmentation. We take a latent-variable model perspective and assume that observations $\mathbf{x}$ (e.g., images) are generated by a *mixing* function $\mathbf{f}$ which takes a latent code $\mathbf{z}$ as input. Importantly, we describe the augmentation process through changes in this latent space as captured by a conditional distribution $p_{\tilde{\mathbf{z}}|\mathbf{z}}$, as opposed to traditionally describing the transformations $\mathbf{t}$ as acting directly at the observation level.

Formally, let $\mathbf{z}$ be a continuous r.v. taking values in an open, simply-connected n-dim. *representation space* $\mathcal{Z} \subseteq \mathbb{R}^n$ with associated probability density $p_{\mathbf{z}}$. Moreover, let $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ be a *smooth and invertible* mapping to an *observation space* $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\mathbf{x}$ be the continuous r.v. defined as $\mathbf{x} = \mathbf{f}(\mathbf{z})$.[4] The generative process for the dataset of original observations of $\mathbf{x}$ is thus given by:

$$\mathbf{z} \sim p_{\mathbf{z}}, \qquad \mathbf{x} = \mathbf{f}(\mathbf{z}). \tag{8.2}$$

Next, we formalise the data augmentation process. As stated above, we take a representation-centric view, i.e., we assume that an augmentation $\tilde{\mathbf{x}}$ of the original $\mathbf{x}$ is obtained by applying the same mixing or rendering

4: While $\mathbf{x}$ may be high-dimensional $n \ll d$, invertibility of $\mathbf{f}$ implies that $\mathcal{X}$ is an n-dim. sub-manifold of $\mathbb{R}^d$.

function $\mathbf{f}$ to a modified representation $\tilde{\mathbf{z}}$ which is (stochastically) related to the original representation $\mathbf{z}$ of $\mathbf{x}$. Specifying the effect of data augmentation thus corresponds to specifying a conditional distribution $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ which captures the relation between $\mathbf{z}$ and $\tilde{\mathbf{z}}$.

In terms of the transformation-centric view presented in § 8.2, we can view the modified representation $\tilde{\mathbf{z}} \in \mathcal{Z}$ as obtained by applying $\mathbf{f}^{-1}$ to a transformed observation $\tilde{\mathbf{x}} = \mathbf{t}(\mathbf{x}) \in \mathcal{X}$ where $\mathbf{t} \sim p_{\mathbf{t}}$, i.e., $\tilde{\mathbf{z}} = \mathbf{f}^{-1}(\tilde{\mathbf{x}})$. The conditional distribution $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ in the representation space can thus be viewed as being induced by the distribution $p_{\mathbf{t}}$ over transformations applied at the observation level.[5]

5: We investigate this correspondence between changes in observation and latent space empirically in § 8.5.

We now encode the notion that the set of transformations $\mathcal{T}$ used for augmentation is typically chosen such that any transformation $\mathbf{t} \in \mathcal{T}$ leaves certain aspects of the data invariant. To this end, we assume that *the representation* $\mathbf{z}$ *can be uniquely partitioned into two disjoint parts*:

(i) an *invariant* part $\mathbf{c}$ which will *always be shared* across $(\mathbf{z}, \tilde{\mathbf{z}})$, and which we refer to as *content*;
(ii) a *varying* part $\mathbf{s}$ which *may change* across $(\mathbf{z}, \tilde{\mathbf{z}})$, and which we refer to as *style*.

We assume that $\mathbf{c}$ and $\mathbf{s}$ take values in content and style subspaces $\mathcal{C} \subseteq \mathbb{R}^{n_c}$ and $\mathcal{S} \subseteq \mathbb{R}^{n_s}$, respectively, i.e., $n = n_c + n_s$ and $\mathcal{Z} = \mathcal{C} \times \mathcal{S}$. W.l.o.g., we let $\mathbf{c}$ corresponds to the first n_c dimensions of $\mathbf{z}$:

$$\mathbf{z} = (\mathbf{c}, \mathbf{s}), \qquad \mathbf{c} := \mathbf{z}_{1:n_c}, \qquad \mathbf{s} := \mathbf{z}_{(n_c+1):n},$$

We formalise the process of data augmentation with content-preserving transformations by defining the conditional $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ such that only a (random) subset of the style variables change at a time.

Assumption 8.3.1 (Content-invariance) *The conditional density $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ over $\mathcal{Z} \times \mathcal{Z}$ takes the form*

$$p_{\tilde{\mathbf{z}}|\mathbf{z}}(\tilde{\mathbf{z}}|\mathbf{z}) = \delta(\tilde{\mathbf{c}} - \mathbf{c}) p_{\tilde{\mathbf{s}}|\mathbf{s}}(\tilde{\mathbf{s}}|\mathbf{s})$$

for some continuous density $p_{\tilde{\mathbf{s}}|\mathbf{s}}$ on $\mathcal{S} \times \mathcal{S}$, where $\delta(\cdot)$ is the Dirac delta function, i.e., $\tilde{\mathbf{c}} = \mathbf{c}$ a.e.

Assumption 8.3.2 (Style changes) *Let $\mathcal{A}$ be the set of subsets of style variables $A \subseteq \{1, \ldots, n_s\}$ and let p_A be a distribution on $\mathcal{A}$. Then, the style conditional $p_{\tilde{\mathbf{s}}|\mathbf{s}}$ is obtained via*

$$A \sim p_A, \qquad p_{\tilde{\mathbf{s}}|\mathbf{s},A}(\tilde{\mathbf{s}}|\mathbf{s}, A) = \delta(\tilde{\mathbf{s}}_{A^c} - \mathbf{s}_{A^c}) p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\tilde{\mathbf{s}}_A|\mathbf{s}_A)\,,$$

where $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ is a continuous density on $\mathcal{S}_A \times \mathcal{S}_A$, $\mathcal{S}_A \subseteq \mathcal{S}$ denotes the subspace of changing style variables specified by A, and $A^c = \{1, \ldots, n_s\} \setminus A$ denotes the complement of A.

Note that Assumption 8.3.2 is less restrictive than assuming that all style variables need to change, since it also allows for only a (possibly different) subset of style variables to change for any given observation. This is in line with the intuition that not all transformations affect all changeable (i.e., style) properties of the data: e.g., a colour distortion should not affect

positional information, and, in the same vein, a (horizontal or vertical) flip should not affect the colour spectrum.

The generative process of an augmentation or transformed observation $\tilde{\mathbf{x}}$ is thus given by

$$A \sim p_A, \qquad \tilde{\mathbf{z}}|\mathbf{z}, A \sim p_{\tilde{\mathbf{z}}|\mathbf{z},A}, \qquad \tilde{\mathbf{x}} = \mathbf{f}(\tilde{\mathbf{z}}). \tag{8.3}$$

Our setting for modelling data augmentation differs from that commonly assumed in (multi-view) disentanglement and ICA in that *we do not assume that the latent factors* $\mathbf{z} = (\mathbf{c}, \mathbf{s})$ *are mutually (or conditionally) independent*, i.e., we allow for *arbitrary* (non-factorised) marginals $p_{\mathbf{z}}$ in (8.2).[6]

6: The recently proposed Independently *Modulated* Component Analysis (IMCA) [85] extension of ICA is a notable exception, but only allows for trivial dependencies across $\mathbf{z}$ in the form of a shared base measure.

Causal interpretation: data augmentation as counterfactuals under soft style intervention. We now provide a causal account of the above data generating process by describing the (allowed) causal dependencies among latent variables using a structural causal model (SCM) [122]. As we will see, this leads to an interpretation of data augmentations as counterfactuals in the underlying latent SCM. The assumption that $\mathbf{c}$ stays invariant as $\mathbf{s}$ changes is consistent with the view that content may causally influence style, $\mathbf{c} \to \mathbf{s}$, but not vice versa, see Fig. 8.1. We therefore formalise their relation as:

$$\mathbf{c} := \mathbf{f}_{\mathbf{c}}(\mathbf{u}_{\mathbf{c}}), \qquad \mathbf{s} := \mathbf{f}_{\mathbf{s}}(\mathbf{c}, \mathbf{u}_{\mathbf{s}}), \qquad (\mathbf{u}_{\mathbf{c}}, \mathbf{u}_{\mathbf{s}}) \sim p_{\mathbf{u}_{\mathbf{c}}} \times p_{\mathbf{u}_{\mathbf{s}}}$$

where $\mathbf{u}_{\mathbf{c}}, \mathbf{u}_{\mathbf{s}}$ are independent exogenous variables, and $\mathbf{f}_{\mathbf{c}}, \mathbf{f}_{\mathbf{s}}$ are deterministic functions. The latent causal variables $(\mathbf{c}, \mathbf{s})$ are subsequently decoded into observations $\mathbf{x} = \mathbf{f}(\mathbf{c}, \mathbf{s})$. Given a factual observation $\mathbf{x}^{\mathsf{F}} = \mathbf{f}(\mathbf{c}^{\mathsf{F}}, \mathbf{s}^{\mathsf{F}})$ which resulted from $(\mathbf{u}_{\mathbf{c}}^{\mathsf{F}}, \mathbf{u}_{\mathbf{s}}^{\mathsf{F}})$, we may ask the counterfactual question: "*what would have happened if the style variables had been (randomly) perturbed, all else being equal?*". Consider, e.g., a *soft intervention* [375] on $\mathbf{s}$, i.e., an intervention that changes the mechanism $\mathbf{f}_{\mathbf{s}}$ to

$$do(\mathbf{s} := \tilde{\mathbf{f}}_{\mathbf{s}}(\mathbf{c}, \mathbf{u}_{\mathbf{s}}, \mathbf{u}_A)),$$

where $\mathbf{u}_A$ is an additional source of stochasticity accounting for the randomness of the augmentation process $(p_A \times p_{\tilde{\mathbf{s}}|\mathbf{s},A})$. The resulting distribution over counterfactual observations $\mathbf{x}^{\mathsf{CF}} = \mathbf{f}(\mathbf{c}^{\mathsf{F}}, \mathbf{s}^{\mathsf{CF}})$ can be computed from the modified SCM by fixing the exogenous variables to their factual values and performing the soft intervention:

$$\mathbf{c}^{\mathsf{CF}} := \mathbf{c}^{\mathsf{F}}, \qquad \mathbf{s}^{\mathsf{CF}} := \tilde{\mathbf{f}}_{\mathbf{s}}(\mathbf{c}^{\mathsf{F}}, \mathbf{u}_{\mathbf{s}}^{\mathsf{F}}, \mathbf{u}_A), \qquad \mathbf{u}_A \sim p_{\mathbf{u}_A}.$$

This aligns with our intuition and assumed problem setting of data augmentations as style corruptions. We note that the notion of augmentation as (hard) style interventions is also at the heart of `ReLIC` [357], a recently proposed, causally-inspired SSL regularisation term for instance-discrimination [324, 348]. However, `ReLIC` assumes independence between content and style and does not address identifiability. For another causal perspective on data augmentation in the context of domain generalisation, c.f. [376].

8.4 Theory: Block-Identifiability of the Invariant Content Partition

Our goal is to prove that we can identify the invariant content partition $\mathbf{c}$ under a distinct, weaker set of assumptions, compared to existing results in nonlinear ICA and disentanglement. We stress again that our primary interest is not to identify or disentangle individual (and independent) latent factors, unlike what discussed in § 2.3 and Defn. 2.4.1. Instead, our aim is to separate content from style, such that the content variables can be subsequently used for downstream tasks. We first define this distinct notion of *block-identifiability*.

Definition 8.4.1 (Block-identifiability) *We say that the true content partition* $\mathbf{c} = \mathbf{f}^{-1}(\mathbf{x})_{1:n_c}$ *is* block-identified *by a function* $\mathbf{g} : \mathcal{X} \to \mathcal{Z}$ *if the inferred content partition* $\hat{\mathbf{c}} = \mathbf{g}(\mathbf{x})_{1:n_c}$ *contains* all *and* only *information about* $\mathbf{c}$*, i.e., if there exists an invertible function* $\mathbf{h} : \mathbb{R}^{n_c} \to \mathbb{R}^{n_c}$ *s.t.* $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{c})$.

Defn. 8.4.1 is related to independent subspace analysis [377–380], which also aims to identify blocks of random variables as opposed to individual factors, though under an *independence assumption across blocks*, and typically not within a multi-view setting as studied in the present work.

8.4.1 Generative Self-Supervised Representation Learning

First, we consider *generative* SSL, i.e., fitting a generative model to pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ of original and augmented views.[7] We show that under our specified data generation and augmentation process (§ 8.3), as well as suitable additional assumptions (stated and discussed in more detail below), it is possible to isolate (i.e., block-identify) the invariant content partition. Full proofs are included in Appendix G.1.

7: For notational simplicity, we present our theory for pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ rather than for two augmented views $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$, as typically used in practice but it also holds for the latter, see § 8.6 for further discussion.

Theorem 8.4.1 (Identifying content with a generative model) *Consider the data generating process described in § 8.3, i.e., the pairs* $(\mathbf{x}, \tilde{\mathbf{x}})$ *of original and augmented views are generated according to* (8.2) *and* (8.3) *with* $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ *as defined in Assumptions 8.3.1 and 8.3.2. Assume further that*

(i) $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ *is smooth and invertible with smooth inverse (i.e., a diffeomorphism);*
(ii) $p_{\mathbf{z}}$ *is a smooth, continuous density on* $\mathcal{Z}$ *with* $p_{\mathbf{z}}(\mathbf{z}) > 0$ *almost everywhere;*
(iii) for any $l \in \{1, \dots, n_s\}$, $\exists A \subseteq \{1, \dots, n_s\}$ *s.t.* $l \in A$; $p_A(A) > 0$; $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ *is smooth w.r.t. both* $\mathbf{s}_A$ *and* $\tilde{\mathbf{s}}_A$*; and for any* $\mathbf{s}_A$, $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\cdot|\mathbf{s}_A) > 0$ *in some open, non-empty subset containing* $\mathbf{s}_A$.

If, for a given n_s $(1 \leq n_s < n)$*, a generative model* $(\hat{p}_{\mathbf{z}}, \hat{p}_A, \hat{p}_{\tilde{\mathbf{s}}|\mathbf{s},A}, \hat{\mathbf{f}})$ *assumes the same generative process (§ 8.3), satisfies the above assumptions (i)-(iii), and matches the data likelihood,*

$$p_{\mathbf{x},\tilde{\mathbf{x}}}(\mathbf{x}, \tilde{\mathbf{x}}) = \hat{p}_{\mathbf{x},\tilde{\mathbf{x}}}(\mathbf{x}, \tilde{\mathbf{x}}) \qquad \forall (\mathbf{x}, \tilde{\mathbf{x}}) \in \mathcal{X} \times \mathcal{X},$$

then it block-identifies the true content variables via $\mathbf{g} = \hat{\mathbf{f}}^{-1}$ *in the sense of Defn. 8.4.1.*

Proof sketch. First, show (using *(i)* and the matching likelihoods) that the representation $\hat{\mathbf{z}} = \mathbf{g}(\mathbf{x})$ extracted by $\mathbf{g}$ is related to the true $\mathbf{z}$ by a smooth invertible mapping $\mathbf{h} = \mathbf{g} \circ \mathbf{f}$ such that $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{z})_{1:n_c}$ is invariant across $(\mathbf{z}, \tilde{\mathbf{z}})$ almost surely w.r.t. $p_{\mathbf{z},\tilde{\mathbf{z}}}$.[8] Second, show by contradiction (using *(ii)*, *(iii)*) that $\mathbf{h}(\cdot)_{1:n_c}$ can, in fact, only depend on the true content $\mathbf{c}$ and not on style $\mathbf{s}$, for otherwise the invariance from step 1 would be violated in a region of the style (sub)space of measure greater than zero.

8: This step is partially inspired by [114]; the technique used to prove the second *main* step is entirely novel.

Intuition. Thm. 8.4.1 assumes that the number of content (n_c) and style (n_s) variables is known, and that there is a positive probability that each *style* variable may change, though not necessarily on its own, according to *(iii)*. In this case, training a generative model of the form specified in § 8.3 (i.e., with an invariant content partition and subsets of changing style variables) by maximum likelihood on pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ will asymptotically (in the limit of infinite data) recover the true invariant content partition up to an invertible function, i.e., it isolates, or unmixes, content from style.

Discussion. The identifiability result of Thm. 8.4.1 for generative SSL is of potential relevance for existing variational autoencoder (VAE) [191] variants such as the `GroupVAE` [381],[9] or its adaptive version `AdaGVAE` [114]. Since, contrary to existing results, Thm. 8.4.1 does not assume independent latents, it may also provide a principled basis for generative causal representation learning algorithms [148, 149, 355]. However, an important limitation to its practical applicability is that generative modelling does not tend to scale very well to complex high-dimensional observations, such as images.

9: which also uses a fixed content-style partition for multi-view data, but assumes that all latent factors are mutually independent, and that all style variables change between views, independent of the original style;

8.4.2 Discriminative Self-Supervised Representation Learning

We therefore next turn to a discriminative approach, i.e., directly learning an encoder function $\mathbf{g}$ which leads to a similar embedding across $(\mathbf{x}, \tilde{\mathbf{x}})$. As discussed in § 8.2, this is much more common for SSL with data augmentations. First, we show that if an invertible encoder $\mathbf{g}$ is used, then learning a representation which is aligned in the first n_c dimensions is sufficient to block-identify content.

Theorem 8.4.2 (Identifying content with an invertible encoder) *Assume the same data generating process (§ 8.3) and conditions (i)-(iv) as in Thm. 8.4.1. Let* $\mathbf{g} : \mathcal{X} \to \mathcal{Z}$ *be any smooth and* invertible *function which minimises the following functional:*

$$\mathcal{L}_{\text{Align}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathbf{x},\tilde{\mathbf{x}}}} \left[\left\| \mathbf{g}(\mathbf{x})_{1:n_c} - \mathbf{g}(\tilde{\mathbf{x}})_{1:n_c} \right\|_2^2 \right] \tag{8.4}$$

Then $\mathbf{g}$ *block-identifies the true content variables in the sense of Definition 8.4.1.*

Proof sketch. First, we show that the global minimum of (8.4) is reached by the smooth invertible function $\mathbf{f}^{-1}$. Thus, any other minimiser $\mathbf{g}$ must satisfy the same invariance across $(\mathbf{x}, \tilde{\mathbf{x}})$ used in step 1 of the proof of Thm. 8.4.1. The second step uses the same argument by contradiction as in Thm. 8.4.1.

Intuition. Thm. 8.4.2 states that if—under the same assumptions on the generative process as in Thm. 8.4.1—we directly learn a representation with an *invertible* encoder, then enforcing alignment between the first n_c latents is sufficient to isolate the invariant content partition. Intuitively, invertibility guarantees that all information is preserved, thus avoiding a collapsed representation.

Discussion. According to Thm. 8.4.2, content can be isolated if, e.g., a flow-based architecture [97, 99, 223, 227, 247] is used, or invertibility is enforced otherwise during training [31, 382]. However, the applicability of this approach is limited as it *places strong constraints on the encoder architecture which makes it hard to scale these methods up to high-dimensional settings*. As discussed in § 8.2, state-of-the-art SSL methods such as `SimCLR` [329], `BYOL` [330], `SimSiam` [331], or `BarlowTwins` [365] do not use invertible encoders, but instead avoid collapsed representations—which would result from naively optimising (8.4) for arbitrary, non-invertible $\mathbf{g}$—using different forms of regularisation.

To close this gap between theory and practice, finally, we investigate how to block-identify content without assuming an invertible encoder. We show that, if we add a regularisation term to (8.4) that encourages maximum entropy of the learnt representation, the invertibility assumption can be dropped.

Theorem 8.4.3 (Identifying content with discriminative learning and a non-invertible encoder) *Assume the same data generating process (§ 8.3) and conditions (i)-(iv) as in Thm. 8.4.1. Let $\mathbf{g} : \mathcal{X} \to (0,1)^{n_c}$ be any smooth function which minimises the following functional:*

$$\mathcal{L}_{\text{AlignMaxEnt}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathbf{x},\tilde{\mathbf{x}}}} \left[\left\| \mathbf{g}(\mathbf{x}) - \mathbf{g}(\tilde{\mathbf{x}}) \right\|_2^2 \right] - H\left(\mathbf{g}(\mathbf{x})\right) \tag{8.5}$$

where $H(\cdot)$ denotes the differential entropy of the random variable $\mathbf{g}(\mathbf{x})$ taking values in $(0,1)^{n_c}$. Then $\mathbf{g}$ block-identifies the true content variables in the sense of Defn. 8.4.1.

Proof sketch. First, use the Darmois construction [69, 70] to build a function $\mathbf{d} : \mathcal{C} \to (0,1)^{n_c}$ mapping $\mathbf{c} = \mathbf{f}^{-1}(\mathbf{x})_{1:n_c}$ to a uniform random variable. Then $\mathbf{g}^\star = \mathbf{d} \circ \mathbf{f}^{-1}_{1:n_c}$ attains the global minimum of (8.5) because $\mathbf{c}$ is invariant across $(\mathbf{x}, \tilde{\mathbf{x}})$ and the uniform distribution is the maximum entropy distribution on $(0,1)^{n_c}$. Thus, any other minimiser $\mathbf{g}$ of (8.5) must satisfy invariance across $(\mathbf{x}, \tilde{\mathbf{x}})$ and map to a uniform r.v. Then, use the same step 2 as in Thms. 8.4.1 and 8.4.2 to show that $\mathbf{h} = \mathbf{g} \circ \mathbf{f} : \mathcal{Z} \to (0,1)^{n_c}$ cannot depend on style, i.e., it is a function from $\mathcal{C}$ to $(0,1)^{n_c}$. Finally, we

show that $\mathbf{h}$ must be invertible since it maps $p_{\mathbf{c}}$ to a uniform distribution, using a result from [88].

Intuition. Thm. 8.4.3 states that if we do not explicitly enforce invertibility of $\mathbf{g}$ as in Thm. 8.4.2, additionally maximising the entropy of the learnt representation (i.e., optimising alignment *and* uniformity [374]) avoids a collapsed representation and recovers the invariant content block. Intuitively, this is because any function that only depends on $\mathbf{c}$ will be invariant across $(\mathbf{x}, \tilde{\mathbf{x}})$, so it is beneficial to preserve all content information to maximise entropy.

Discussion. Of our theoretical results, Thm. 8.4.3 requires the weakest set of assumptions, and is most closely aligned with common SSL practice. As discussed in § 8.2, contrastive SSL with negative samples using InfoNCE (8.1) as an objective can asymptotically be understood as alignment with entropy regularisation [374], i.e., objective (8.5). *Thm. 8.4.3 thus provides a theoretical justification for the empirically observed effectiveness of CL with InfoNCE*: subject to our assumptions, CL with InfoNCE asymptotically isolates content, i.e., the part of the representation that is always left invariant by augmentation. For example, the strong image classification performance based on representations learnt by `SimCLR` [329], which uses color distortion and random crops as augmentations, can be explained in that object class is a content variable in this case. We extensively evaluate the effect of various augmentation techniques on different ground-truth latent factors in our experiments in § 8.5. There is also an interesting connection between Thm. 8.4.3 and `BarlowTwins` [365], which only uses positive pairs and combines alignment with a redundancy reduction regulariser that enforces decorrelation between the inferred latents. Intuitively, redundancy reduction is related to increased entropy: $\mathbf{g}^\star$ constructed in the proof of Thm. 8.4.3—and thus also any other minimiser of (8.5)—attains the global optimum of the `BarlowTwins` objective, though the reverse implication may not hold.

8.5 Experiments

We perform two main experiments. First, we numerically test our main result, Thm. 8.4.3, in a *fully-controlled*, finite sample setting (§ 8.5.1), using CL to estimate the entropy term in (8.5). Second, we seek to better understand the effect of data augmentations used *in practice* (§ 8.5.2). To this end, we introduce a new dataset of 3D objects with dependencies between a number of known ground-truth factors, and use it to evaluate the effect of different augmentation techniques on what is identified as content. Additional experiments are summarised in § 8.5.3 and described in more detail in Appendix G.3. Code to reproduce the experiments is available at: https://www.github.com/ysharma1126/ssl_identifiability.

8.5.1 Numerical Data

Experimental setup. We generate synthetic data as described in § 8.3. We consider $n_c = n_s = 5$, with content and style latents distributed as

$\mathbf{c} \sim \mathcal{N}(0, \Sigma_c)$ and $\mathbf{s}|\mathbf{c} \sim \mathcal{N}(\mathbf{a} + B\mathbf{c}, \Sigma_s)$, thus allowing for *statistical dependence* within the two blocks (via Σ_c and Σ_s) and *causal dependence* between content and style (via B). For $\mathbf{f}$, we use a 3-layer MLP with LeakyReLU activation functions.[10] The distribution p_A over subsets of changing style variables is obtained by independently flipping the same biased coin for each s_i. The conditional style distribution is taken as $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A} = \mathcal{N}(\mathbf{s}_A, \Sigma_A)$. We train an encoder $\mathbf{g}$ on pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ with InfoNCE using the negative L2 loss as the similarity measure, i.e., we approximate (8.5) using empirical averages and negative samples. For evaluation, we use kernel ridge regression [67] to predict the ground truth $\mathbf{c}$ and $\mathbf{s}$ from the learnt representation $\hat{\mathbf{c}} = \mathbf{g}(\mathbf{x})$ and report the R^2 coefficient of determination. For a more detailed account, we refer to Appendix G.4.

10: chosen to lead to invertibility almost surely by following the settings used by previous work [64, 65]

Results. In Tab. 8.1, we report mean ± std. dev. over 3 random seeds across four generative processes of increasing complexity covered by Thm. 8.4.3: "p(chg.)", "Stat.", and "Cau." denote respectively the change probability for each s_i, statistical dependence within blocks ($\Sigma_c \neq \mathbf{I} \neq \Sigma_s$), and causal dependence of style on content ($B \neq 0$). An R^2 close to one indicates that almost all variation is explained by $\hat{\mathbf{c}}$, i.e., that there is a 1-1 mapping, as required by Defn. 8.4.1. As can be seen, *across all settings, content is block-identified*. Regarding style, we observe an increased score with the introduction of dependencies, which we explain in an extended discussion in Appendix G.3.1. Finally, we show in Appendix G.3.1 that a high R^2 score can be obtained even if we use linear regression to predict $\mathbf{c}$ from $\hat{\mathbf{c}}$ ($R^2 = 0.98 \pm 0.01$, for the last row).

Generative process			R^2 (nonlinear)	
p(chg.)	Stat.	Cau.	Content c	Style s
1.0	✗	✗	**1.00** ± 0.00	0.07 ± 0.00
0.75	✗	✗	**1.00** ± 0.00	0.06 ± 0.05
0.75	✓	✗	**0.98** ± 0.03	0.37 ± 0.05
0.75	✓	✓	**0.99** ± 0.01	**0.80** ± 0.08

Table 8.1: R^2 scores for content and style in our synthetic experiment.

8.5.2 High-Dimensional Images: *Causal3DIdent*

***Causal3DIdent* dataset.** *3DIdent* [88] is a benchmark for evaluating identifiability with rendered 224×224 images which contains hallmarks of natural environments (e.g. shadows, different lighting conditions, a 3D object). For influence of the latent factors on the renderings, see Fig. 2 of [88]. In *3DIdent*, there is a single object class (Teapot [383]), and all 10 latents are sampled independently. For *Causal3DIdent*, we introduce **six** additional classes: Hare [384], Dragon [385], Cow [386], Armadillo [387], Horse [388], and Head [389]; and impose a causal graph over the latent variables, see Fig. 8.2. While object class and all environment variables (spotlight position & hue, background hue) are sampled independently, all object latents are dependent,[11] see Appendix G.2 for details. The Causal3DIdent dataset is publicly available at https://zenodo.org/record/4784282.

11: e.g., our causal graph entails hares blend into the environment (object hue centered about background & spotlight hue), a form of active camouflage observed in Alaskan [390], Arctic [391], & Snowshoe hares.

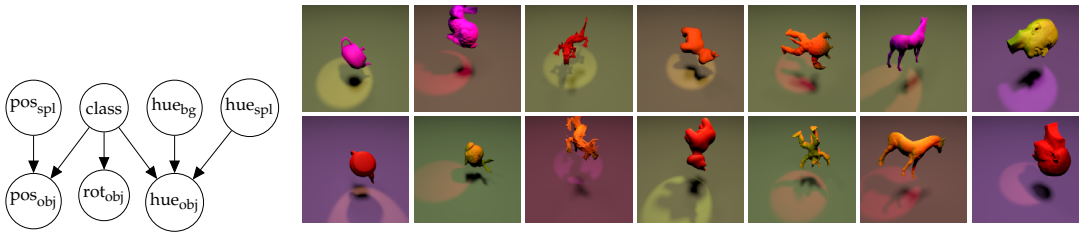


Figure 8.2: *(Left)* Causal graph for the *Causal3DIdent* dataset. *(Right)* Two samples from each object class.

Experimental setup. For $\mathbf{g}$, we train a convolutional encoder composed of a ResNet18 [392] and an additional fully-connected layer, with LeakyReLU activation. As in `SimCLR` [329], we use InfoNCE with cosine similarity, and train on pairs of augmented examples $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$. As n_c is unknown and variable depending on the augmentation, we fix $\dim(\hat{\mathbf{c}}) = 8$ throughout. Note that we find the results to be, for the most part, robust to the choice of $\dim(\hat{\mathbf{c}})$, see Fig. 8.3. We consider the following data augmentations (DA): crop, resize & flip; colour distortion (jitter & drop); and rotation $\in \{90°, 180°, 270°\}$. For comparison, we also consider directly imposing a content-style partition by performing a latent transformation (LT) to generate views. For evaluation, we use linear logistic regression to predict object class, and kernel ridge to predict the other latents from $\hat{\mathbf{c}}$. See Appendix G.3.2 for results with linear regression, as well as evaluation using a higher-dimensional intermediate layer by considering a projection head [329].

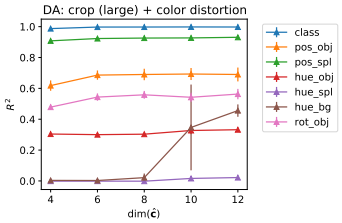


Figure 8.3: R^2 score as a function of $\dim(\hat{\mathbf{c}})$ for various latent factors.

Results. The results are presented in Tab. 8.2. Overall, our main findings can be summarised as:

(i) it can be difficult to design image-level augmentations that leave *specific* latent factors invariant;
(ii) augmentations & latent transformations generally have a similar effect on groups of latents;
(iii) augmentations that yield good classification performance induce variation in all other latents.

We observe that, similar to directly varying the hue latents, colour distortion leads to a discarding of hue information as style, and a preservation of

Table 8.2: *Causal3DIdent* results: R^2 mean ± std. dev. over 3 random seeds. DA: data augmentation, LT: latent transformation, bold: $R^2 \geq 0.5$, red: $R^2 < 0.25$. Results for individual axes of object position & rotation are aggregated, see Appendix G.3 for the full table.

Views generated by	Class	Positions		Hues			Rotations
		object	spotlight	object	spotlight	background	
DA: colour distortion	0.42 ± 0.01	**0.61** ± 0.10	0.17 ± 0.00	0.10 ± 0.01	0.01 ± 0.00	0.01 ± 0.00	0.33 ± 0.02
LT: change hues	**1.00** ± 0.00	**0.59** ± 0.33	**0.91** ± 0.00	0.30 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.30 ± 0.01
DA: crop (large)	0.28 ± 0.04	0.09 ± 0.08	0.21 ± 0.13	**0.87** ± 0.00	0.09 ± 0.02	**1.00** ± 0.00	0.02 ± 0.02
DA: crop (small)	0.14 ± 0.00	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	**1.00** ± 0.00	0.00 ± 0.00
LT: change positions	**1.00** ± 0.00	0.16 ± 0.23	0.00 ± 0.01	0.46 ± 0.02	0.00 ± 0.00	**0.97** ± 0.00	0.29 ± 0.01
DA: crop (large) + colour distortion	**0.97** ± 0.00	**0.59** ± 0.07	**0.59** ± 0.05	0.28 ± 0.00	0.01 ± 0.01	0.01 ± 0.00	**0.74** ± 0.03
DA: crop (small) + colour distortion	**1.00** ± 0.00	**0.69** ± 0.04	**0.93** ± 0.00	0.30 ± 0.01	0.00 ± 0.00	0.02 ± 0.03	**0.56** ± 0.03
LT: change positions + hues	**1.00** ± 0.00	0.22 ± 0.22	0.07 ± 0.08	0.32 ± 0.02	0.00 ± 0.01	0.02 ± 0.03	0.34 ± 0.06
DA: rotation	0.33 ± 0.06	0.17 ± 0.09	0.23 ± 0.12	**0.83** ± 0.01	0.30 ± 0.12	**0.99** ± 0.00	0.05 ± 0.03
LT: change rotations	**1.00** ± 0.00	**0.53** ± 0.33	**0.90** ± 0.00	0.41 ± 0.00	0.00 ± 0.00	**0.97** ± 0.00	0.28 ± 0.00
DA: rotation + colour distortion	**0.59** ± 0.01	**0.58** ± 0.06	0.21 ± 0.01	0.12 ± 0.02	0.01 ± 0.00	0.01 ± 0.00	0.33 ± 0.04
LT: change rotations + hues	**1.00** ± 0.00	**0.57** ± 0.34	**0.91** ± 0.00	0.30 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.28 ± 0.00

(object) position as content. Crops, similar to varying the position latents, lead to a discarding of position as style, and a preservation of background and object hue as content, the latter assuming crops are sufficiently large. In contrast, image-level rotation affects both the object rotation and position, and thus deviates from only varying the rotation latents.

Whereas class is always preserved as content when generating views with latent transformations, when using data augmentations, we can only reliably decode class when crops & colour distortion are used in conjunction—a result which mirrors evaluation on ImageNet [329]. As can be seen by our evaluation of crops & colour distortion in isolation, while colour distortion leads to a discarding of hues as style, crops lead to a discarding of position & rotation as style. Thus, when used in conjunction, class is isolated as the sole content variable. See Appendix G.3.2 for additional analysis.

8.5.3 Additional Experiments and Ablations

We also perform an ablation on $\dim(\hat{\mathbf{c}})$ for the synthetic setting from § 8.5.1, see Appendix G.3.1 for details. Generally, we find that if $\dim(\hat{\mathbf{c}}) < n_c$, there is insufficient capacity to encode all content, so a lower-dimensional mixture of content is learnt. Conversely, if $\dim(\hat{\mathbf{c}}) > n_c$, the excess capacity is used to encode some style information (as that increases entropy). Further, we repeat our analysis from § 8.5.2 using `BarlowTwins` [365] (instead of `SimCLR`) which, as discussed at the end of § 8.4.2, is also loosely related to Thm. 8.4.3. The results mostly mirror those obtained for `SimCLR` and presented in Tab. 8.2, see Appendix G.3.2 for details. Finally, we ran the same experimental setup as in § 8.5.2 also on the *MPI3D-real* dataset [393] containing > 1 million *real* images with ground-truth annotations of 3D objects being moved by a robotic arm. Subject to some caveats, the results show a similar trend as those on *Causal3DIdent*, see Appendix G.3.3 for details.

8.6 Discussion

Theory vs practice. We have made an effort to tailor our problem formulation (§ 8.3) to the setting of data augmentation with content-preserving transformations. However, some of our more technical assumptions, which are necessary to prove block-identifiability of the invariant content partition, may not hold exactly in practice. This is apparent, e.g., from our second experiment (§ 8.5.2), where we observe that—while class should, in principle, always be invariant across views (i.e., content)—when using *only* crops, colour distortion, or rotation, $\mathbf{g}$ appears to encode *shortcuts* [394, 395].[12] Data augmentation, unlike latent transformations, generates views $\tilde{\mathbf{x}}$ which are not restricted to the 11-dim. image manifold $\mathcal{X}$ corresponding to the generative process of *Causal3DIdent*, but may introduce additional variation: e.g., colour distortion leads to a rich combination of colours, typically a 3-dim. feature, whereas *Causal3DIdent* only contains one degree of freedom (hue). With additional factors, any introduced invariances may be encoded as content in place of class. Image-level augmentations also tend to change multiple latent factors in a correlated way, which may violate assumption *(iii)* of our theorems,

12: class is distinguished by shape, a feature commonly unused by convolutional neural networks in downstream tasks on natural images [396]

i.e., that $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ is fully-supported locally. We also assume that $\mathbf{z}$ is continuous, even though *Causal3DIdent* and most disentanglement datasets also contain discrete latents. This is a very common assumption in the related literature [62, 64, 66, 70, 72, 83, 88, 114, 359] that may be relaxed in future work. Moreover, our theory holds asymptotically and at the global optimum, whereas in practice we solve a non-convex optimisation problem with a finite sample and need to approximate the entropy term in (8.5), e.g., using a finite number of negative pairs. The resulting challenges for optimisation may be further accentuated by the higher dimensionality of $\mathcal{X}$ induced by image-level augmentations. Finally, we remark that while, for simplicity, we have presented our theory for pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ of original and augmented examples, in practice, using pairs $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$ of two augmented views typically yields better performance. All of our assumptions (content invariance, changing style, etc) and theoretical results still apply to the latter case. We believe that using two augmented views helps because it leads to *increased variability* across the pair: for if $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}'$ differ from $\mathbf{x}$ in style subsets A and A', respectively, then $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$ differ from each other (a.s.) in the union $A \cup A'$.

Beyond entropy regularisation. We have shown a clear link between an identifiable maximum entropy approach to SSL (Thm. 8.4.3) and `SimCLR` [329] based on the analysis of [374], and have discussed an intuitive connection to the notion of redundancy reduction used in `BarlowTwins` [365]. Whether other types of regularisation such as the architectural approach pursued in `BYOL` [330] and `SimSiam` [331] can also be linked to entropy maximisation, remains an open question. Deriving similar results to Thm. 8.4.3 with other regularisers is a promising direction for future research, c.f. [397].

The choice of augmentation technique implicitly defines content and style. As we have defined content as the part of the representation which is always left invariant across views, the choice of augmentation implicitly determines the content-style partition. This is particularly important to keep in mind when applying SSL with data augmentation to safety-critical domains, such as medical imaging. We also advise caution when using data augmentation to identify specific latent properties, since, as observed in § 8.5.2, image-level transformations may affect the underlying ground-truth factors in unanticipated ways. Also note that, *for a given downstream task*, we may not want to discard all style information since style variables may still be correlated with the task of interest and may thus help improve predictive performance. *For arbitrary downstream tasks*, however, where style may change in an adversarial way, it can be shown that only using content is optimal [134].

***What* vs *how* information is encoded.** In this chapter, we focused on *what* information is retained in representations learnt by SSL with data augmentations: block-identifying the content variables corresponds to discarding information about the style variables. The notion of block identifiability Defn. 8.4.1 is not informative on *how* the content variables are encoded (individual causes may be arbitrarily entangled in the learnt representation). Orthogonal to our contribution, a different line of work instead studies *how* information is encoded in a task-dependent

manner, in the sense of analysing the sample complexity needed to solve *a given downstream task* using a *linear* predictor [397–402]. Provided that downstream tasks only involve content, we can draw some comparisons. Whereas our results recover content only up to arbitrary invertible nonlinear functions (see Defn. 8.4.1), our problem setting is more general: [398, 400] assume (approximate) independence of views $(\mathbf{x}, \tilde{\mathbf{x}})$ given the task (content), while [401, 402] assume (approximate) independence between one view and the task (content) given the other view, neither of which hold in our setting.

Conclusion. Existing representation learning approaches typically assume mutually independent latents, though dependencies clearly exist in nature [13]. We demonstrate that in a *non-i.i.d.* scenario, e.g., by constructing multiple views of the same example with data augmentation, we can learn useful representations in the presence of this neglected phenomenon. More specifically, the contribution in this chapter is, to the best of our knowledge, the first: (i) identifiability result under *arbitrary dependence* between latents; and (ii) empirical study that evaluates the effect of data augmentations not only on classification, but also on other *continuous* ground-truth latents. Unlike existing identifiability results which rely on *change* as a learning signal, our approach aims to identify what is always shared across views, i.e., also using *invariance* as a learning signal. We hope that this change in perspective will be helpful for applications such as optimal style transfer or disentangling shape from pose in vision, and inspire other types of *counterfactual training* to recover a more fine-grained causal representation.

Conclusions & Future Perspectives

9 Closing Remarks

In § 9.1, we summarise some open questions and potential avenues for future research based on the contributions presented in this thesis. We then present some concluding remarks on identifiability, mentioning its role in causal inference and how it connects to the perspective presented in this manuscript (§ 9.2); finally, in § 9.3, we present some reflections on its significance in current machine learning practice.

9.1 Potential Avenues of Future Research

9.1.1 Extensions of Independent Mechanism Analysis

Identifiability. One important open question is the full characterisation of identifiability of the model proposed in Chapter 3. While the IMA function class has not yet been shown to be identifiabile, such results exist for special cases such as conformal maps (the case $n = 2$ was discussed in [70]), isometries [201] and for closely-related unsupervised nonlinear ICA models [89].

Since the original work was published, there have already been works providing initial answers to this, see [89, 90], or proposing related models, see [201].[1] [90] proved a weaker form of identifiability (termed *"local identifiability"*) for the IMA function class, and characterised a set of corner cases where identifiability is impossible (similar to the Gaussian case in linear ICA). This progress will hopefully lead to a more thorough characterisation of IMA and to an explanation of its empirical success in blind source separation, as observed in Chapter 3 and Chapter 4.

1: [201] was published at the same conference as our work [63], and considers a related, but more restrictive class of mixing functions, i.e., isometries.

Estimation. Another question concerns efficient estimation of IMA: as discussed in Chapter 5, optimisation of the Jacobian term is hard, and the IMA regularisation might not scale well to high-dimensional data. In part, this is already implicitly solved in Chapter 4 by using VAEs (which sidestep expensive Jacobian computations) to estimate IMA. However, VAEs may not be the most efficient way to enforce column orthogonality of the Jacobian. A direction for future research would be to develop efficient ways of enforcing orthogonality of the Jacobian which avoid expensive computations via automatic differentiation,[2] in a similar spirit to the contribution in Chapter 5 for unconstrained Jacobians.

2: For example, [403] introduces a stochastic estimator based on Hutchinson's trace estimator and finite differences.

Fewer sources than observed components. It may be interesting to study undercomplete settings (where the number of latent sources is lower than the number of observed components) for IMA, since this is arguably the most typical setting in representation learning. In Chapter 4, and particularly in the experiments of § 4.4.3, we showed empirical results where VAEs appear to achieve blind source separation on high-dimensional, image data: however, an extension of the IMA theory to such setting is still missing.[3]

3: [404] studied an objective a special case of which is equivalent to IMA, and discussed how to combine it with dimensionality reduction [404, Sec. 3.2].

Robustness of IMA to model misspecification. Another open question concerns robustness of IMA to model misspecification. While the results presented in Chapter 3 assume a certain ground truth model to hold, in many practical situations we may expect some degree of violation of the assumptions. A first empirical study [405] gave encouraging results, showing a degree of robustness of IMA to violations of the underlying assumptions. A theoretical characterisation of the model's performance in these cases might, however, require a novel approach—see, e.g., [406].

9.1.2 Other avenues

Multiple layers of representation. While most of the work on identifiability we presented concerns the *final layer* representations of neural networks, fewer works investigate the intermediate layer representations (in § 1.3.2, one of our motivations for studying representation learning was that deep learning extracts *"multiple layers of representation"* [10]). For example, it is well-known that neural networks trained on natural images consistently recover Gabor filters in the early layers [407, 408]. Previous work also analysed the transferability of features extracted at different layers [409], finding that it decreases with depth.

The identifiability theory we presented does not directly account for such intermediate representations extracted by end-to-end training—as it would only concern the final layers. Given the observed reproducibility of some intermediate layer features, this seems like an interesting problem to be studied through the lens of identifiability, as also observed in [410].

Hierarchical latent variable models for subgroup differences. An extension of latent variable models which would be interesting to consider in future work are *hierarchical* latent variable models [411–416]. For example, for statistical analysis of experimental studies (e.g., when a treatment and a control group are involved), it would be interesting to extend the work in Chapter 6 to model subgroup-specific variability (besides differences across individuals, as in Chapter 6) in a hierarchical fashion. Interesting questions may arise both with respect to model identifiability and estimation.

Statistical efficiency and finite-sample analysis of identifiable methods. A question we did not address in this thesis is finite-sample performance of different estimation procedures for identifiable methods. For nonlinear ICA, a recent contribution in this direction was [417], which characterised finite-sample behaviour of contrastive-learning based estimation.[4] Another paper, which studies a model closely related to the one presented in Chapter 8, is [419], in particular Theorem 3 on Sample Complexity.

4: In the context of contrastive learning, an analysis which also takes into consideration the computational budget, and characterises the optimal noise distribution for a fixed budget, was presented in [418].

9.2 Identifiability in Representation Learning and Causal Inference

In Chapter 2, we described the ICA approach to representation learning, and mentioned (§ 2.2.4) that it postulates a separation between the

problems of identification and estimation. According to [66],

> The essential difference [between nonlinear ICA and] most methods for unsupervised representation learning is that the approach starts by defining a generative model in which the original latent variables can be recovered, i.e. the model is identifiable by design.

[66]: Hyvärinen et al. (2019), 'Nonlinear ICA using auxiliary variables and generalized contrastive learning'

Here we want to briefly review how this perspective on identifiability is related to the one in causal inference, and in particular to the approach presented in [122].

[122]: Pearl (2009), *Causality*

9.2.1 Identifiability in Causal Inference

Following [122, 420], we postulate that the ground truth data generating process can be represented in the form of a Structural Causal Model (SCM).[5] Each SCM induces a causal hierarchy [420], also termed the *ladder of causation* [421], which can be thought of as a taxonomy of causal questions organised in three distinct levels or rungs:

(i) **Rung 1:** Association (*"What does a symptom tell me about a disease?"*);
(ii) **Rung 2:** Intervention (*"If I take an aspirin, will my headache be cured?"*);
(iii) **Rung 3:** Counterfactuals (*"Was it the aspirin that stopped my headache?"*).[6]

5: See, e.g., [420, Definition 1] for a technical definition. [420] remark that a ground truth SCM is postulated *"whether or not we, as an epistemological matter, know much about it"*.

[420]: Bareinboim et al. (2022), 'On Pearl's hierarchy and the foundations of causal inference'

[421]: Pearl et al. (2018), *The book of Why: the new science of cause and effect*

6: The example questions are taken from [421].

Causal inference is especially difficult since we typically only have measurements from lower rungs, but want to reason about higher ones: for example, knowledge about statistical associations (rung 1 knowledge) within some observed variables, may be insufficient to identify causal effects and answer questions on how their distribution would change under interventions (rung 2 queries). Moreover, counterfactual (rung 3) questions (i.e., questions of the form "what would have happened if") might not be uniquely answered even if experiments (rung 2) are at hand. As argued in [420], *"it is generically impossible to draw higher-layer inferences using only lower-layer information"*.

A key question of causal inference is what assumptions and measurements are required to unambiguously answer, or *identify*, a given causal query.[7] A central contribution in this field is the *do-calculus* [423, 424], which provides a way to determine whether a given causal query can be unambiguously answered based on the available measurements (possibly a combination of experimental and observational data), based on assumptions including a directed acyclic graph (DAG) summarising the causal relations occurring among the considered variables.

7: In this context, the identifiability requirement is articulated in [422, Def. 2]. Note that this definition is closer in spirit to the one in (2.10) than to the one based on equivalence relations (2.11).

In a recent interview, Judea Pearl described his contribution to causality as follows [425]:

> I have focused on the problem of identification, rather than estimation. This calls for transforming the desired causal quantity into an equivalent probabilistic expression (called estimand) that can be estimated from data. Once an estimand is derived, the actual estimation step is no longer causal, and can be accomplished by standard statistical methods. This is indeed where machine learning excels, unlike the identification step in which machine learning and standard statistical methods are almost helpless. It is for this reason

that I focus on identification – this is where the novelty of causal thinking lies, and where a new calculus had to be developed.

The focus on identifiability is therefore a common theme between this approach to causal inference and nonlinear ICA.[8] Besides this analogy, it is worth making some further remarks:

- Causal inference typically focuses on identifiability of *queries* (e.g., "Will I get a headache if I take an aspirin?"), not models (but see, e.g., [121]); whereas nonlinear ICA focuses on identification of (un)mixing functions and latent variables (these latter two are equivalent in the noiseless case we introduced in (2.1)).
- Moreover, for causal queries, when full identification is not achievable, partial identification sometimes still yields informative bounds based on empirically observable quantities [430–433]. To the best of our knowledge, similar *partial identification* results are missing in nonlinear ICA literature.[9]

As a final remark, identifiability is *easier* to assess in the graphical approach to causal inference than it is in representation learning. In fact, under suitable assumptions, the do-calculus provides a way to translate a question about identification into a question about graphical properties of an underlying DAG, which can be solved algorithmically. A similar tool, through which a question regarding identifiability could be translated into a graphical or diagrammatic one, would be quite helpful also in the context of representation learning theory.[10]

Identifiability and estimability. Identifiability of a query or model may however be insufficient to ensure that it can be estimated from data: for example, [49, 50] distinguish the notions of *identifiability* from the one of *estimability*, and argue that estimability is more suited to answer the question of what can be estimated from data. In short, estimability may require some additional smoothness assumptions on the considered distributions.[11]

Identifiability and falsifiability. Identifiability is also related to falsifiability. For example, any interventional model may be consistent with many structural causal models [420]; and it may be impossible to distinguish the true SCM within a family of (observationally or interventionally) equivalent ones based on empirical measurements. These alternative SCMs might entail entirely different counterfactual inferences, but due to their equivalence on any statistical or interventional question they are to all effects unfalsifiable. In turn, when a model or query is identifiable, all but one of the solutions (or a subset thereof in case of partial identifiability) are ruled out, i.e., falsified—see [121, Sec. 7] and discussion therein).

9.2.2 ICA for Causal Inference & Causality for ICA

ICA can be used in causal inference as a method to solve the problem of *causal discovery*—that is, the problem of discovering the DAG representing

8: Identifiability is a central aspect of causal inference beyond [122] (e.g., see [426–428]), but the special focus on algorithms to infer identifiability given a complex model is arguably a distinguishing feature of the approach in [122]; see also [429] and discussion therein.

9: Note that in these partial identifiability results bounds are derived for the population limit, i.e., in the context where infinite datapoints are available: they are conceptually different from finite-sample confidence intervals.

10: Some recent results of identifiability for latent causal models do exploit graphical criteria in their identifiability proofs [434, 435]. Also, in a loose analogy, diagrammatic proofs of identifiabilty have been given in [94] based on category theory and string diagrams.

11: These are often implicitly assumed in ICA literature.

causal relations among some considered variables, instead of assuming it a priori based on domain knowledge [436].

In this context, identifiability results establish whether the DAG summarising the causal relations can be unambiguously determined based on assumptions on the data generating process. While causal discovery from observational data is only possible up to the Markov equivalence class of the true DAG [437], causal discovery methods based on ICA allow going beyond this—subject to additional assumptions mirroring those used for identifiability of ICA [102, 104, 105, 280].[12]

In turn, causality might provide a different perspective to think of problems in ICA and representation learning.[13] The contribution presented in Chapter 3 constitutes, to the best of our knowledge, the first effort to use ideas from causality (specifically ICM) to make progress on the challenging problem of nonlinear blind source separation.

12: Interestingly, these methods identify the underlying SCMs, something which is impossible in general but which may become possible under certain restrictions on the model class [102, 105] or when diverse enough data (e.g., nonstationary) is available [104].

13: In this spirit, Aapo Hyvärinen suggested [438] that the ICA model (2.1) should be written as $\mathbf{x} := \mathbf{f}(\mathbf{s})$, adopting the notation used to define structural assignments in SCMs, thus implicitly considering $\mathbf{x}$ as the effect of $\mathbf{s}$ through the mechanism $\mathbf{f}$.

9.2.3 Toward Causal Representation Learning

As mentioned already in § 2.4.3, the study of latent variable models with non-independent components, and in particular those where the latent components are connected by causal relations and support causal reasoning [13], is an exciting avenue of research [93, 94, 434, 439]. Based on the works presented in this thesis, the problem of causal representation learning may be further explored in different ways.

[13]: Schölkopf et al. (2021), 'Toward causal representation learning'

On the one hand, the multi-view setting discussed in Chapter 6 may be extended by allowing causal dependence among the latent variables, and considering pairs of views $(\mathbf{x}, \tilde{\mathbf{x}})$ as observations of the same system before and after a given intervention. In the context of data augmentation, a first step was discussed in Chapter 8, where a causal interpretation of the data-augmentation process was presented (§ 8.3). This may be further explored and extended to model other kinds of actions or interventions; for example, the interactions between an agent and a system, thus grounding representation learning in the comparison between measurements of the same system before and after an intervention [93, 94, 440]. This may be relevant in the context of reinforcement learning [441], where the role of different representations is a topic of active research, with new datasets and benchmarks [442, 443].

On the other hand, an application of the IMA principle in causal representation learning [13] may also be an interesting direction for future work. The IMA principle enforces constraints on the mixing function class, which are orthogonal to constraints on the source distribution, and could therefore be combined with a different assumption from independence on the latent components,[14] including latent causal models. It is possible that the IMA constraint on the mixing function may be helpful to solve the representation learning problem even with non-independent latent components. In fact, many existing works on identifiability test their results on high-dimensional datasets through estimation procedures based on modifications of the VAE model, e.g., [94, 114, 359]. Such works provide identifiability proofs based on principles such as, e.g., weak supervision [94, 114], and completely unrelated to the implicit functional constraint of VAEs discussed in Chapter 3. It would therefore be interesting to investigate whether the implicit IMA regularisation of VAEs

14: In the different setting of post-hoc concept discovery, related ideas have also been explored in [444].

described in Chapter 3 is also present in these modified VAE models (and others, e.g., [101]), and whether it might play a role in explaining their empirical effectiveness.

9.3 Identifiability and Current Empirical Practice in Machine learning

Underspecification. Besides representation learning, a problem in current machine learning practice which may be related to identifiability is *underspecification* [445]. A machine learning pipeline is underspecified when it can return many predictors with equivalently strong held-out performance in the training domain.[15] This may be problematic when the model is required to encode some essential structure of the problem at hand, allowing it to perform well beyond its training set.

The problem diagnosed in [445] is that many existing machine learning pipelines allow highly non-unique predictors: that is, predictors trained to the same level of held-out performance on some training data can show widely divergent behaviour when applied to real-world settings. The authors argue that underspecification in machine learning pipelines is a key obstacle to reliable training of models that behave as expected in deployment; and that it is ubiquitous in modern applications of machine learning, with substantial practical implications undermining its credibility.

It would therefore be interesting to explore whether the characterisation of identifiability discussed in this manuscript might be helpful to address this problem from a theoretical perspective.[16]

[445]: D'Amour et al. (2020), 'Underspecification presents challenges for credibility in modern machine learning'

15: Following [445], we consider a supervised learning setting where the goal is to learn a predictor $f : \mathcal{X} \mapsto \mathcal{Y}$ mapping inputs $x \in \mathcal{X}$ to labels $y \in \mathcal{Y}$. A *model* is specified by a function class $\mathcal{F}$ from which the predictor f will be chosen; a *pipeline* takes training data $\mathcal{D}$ from a given distribution $\mathbb{P}$ and produces a trained model, or *predictor*, $f(x)$.

16: In the authors' own words [445], for the analysis of the underspecification phenomenon *"there are opportunities to import ideas from the sensitivity analysis and partial identification subfields in causal inference and inverse problems"*.

Empirical practice in representation learning. Advancements in the fields of representation learning and disentanglement have been largely driven by an empirical perspective. While many effective self-supervised or fully unsupervised methods were developed without being explicitly grounded in identifiability theory, their empirical success might a posteriori be interpreted in this light.

In the case of self-supervised learning methods, it has been argued [66] that nonlinear ICA, and specifically the identifiability theory based on auxiliary variables (§ 2.4.2), establishes mathematical principles underlying a strand of self-supervised approaches which rely on videos and exploit the correspondence between the visual and audio streams [446, 447] .We believe that the contribution in Chapter 8 adds another element to the puzzle by investigating commonly deployed data augmentation strategies [329].

In unsupervised representation learning, an interesting empirical finding is that, despite the intrinsic limitations and impossibility results for fully unsupervised representation learning [70, 72], VAE and β-VAE architectures appear in some cases to have the ability to transform raw, unstructured data into a semantically meaningful set of latent variables (e.g., [194, 199] and § 4.4.3). As argued in, e.g., [448], this contributed to making VAEs part of some state-of-the-art world-models in the context of reinforcement learning [449, 450]. The contribution in Chapter 4 aims at

providing a partial explanation of this phenomenon, by linking the ELBO gap to the IMA-regularisation introduced in Chapter 3 (although, as argued in § 4.6, further work might be required to elucidate the specificity of β-VAE models).

Summarising, empirical practice in representation learning appears to have developed methods whose success may be interpreted through the lens of identifiability theory, even though they were not originally conceived based on it. In the future, it will be interesting to see whether the study of identifiability will be useful to explain success of novel machine learning methods (e.g., recent self-supervised approaches such as masked autoencoders [451]; or popular models such as the VQ-VAE [452]), improve them or suggest new approaches to representation learning.

Appendix

A

Additional Material for Chapter 2

A.1 Whitening in the context of linear ICA

We give a brief account of the role of *whitening in linear ICA*, which was mentioned in § 2.2.2 and which again plays a role in B.1.1. The following exposition is largely based on [1], §7.4.2.

[1]: Hyvärinen et al. (2001), *Independent Component Analysis*

A zero-mean random vector, say $\mathbf{y}$, is said to be *white* if its components are uncorrelated and their variances equal unity. In other words, the covariance matrix of $\mathbf{y}$ is equal to the identity matrix:

$$\mathbb{E}\left[\mathbf{y}\mathbf{y}^\top\right] = \mathbf{I}\,.$$

It is always possible to whiten a zero-mean random vector $\mathbf{x}$ through a linear operation,

$$\mathbf{z} = \mathbf{V}\mathbf{x}\,. \tag{A.1}$$

As an example, a popular method for whitening uses the eigenvalue decomposition (EVD) of the covariance matrix,

$$\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right] = \mathbf{E}\mathbf{D}\mathbf{E}^\top$$

where $\mathbf{E}$ is the orthogonal matrix of eigenvectors of $\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right]$ and $\mathbf{D}$ is the diagonal matrix of its eigenvalues, $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Note that the covariance matrix is a symmetric matrix, therefore it is diagonalisable. Whitening can then be performed by substituting in (A.1) the matrix

$$\mathbf{V} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^\top\,. \tag{A.2}$$

so that

$$\mathbb{E}[\mathbf{z}\mathbf{z}^\top] = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^\top\mathbf{E}\mathbf{D}\mathbf{E}^\top\mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^\top = \mathbf{I}$$

Whitening is only half ICA. Assume a linear ICA model,

$$\mathbf{x} = \mathbf{A}\mathbf{s}\,. \tag{A.3}$$

and suppose that the observed data is whitened, for example, by the matrix $\mathbf{V}$ given in (A.2). Whitening transforms the mixing matrix into a new one, $\tilde{\mathbf{A}} = \mathbf{V}\mathbf{A}$. We have from (A.3) and (A.2)

$$\mathbf{z} = \mathbf{V}\mathbf{A}\mathbf{s} = \tilde{\mathbf{A}}\mathbf{s}$$

Note that whitening does not solve linear ICA, since *uncorrelatedness is weaker than independence*. To see this, consider any orthogonal transformation $\mathbf{U}$ of $\mathbf{z}$:

$$\mathbf{y} = \mathbf{U}\mathbf{z}.$$

Due to the orthogonality of $\mathbf{U}$, we have

$$\mathbb{E}\left[\mathbf{y}\mathbf{y}^\top\right] = \mathbb{E}\left[\mathbf{U}\mathbf{z}\mathbf{z}^\top\mathbf{U}^\top\right] = \mathbf{U}\mathbb{E}\left[\mathbf{z}\mathbf{z}^\top\right]\mathbf{U}^\top = \mathbf{U}\mathbf{I}\mathbf{U}^\top = \mathbf{I}\,,$$

so, $\mathbf{y}$ is white as well. Thus, we cannot tell if the independent components are given by $\mathbf{z}$ or $\mathbf{y}$ using the whiteness property alone. Since $\mathbf{y}$ could be any orthogonal transformation of $\mathbf{z}$, whitening gives the independent components only up to an orthogonal transformation.

On the other hand, whitening is useful as a pre-processing step in ICA: its utility resides in the fact that the new mixing matrix $\tilde{\mathbf{A}} = \mathbf{V}\mathbf{A}$ is orthogonal. This can be seen from

$$\mathbb{E}\left[\mathbf{z}\mathbf{z}^\top\right] = \tilde{\mathbf{A}}\mathbb{E}\left[\mathbf{s}\mathbf{s}^\top\right]\tilde{\mathbf{A}}^\top = \tilde{\mathbf{A}}\tilde{\mathbf{A}}^\top = \mathbf{I}.$$

We can thus restrict the search for the (un)mixing matrix to the space of orthogonal matrices. Instead of having to estimate n^2 parameters (the elements of the original matrix $\mathbf{A}$), we only need to estimate an orthogonal mixing matrix $\tilde{\mathbf{A}}$ which contains $n(n-1)/2$ degrees of freedom; e.g., in two dimensions, an orthogonal transformation is determined by a single angle parameter. For larger n, an orthogonal matrix contains only about half of the number of parameters of an arbitrary matrix.

Whitening thus "solves half of the problem of ICA". Because whitening is a very simple and standard procedure—much simpler than any ICA algorithm—it is a good idea to reduce the complexity of the problem this way. The remaining half of the parameters has to be estimated by some other method.

A.2 The variability assumption [66]

Here we report the definition of the assumption of variability, presented in [66]:

Definition A.2.1 (Assumption of Variability) *For any $\mathbf{y} \in \mathbb{R}^n$, there exist $2n+1$ values for $\mathbf{u}$, denoted by $\mathbf{u}_j$, $j = 0 \ldots 2n$ such that the $2n$ vectors in $\mathbb{R}^{2n}$ given by*

$$\begin{aligned}(\mathbf{w}(\mathbf{y},\mathbf{u}_1) - \mathbf{w}(\mathbf{y},\mathbf{u}_0))\,,(\mathbf{w}(\mathbf{y},\mathbf{u}_2) - \mathbf{w}(\mathbf{y},\mathbf{u}_0))\,,\\ \ldots,(\mathbf{w}(\mathbf{y},\mathbf{u}_{2n}) - \mathbf{w}(\mathbf{y},\mathbf{u}_0))\end{aligned}$$

with

$$\begin{aligned}\mathbf{w}(\mathbf{y},\mathbf{u}) = \Bigg(&\frac{\partial q_1(y_1,\mathbf{u})}{\partial y_1},\ldots,\frac{\partial q_n(y_n,\mathbf{u})}{\partial y_n}\\ &\frac{\partial^2 q_1(y_1,\mathbf{u})}{\partial y_1^2},\ldots,\frac{\partial^2 q_n(y_n,\mathbf{u})}{\partial y_n^2}\Bigg)\end{aligned}$$

B

Additional Material on Chapter 3

Overview

- Appendix B.1 contains additional discussion of existing ICM criteria and their relation to IMA.
- Appendix B.2 presents the full proofs for all theoretical results from Chapter 3.
- Appendix B.3 contains a worked out computation of the value of C_{IMA} for the mapping from radial to Cartesian coordinates.
- Appendix B.4 contains experimental details and additional results.
- Appendix B.5 contains additional background on conformal maps and Möbius transformations

B.1 Existing ICM criteria and their relationship to ICA and IMA

Here, we provide additional discussion of the ICM principle and its connection to ICA and IMA. First, we introduce a linear ICM criterion and discuss its relation with linear ICA in Appendix B.1.1.

B.1.1 Trace method

As mentioned in § 3.2.1, besides IGCI, another existing ICM criterion that is closely related to ICA due to also assuming a deterministic relation between cause $\mathbf{c}$ and effect $\mathbf{e}$ is the *trace method* [157, 158]. The trace method assumes a linear relationship,

$$\mathbf{e} = \mathbf{A}\mathbf{c}, \tag{B.1}$$

and formulates ICM as an "independence" between the covariance matrix $\mathbf{\Sigma}$ of $\mathbf{c}$ and the mechanism $\mathbf{A}$ (which, as for IGCI, we can again think of as a degenerate conditional $p_{\mathbf{e}|\mathbf{c}}$) via the condition

$$\tau(\mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top) = \tau(\mathbf{\Sigma})\tau(\mathbf{A}\mathbf{A}^\top) \tag{B.2}$$

where $\tau(\cdot)$ denotes the renormalized trace. Intuitively, this condition (B.2) rules out a fine-tuning of $\mathbf{A}$ to the eigenvectors of $\mathbf{\Sigma}$ which would violate the assumption of no shared information between the cause distribution (specifically, its covariance structure) and the mechanism.

As with IGCI and nonlinear ICA, it can be seen by comparing (B.1) and (2.2) that *the trace method assumes the same generative model as linear ICA* (where the cause $\mathbf{c}$ corresponds to the independent sources $\mathbf{s}$ and the

effect to the observed mixtures $\mathbf{x}$). While the focus of the present work is on nonlinear ICA, we briefly discuss the usefulness of the trace method as a constraint for achieving identifiability in a linear ICA setting.

As is clear from (B.2), the trace condition is trivially satisfied if the covariance matrix of the sources (causes) is the identity, $\mathbf{\Sigma} = \mathbf{I}$. However, as explained in Appendix A.1, in the context of linear ICA this can easily be achieved by whitening the data. As with IGCI, the trace method was developed for cause-effect inference where both variables are observed, and thus relies on the observed cause distribution being informative. This renders is unsuitable (on its own) to constrain the unsupervised representation learning problem of linear ICA problem where the sources are unobserved.

Note, however, that this is qualitatively different from the IGCI argument presented in § 3.3, as whitening on its own does not necessarily lead to independent variables, but only uncorrelated ones, and thus does not solve linear ICA—unlike the Darmois construction in the case of nonlinear ICA which also yields independent components.

B.1.2 Information geometric interpretation of the ICM principle

There is a well-established connection between IGCI and the trace method [156]. At the heart of this derivation lies an information-geometric interpretation of the ICM principle for probability distributions, which we sketch in this section. First, we need to review some basic concepts.

Background on information geometry. Information geometry [488, 489] is a discipline in which ideas from differential geometry are applied to probability theory. Probability distributions correspond to points on a Riemannian manifold, known as *statistical manifold*. Equipped with the Kullback-Leibler (KL) divergence, also called the relative entropy distance, as a premetric,[1] one can study the geometrical properties of the statistical manifold. For two probability distributions P and Q, we denote their KL divergence by $D_{KL}(P\|Q)$, which is defined for P absolutely continuous with respect to Q as:

$$D_{KL}(P\|Q) = \int dP \log \frac{dP}{dQ} \,.$$

1: A premetric on a set $\mathfrak{X}$ is a function $d : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}^+ \cup \{0\}$ such that (i) $d(x, y) \geq 0$ for all x and y in $\mathfrak{X}$ and (ii) $d(x, x) = 0$ for all $x \in \mathfrak{X}$.

An interesting property of the KL divergence is its invariance to reparametrisation. Consider an invertible transformation h, mapping random variables X and Y to $h(X)$ and $h(Y)$, respectively (the domains and codomains being arbitrary spaces, e.g., discrete or Euclidean of arbitrary dimension). Then the KL divergence between P_X and P_Y is preserved by the pushforward operation implemented by h, such that

$$D_{KL}(P_{h(X)}\|P_{h(Y)}) = D_{KL}(P_X\|P_Y) \,. \tag{B.3}$$

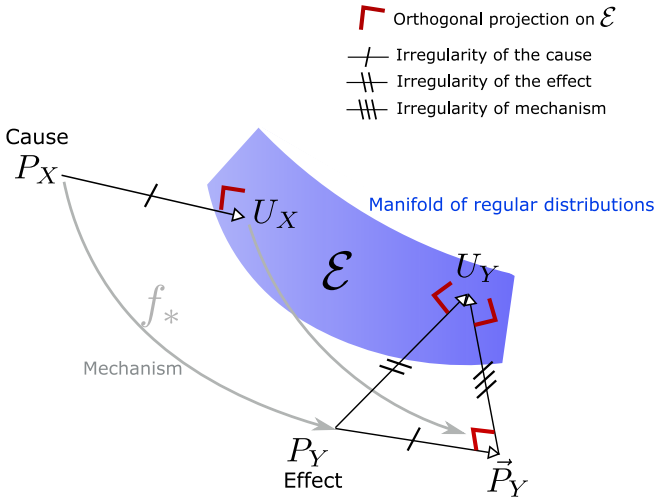


Figure B.1: Interpretation of the ICM principle as an orthogonality principle in information space. The irregularity of the effect distribution, as measured by $D_{KL}(P_Y\|U_Y)$, can be decomposed into the irregularities of the cause, as measured by $D_{KL}(P_X\|U_X)$, and the irregularity of the mechanism f, as measured by $D_{KL}(\vec{P}_Y\|U_Y)$. Here, U_X and U_Y denote the orthogonal projections of P_X and P_Y onto the manifold $\mathcal{E}$ of regular distributions, and $\vec{P}_Y$ denotes the pushforward of the regular distribution U_X via f. Note that the KL divergence is invariant to reparametrisation by invertible functions.

Interpretation of ICM as orthogonality condition in information space.

Consider a deterministic causal relationship of the form $Y := f(X)$, and denote by P_X and P_Y the marginal distributions of the cause X and the effect Y, respectively. The "irregularity" of each distribution can be quantified by evaluating their divergence to a reference set $\mathcal{E}$ of "regular" distributions,[2]

$$D_{KL}(P_X\|\mathcal{E}) = \inf_{U\in\mathcal{E}} D_{KL}(P_X\|U), \qquad D_{KL}(P_Y\|\mathcal{E}) = \inf_{U\in\mathcal{E}} D_{KL}(P_Y\|U).$$

Let us assume that these infima are reached at a unique point, their projections onto $\mathcal{E}$:

$$U_X = \arg\min_{U\in\mathcal{E}} D_{KL}(P_X\|U), \qquad U_Y = \arg\min_{U\in\mathcal{E}} D_{KL}(P_Y\|U).$$

As elaborated in [156, §4], the choice of $\mathcal{E}$ is context-dependent. For example, in the context of the trace method [157], X and Y are assumed to be n-dimensional multivariate Gaussian random vectors, and $\mathcal{E}$ is taken as the set of multivariate *isotropic* Gaussian distributions. In contrast, when IGCI is applied in contexts where the considered mechanism is a deterministic non-linear diffeomorphism, the reference distributions are typically uniform distributions [155, 490].

Overall, it can be shown that the independence postulate underlying these approaches leads to the following decomposition of the irregularity of P_Y (see [156, Thm. 2]):

$$D_{KL}(P_Y\|U_Y) = D_{KL}(P_Y\|\vec{P}_Y) + D_{KL}(\vec{P}_Y\|U_Y)$$

where $\vec{P}_Y$ denotes the distribution of $f(U_X)$, i.e., the hypothetical distribution of the effect that would be obtained if the cause X were replaced by the random variable U_X (which corresponds to the closest regularly distributed random variable to X).

Since applying the bijection f^{-1} preserves the KL divergences, see (B.3), we can obtain the equivalent relation

$$D_{KL}(P_Y\|U_Y) = D_{KL}(P_X\|U_X) + D_{KL}(\vec{P}_Y\|U_Y)\,. \tag{B.4}$$

2: Here "regular" is only meant in an intuitive sense, not implying any further mathematical notion. If $\mathcal{E}$ is the set of Gaussians, for instance, the distance from $\mathcal{E}$ measures non-Gaussianity.

This relation can be interpreted as an *orthogonality principle* in information space by considering the KL divergences as a generalization of the squared Euclidean norm for the difference vectors $\overrightarrow{P_Y U_Y}$, $\overrightarrow{P_Y \vec{P}_Y}$ and $\overrightarrow{\vec{P}_Y U_Y}$. It can thus be viewed as a Pythagorean theorem in the space of distributions, see Fig. B.1 for an illustration.

The orthogonality principle (B.4) thus captures a decomposition of the irregularity $D_{KL}(P_Y \| U_Y)$ of P_Y on the LHS into the sum of two irregularities on the RHS: the irregularity $D_{KL}(P_X \| U_X)$ of P_X, and the term $D_{KL}(\vec{P}_Y \| U_Y)$ which measures the irregularity of the mechanism f indirectly, via the "irregularity" of the distribution resulting from applying f to a regular distribution U_Y.

Overall, the decomposition (B.4) links the postulate of independence between the cause distribution, on the one hand, and the mechanism, on the other hand, to an *orthogonality of their irregularities in information space* (namely the statistical manifold of information geometry). As proposed in [156], this can be intuitively interpreted as a geometric form of independence if we assume that Nature chooses such irregularities independently of each other, and "isotropically" in a high-dimensional subspace of irregularities.

While, to date, we are not aware of similar results in the context of information geometry (i.e., on the statistical manifold), this intuition is supported by concentration of measure results in Euclidean spaces. Indeed, in high-dimensions, it is likely that two vectors are close to orthogonal if they are chosen independently according to a uniform prior [491].

We will take inspiration of the decomposition (B.4) to justify IMA in the following section.

B.1.3 Decoupling of the influences in IMA and comparison with IGCI

In contrast to Appendix B.1.2, in this section we will, for notational consistency with the main paper, assume that all distributions have a density with respect to the Lebesgue measure, and thus consider, with a slight abuse of notation, that the KL divergence is a distance between two densities on the relevant support, such that

$$D_{KL}(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx \,.$$

Overview. In line with the information-geometric interpretation of IGCI presented in Appendix B.1.2, we also consider an interpretation of IMA in information space. We consider the KL-divergence between the observed density $p_{\mathbf{x}}$ of $\mathbf{x} = \mathbf{f}(\mathbf{s})$ and an *interventional* distribution $p_{\widehat{\mathbf{x}}}$ of $\widehat{\mathbf{x}} = \widehat{\mathbf{f}}(\mathbf{s})$, resulting from a soft intervention that replaces the mixing function $\mathbf{f}$ with another mixing $\widehat{\mathbf{f}}$. We take $D_{KL}(p_{\mathbf{x}} \| p_{\widehat{\mathbf{x}}})$ as a measure of the causal effect of the soft intervention (or perturbation) that turns $\mathbf{f}$ into $\widehat{\mathbf{f}}$—similarly to how $D_{KL}(P_Y \| U_Y)$ is used as a measure of the irregularity of the effect distribution in the context of IGCI (Appendix B.1.2).

As we will show, under suitable assumptions, the functional form imposed on $\mathbf{f}$ by the IMA Principle 3.4.1 can lead to a decomposition of the *causal effect* of an intervention on the mechanism into a sum of terms, corresponding to the causal effects of separate soft interventions on the mechanisms associated to each source. In contrast, IGCI decomposes *irregularities* of the effect distribution into *two terms, one irregularity of the cause and one irregularity of the mechanism*.

Soft-interventions on the individual mechanisms. Assume $\mathbf{f}$ satisfies the IMA principle. We consider interventions performed through the element-wise transformation $\boldsymbol{\sigma}$ such that

$$\boldsymbol{\sigma} : \mathbf{s} \mapsto \begin{bmatrix} \sigma_1(s_1) \\ \vdots \\ \sigma_j(s_j) \\ \vdots \\ \sigma_n(s_n) \end{bmatrix} .$$

This can be seen as a composition of n soft interventions $\{\boldsymbol{\sigma}_j\}$ on each individual source component j, implemented through univariate smooth diffeomorphisms σ_j, such that

$$\boldsymbol{\sigma}_j : \mathbf{s} \mapsto \begin{bmatrix} s_1 \\ \vdots \\ \sigma_j(s_j) \\ \vdots \\ s_n \end{bmatrix} ,$$

and $\boldsymbol{\sigma} = \boldsymbol{\sigma}_n \circ \cdots \circ \boldsymbol{\sigma}_1$ (in arbitrary order, since the individual $\boldsymbol{\sigma}_j$ commute). This soft intervention can be seen as turning the random variable $\mathbf{s}$ into $\widehat{\mathbf{s}}$, yielding the intervened observations $\widehat{\mathbf{x}} = \mathbf{f}(\widehat{\mathbf{s}})$. Alternatively, the intervention on $\mathbf{x}$ can be implemented by replacing $\mathbf{f}$ by $\widehat{\mathbf{f}} = \mathbf{f} \circ \boldsymbol{\sigma}$—i.e., $\widehat{\mathbf{x}} = \widehat{\mathbf{f}}(\mathbf{s})$. Notably, since $\mathbf{f}$ satisfies the IMA principle, so does $\widehat{\mathbf{f}}$ (due to Proposition 3.4.2, *(ii)*, since $\boldsymbol{\sigma}$ is an element-wise nonlinearity). Moreover, the partial derivatives of the intervened function are given by

$$\frac{\partial \widehat{\mathbf{f}}}{\partial s_i}(\mathbf{s}) = \frac{\partial \mathbf{f}}{\partial s_i}(\boldsymbol{\sigma}(\mathbf{s})) \left| \frac{d\sigma_i}{ds_i} \right| (s_i) .$$

The classical change of variable formula for bijection $\mathbf{f}$ yields the expression of the pushforward density of $\mathbf{x}$ as

$$p_{\mathbf{x}}(\mathbf{x}) = |J_{\mathbf{f}}(\mathbf{f}^{-1}(\mathbf{x}))|^{-1} p_{\mathbf{s}}(\mathbf{f}^{-1}(\mathbf{x})) ,$$

and for $\widehat{\mathbf{x}}$ we get

$$p_{\widehat{\mathbf{x}}}(\widehat{\mathbf{x}}) = |J_{\widehat{\mathbf{f}}}(\widehat{\mathbf{f}}^{-1}(\widehat{\mathbf{x}}))|^{-1} p_{\mathbf{s}}(\widehat{\mathbf{f}}^{-1}(\widehat{\mathbf{x}})) ,$$

Information geometric interpretation of IMA. Let us now compute the KL divergence between the intervened and observed distribution,

$$D_{KL}(p_{\mathbf{x}} \| p_{\widehat{\mathbf{x}}}) = \int p_{\mathbf{x}}(\mathbf{x}) \log \frac{p_{\mathbf{x}}(\mathbf{x})}{p_{\widehat{\mathbf{x}}}(\mathbf{x})} d\mathbf{x} \,. \tag{B.5}$$

Expressing the density of the observed variables as a pushforward of the density of the sources, and without additional assumptions on $\mathbf{f}$ and $\widehat{\mathbf{f}}$ besides smoothness and invertibility, we get,

$$D_{KL}(p_{\mathbf{x}} \| p_{\widehat{\mathbf{x}}}) = \int \left|\mathbf{J}_{\mathbf{f}}(\mathbf{f}^{-1}(\mathbf{x}))\right|^{-1} p_{\mathbf{s}}(\mathbf{f}^{-1}(\mathbf{x})) \log \frac{\left|\mathbf{J}_{\mathbf{f}}(\mathbf{f}^{-1}(\mathbf{x}))\right|^{-1} p_{\mathbf{s}}(\mathbf{f}^{-1}(\mathbf{x}))}{\left|\mathbf{J}_{\widehat{\mathbf{f}}}(\widehat{\mathbf{f}}^{-1}(\mathbf{x}))\right|^{-1} p_{\mathbf{s}}(\widehat{\mathbf{f}}^{-1}(\mathbf{x}))} d\mathbf{x} \,.$$

We now consider a factorization of $\mathbf{s}$ over a directed acyclic graph (DAG), such that

$$p_{\mathbf{s}}(\mathbf{s}) = \prod_j p_j(s_j | \mathrm{pa}(s_j)) \,,$$

where $\mathrm{pa}(s_j)$ denotes the components associated to the parents of node j in the DAG. Because $\boldsymbol{\sigma}$ is an element-wise transformation the factorization will be the same for $p_{\widehat{\mathbf{s}}}$.

If we now additionally assume that $\mathbf{f}$ and $\widehat{\mathbf{f}}$ satisfy the IMA postulate, we get

$$D_{KL}(p_{\mathbf{x}} \| p_{\widehat{\mathbf{x}}}) = \int \left|\mathbf{J}_{\mathbf{f}}(\mathbf{f}^{-1}(\mathbf{x}))\right|^{-1} p_{\mathbf{s}}(\mathbf{f}^{-1}(\mathbf{x})) \sum_{i=1}^{n} \log \frac{\left\|\frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{f}^{-1}(\mathbf{x}))\right\|^{-1} p_i(\mathbf{f}^{-1}(\mathbf{x})_i | \mathrm{pa}(\mathbf{f}^{-1}(\mathbf{x})_i))}{\left\|\frac{\partial \widehat{\mathbf{f}}}{\partial s_i}(\widehat{\mathbf{f}}^{-1}(\mathbf{x}))\right\|^{-1} p_i(\widehat{\mathbf{f}}^{-1}(\mathbf{x})_i | \mathrm{pa}(\widehat{\mathbf{f}}^{-1}(\mathbf{x})_i))} d\mathbf{x} \,.$$

By reparameterizing the integral in terms of the source coordinates, we get (using $\widehat{\mathbf{f}}^{-1} = \boldsymbol{\sigma}^{-1} \circ \mathbf{f}^{-1}$)

$$D_{KL}(p_{\mathbf{x}} \| p_{\widehat{\mathbf{x}}}) = \sum_{i=1}^{n} \int p_{\mathbf{s}}(\mathbf{s}) \log \frac{\left\|\frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{s})\right\|^{-1} p_i(\mathbf{s}_i | \mathrm{pa}(\mathbf{s}_i))}{\left\|\frac{\partial \widehat{\mathbf{f}}}{\partial s_i}(\boldsymbol{\sigma}^{-1}(\mathbf{s}))\right\|^{-1} p_i\left(\boldsymbol{\sigma}^{-1}(\mathbf{s})_i | \mathrm{pa}(\boldsymbol{\sigma}^{-1}(\mathbf{s})_i)\right)} d\mathbf{s} \,. \tag{B.6}$$

such that the KL divergence can be written as a sum of n terms, each associated to the intervention on a mechanism $\frac{\partial \mathbf{f}}{\partial s_i}$. Positivity of these terms would suggest that we can interpret each of them as quantifying the individual contribution of a soft intervention $\boldsymbol{\sigma}_j$ applied to the original sources.

In the following, we propose a justification for the positivity of these terms *in a restricted setting where only the m leaf nodes of the graph are intervened on (with $1 \leq m \leq n$).*[3] In the special case of independent sources, all nodes are leaves and $m = n$.

3: A leaf node in a DAG is one that does not have any descendants.

Under this assumption, we consider (without loss of generality) an ordering of the nodes such that the m first nodes are the leaf nodes in the DAG. Then we argue that the terms of the right-hand side of (B.6) associated to leaf nodes ($i \leq m$) are positive, as they correspond to the expectations of KL-divergences. Indeed, taking one of the first m terms,

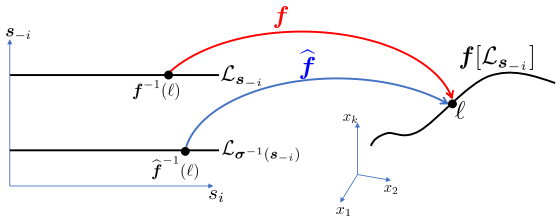


Figure B.2: Illustration of the mapping between lines in source space to a curve in observation space. $\mathscr{L}_{\mathbf{s}_{-i}}$ is the line obtained by varying s_i while keeping the value of all other sources fixed to $\mathbf{s}_{-i}$. $\mathscr{L}_{\boldsymbol{\sigma}^{-1}(\mathbf{s}_{-i})}$ is then defined by applying the transformations in $[\boldsymbol{\sigma}^{-1}]_{-i}$ to $\mathscr{L}_{\mathbf{s}_{-i}}$. Both lines are mapped to the same image line $\mathbf{f}[\mathscr{L}_{\mathbf{s}_{-i}}]$.

denoted i, we have the factorization

$$p_{\mathbf{s}}(\mathbf{s}) = p_i(s_i|\mathrm{pa}(s_i)) \prod_{j \neq i} p_j(s_j|\mathrm{pa}(s_j)) \,,$$

where $\prod_{j\neq i} p_j(s_j|\mathrm{pa}(s_j))$ does not depend on s_i because node i is a leaf node. Moreover, as non-leaf nodes are not intervened on, the transformation $\boldsymbol{\sigma}$ does not modify the value of any parent variables in these factorizations. As a consequence, the integral can be computed as an iterated integral with respect to s_i and $\mathbf{s}_{-i}$, where $\mathbf{s}_{-i}$ denotes the vector including all source variables but s_i, such that

$$\int p_{\mathbf{s}}(\mathbf{s}) \log \frac{\left\| \frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{s}) \right\|^{-1} p_i(\mathbf{s}_i|\mathrm{pa}(\mathbf{s}_i))}{\left\| \frac{\partial \widehat{\mathbf{f}}}{\partial s_i}(\boldsymbol{\sigma}^{-1}(\mathbf{s})) \right\|^{-1} p_i(\boldsymbol{\sigma}^{-1}(\mathbf{s})_i|\mathrm{pa}(\boldsymbol{\sigma}^{-1}(\mathbf{s})_i))} d\mathbf{s}$$

$$= \mathbb{E}_{\mathbf{s}_{-i}\sim\prod_{j\neq i} p_j(s_j|\mathrm{pa}(s_j))} \left[\int p(s_i|\mathrm{pa}(s_i)) \log \frac{\left\| \frac{\partial \mathbf{f}}{\partial s_i}(s_i, \mathbf{s}_{-i}) \right\|^{-1} p_i(\mathbf{s}_i|\mathrm{pa}(\mathbf{s}_i))}{\left\| \frac{\partial \widehat{\mathbf{f}}}{\partial s_i}(\sigma_i^{-1}(s_i), \boldsymbol{\sigma}^{-1}(\mathbf{s})_{-i}) \right\|^{-1} p_i(\sigma_i^{-1}(s_i)|\mathrm{pa}(s_i))} ds_i \right] .$$

As illustrated in Fig. B.2, for a fixed $\mathbf{s}_{-i}$, consider the straight line $\mathscr{L}_{\mathbf{s}_{-i}} = \{(s_i, \mathbf{s}_{-i}) : s_i \in \mathbb{R}\}$ in source space (parallel to the s_i coordinate axis). This line is mapped in observation space to the smooth curve $\mathbf{f}[\mathscr{L}_{\mathbf{s}_{-i}}]$, by $\mathbf{f}$ in a smooth invertible way. Similarly, $\widehat{\mathbf{f}} = \mathbf{f} \circ \boldsymbol{\sigma}$ maps $\mathscr{L}_{\boldsymbol{\sigma}^{-1}(\mathbf{s}_{-i})}$ to the same image curve, since $\widehat{\mathbf{f}}[\mathscr{L}_{\boldsymbol{\sigma}^{-1}(\mathbf{s}_{-i})}] = \mathbf{f} \circ \boldsymbol{\sigma}[\mathscr{L}_{\boldsymbol{\sigma}^{-1}(\mathbf{s}_{-i})}] = \mathbf{f}[\mathscr{L}_{\mathbf{s}_{-i}}]$.

By using the change of variable formula to represent the integral on $\mathbf{f}[\mathscr{L}_{\mathbf{s}_{-i}}]$ indexed by the curvilinear coordinate ℓ, we get the expression of the pushfoward distribution $\mathbf{f}_* p_i(\,.\,|\mathrm{pa}(s_i))$ on the curve $\mathbf{f}[\mathscr{L}_{\mathbf{s}_{-i}}]$

$$\left[\mathbf{f}_* p_i\left(\,.\,|\mathrm{pa}(s_i)\right)\right](\ell) = \left\| \frac{\partial \mathbf{f}}{\partial s_i}\left(\mathbf{f}^{-1}(\ell), \mathbf{s}_{-i}\right) \right\|^{-1} p_i\left(\mathbf{f}^{-1}(\ell)|\mathrm{pa}(s_i)\right) \,.$$

where, to simplify notation, $\mathbf{f}^{-1}(\ell)$ denotes in this context the coordinate s_i on $\mathscr{L}_{\mathbf{s}_{-i}}$ in bijection with the curvilinear coordinate ℓ on $\mathbf{f}[\mathscr{L}_{\mathbf{s}_{-i}}]$.

Similarly, we get the expression of the pushfoward distribution $\widehat{\mathbf{f}}_* p_i(\,.\,|\boldsymbol{\sigma}^{-1}(\mathrm{pa}(s_i)))$ from $\mathscr{L}_{\boldsymbol{\sigma}^{-1}(\mathbf{s}_{-i})}$ to the curve $\mathbf{f}[\mathscr{L}_{\mathbf{s}_{-i}}]$ (using again the fact that parent variables are not intervened on, and thus left unchanged by $\boldsymbol{\sigma}$)

$$\left[\widehat{\mathbf{f}}_* p_i(\,.\,|\boldsymbol{\sigma}^{-1}(\mathrm{pa}(s_i)))\right](\ell) = \left\| \frac{\partial \widehat{\mathbf{f}}}{\partial s_i}\left(\widehat{\mathbf{f}}^{-1}(\ell), \boldsymbol{\sigma}^{-1}(\mathbf{s})_{-i}\right) \right\|^{-1} p_i\left(\widehat{\mathbf{f}}^{-1}(\ell)|\mathrm{pa}(s_i)\right) \,.$$

These terms appear when rewriting the i-th term (for a leaf variable)

in (B.6) as a curvilinear integral:

$$\int p_{\mathbf{s}}(\mathbf{s}) \log \frac{\left\| \frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{s}) \right\|^{-1} p_i(\mathbf{s}_i|\mathrm{pa}(\mathbf{s}_i))}{\left\| \frac{\partial \widehat{\mathbf{f}}}{\partial s_i}(\boldsymbol{\sigma}^{-1}(\mathbf{s})) \right\|^{-1} p_i\left(\boldsymbol{\sigma}^{-1}(\mathbf{s})_i|\mathrm{pa}(\boldsymbol{\sigma}^{-1}(\mathbf{s})_i)\right)} d\mathbf{s}$$
$$= \mathbb{E}_{\mathbf{s}_{-i}\sim\prod_{j\neq i} p_j(s_j|\mathrm{pa}(s_j))} \left[\int \left\| \frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{f}^{-1}(\ell), \mathbf{s}_{-i}) \right\|^{-1} p_i(\mathbf{f}^{-1}(\ell) \mid \mathrm{pa}(s_i)) \right.$$
$$\left. \log \frac{\left\| \frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{f}^{-1}(\ell), \mathbf{s}_{-i}) \right\|^{-1} p_i\left(\mathbf{f}^{-1}(\ell) \mid \mathrm{pa}(s_i)\right)}{\left\| \frac{\partial \widehat{\mathbf{f}}}{\partial s_i}(\widehat{\mathbf{f}}^{-1}(\ell), \boldsymbol{\sigma}^{-1}(\mathbf{s})_{-i}) \right\|^{-1} p_i\left(\widehat{\mathbf{f}}^{-1}(\ell) \mid \mathrm{pa}(s_i)\right)} d\ell \right].$$

The inner integral term can thus be interpreted as the KL divergence between two pushforward measures defined on $\mathbf{f}_*[\mathcal{L}_{\mathbf{s}_{-i}}]$ by $\mathbf{f}$ and $\widehat{\mathbf{f}}$, that we can denote by

$$D_{KL}\left(\mathbf{f}_* p_i\left(\,.\,|\mathrm{pa}(s_i)\right) \,\|\, \widehat{\mathbf{f}}_* p\left(\,.\,|\boldsymbol{\sigma}^{-1}(\mathrm{pa}(\mathbf{s}_i))\right)\right).$$

To conclude, this implies that the causal effect of the soft intervention $\mathbf{f} \to \widehat{\mathbf{f}}$ can be decomposed as the following sum of m positive terms associated to interventions on each leaf variable, plus an additional term for the remaining non-leaf variables, which further simplifies (in comparison to (B.6)) due to the assumption that those variables are unintervened.

$$D_{KL}(p_{\mathbf{x}} \,\|\, p_{\widehat{\mathbf{x}}}) = \sum_{i=1}^{m} \mathbb{E}_{\mathbf{s}_{-i}\sim\prod_{j\neq i} p_j(s_j|\mathrm{pa}(s_j))} \left[D_{KL}\left(\mathbf{f}_* p\left(\,.\,|\mathrm{pa}(s_i)\right) \,\|\, \widehat{\mathbf{f}}_* p\left(\,.\,|\boldsymbol{\sigma}^{-1}(\mathrm{pa}(s_i))\right)\right)\right]$$
$$+ \sum_{i>m} \int p_{\mathbf{s}}(\mathbf{s}) \log \frac{\left\| \frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{s}) \right\|^{-1}}{\left\| \frac{\partial \widehat{\mathbf{f}}}{\partial s_i}(\boldsymbol{\sigma}^{-1}(\mathbf{s})) \right\|^{-1}} d\mathbf{s}\,. \quad \text{(B.7)}$$

This expression suggests that the KL-divergences appearing in the first m terms each reflect the causal effect of an intervention on the mechanism at the level of one single source coordinate i, turning $\frac{\partial \mathbf{f}}{\partial s_i}$ into $\frac{\partial \widehat{\mathbf{f}}}{\partial s_i}$. When the sources are jointly independent, we have $m = n$ and the right hand side of (B.7) contains only positive terms. An interesting direction for future work would be to analyse the remaining term in the case of non unconditionally independent sources.

In contrast to the decomposition (B.4) in the context of IGCI, the IMA decomposition (B.7) involves m (expectations of) KL-divergence terms instead of two, each related to the intervention on the part of the mechanism $\frac{\partial \mathbf{f}}{\partial s_i}$ that reflects the influence of a single source.

B.1.4 Independence of cause and mechanism and IMA

We now discuss an example in which a formalisation of the principle of independence of cause and mechanism [157] is violated, and one in which the IMA principle is violated.

Violations of independence of cause and mechanism

In the context of the Trace method [157], used in causal discovery, a technical example of fine-tuning can be constructed by taking a vector of i.i.d. random variables with arbitrary (not diagonal) covariance matrix Σ as the cause, and by constructing the mechanism as a whitening matrix, turning the cause variables into uncorrelated (effect) variables. By doing so, the singular values and singular vectors of the matrix (the mechanism) are fine-tuned to the input covariance matrix (a property of the cause distribution), and such fine-tuning can be quantified via the Trace method (see [157], Section 1).

Violations of the IMA principle

Technical example. As mentioned in § 3.3, an example of a mixing function $\mathbf{f}$ which is non-generic according to the IMA principle is an autoregressive function, for example an autoregressive normalising flow [71], where the k-th component of the observations only depends on the k-th sources: intuitively, this would correspond to the unlikely cocktail-party setting where the k-th microphone only picks up the voices of the first speakers. More precisely, as we show in Lemma B.2.1, this leads to positive C_{IMA} value for such mixing.

Pictorial example: Violations of the IMA principle in a cocktail party. A cocktail party (Fig. 3.2, left) may violate our IMA principle when the locations of several speakers and the room acoustics have been fine tuned to one another. This is for example the case in concert halls where the acoustics of the room have been fine-tuned to the position and configuration of multiple locations on the stage, where the sources (i.e., the voices of the actors or singers) are emitted—in order to make the listening experience as homogeneous as possible across the spectators (that is, the influence of each of the sources on the different listeners should not differ too much). This would lead to an increase in collinearity between the columns of the mixing's Jacobian, thus violating the IMA principle.

Additionally, we recall that the ICM principle is often informally introduced by referencing the fine-tuning and non-generic viewpoints giving rise to certain visual illusions, such as the Beuchet chair (see [106], Section 2); in a similar vein, we can imagine that violations of the IMA principle in the cocktail-party setting may be related to illusions in binaural hearing such as for example the Franssen effect, where the listener is tricked into incorrectly localizing a sound [492].

B.2 Proofs

We now provide the proofs of all theoretical results in Chapter 3.

B.2.1 Proof of Proposition 3.4.1

Before giving the proof, it is useful to rewrite the local IMA constrast (3.3) as follows:

$$\begin{aligned} c_{\text{IMA}}(\mathbf{f},) &= \sum_{i=1}^{n} \log \left\| \frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{s}) \right\| - \log |\mathbf{J_f}(\mathbf{s})| \\ &= \frac{1}{2} \left(\log \left| diag \left(\mathbf{J}_\mathbf{f}^\top(\mathbf{s}) \mathbf{J_f}(\mathbf{s}) \right) \right| - \log \left| \mathbf{J}_\mathbf{f}^\top(\mathbf{s}) \mathbf{J_f}(\mathbf{s}) \right| \right) \\ &= \frac{1}{2} D_{KL}^{\text{left}} \left(\mathbf{J}_\mathbf{f}^\top(\mathbf{s}) \mathbf{J_f}(\mathbf{s}) \right) , \end{aligned} \tag{B.8}$$

where the quantity in (B.8) is called the left KL measure of diagonality of the matrix $\mathbf{J}_\mathbf{f}^\top(\mathbf{s})\mathbf{J_f}(\mathbf{s})$ [172] (see Remark 3.4.1):

$$\begin{aligned} D_{KL}^{\text{left}}(\mathbf{A}) &= -\log |(\mathrm{d}iag(\mathbf{A}))^{-\frac{1}{2}} \mathbf{A} (\mathrm{d}iag(\mathbf{A}))^{-\frac{1}{2}}| \\ &= \log |\mathrm{d}iag(\mathbf{A})| - \log |\mathbf{A}| . \end{aligned}$$

From (B.8), it can be seen that $c_{\text{IMA}}(\mathbf{f},)$ is a function of $\mathbf{J_f}(\mathbf{s})$ only through $\mathbf{J}_\mathbf{f}^\top(\mathbf{s})\mathbf{J_f}(\mathbf{s})$.

Proposition 3.4.1 (Properties of $c_{\text{IMA}}(\mathbf{f},)$) *The local IMA contrast $c_{IMA}(\mathbf{f},)$ defined in (3.3) satisfies:*

(i) $c_{IMA}(\mathbf{f},) \geq 0$, with equality if and only if all columns $\partial \mathbf{f}/\partial s_i(\mathbf{s})$ of $\mathbf{J_f}(\mathbf{s})$ are orthogonal.

(ii) $c_{IMA}(\mathbf{f},)$ is invariant to left multiplication of $\mathbf{J_f}(\mathbf{s})$ by an orthogonal matrix and to right multiplication by permutation and diagonal matrices.

Proof. For ease of exposition, we denote the value of the Jacobian of $\mathbf{f}$ evaluated at the point $\mathbf{s}$ by $\mathbf{J_f}(\mathbf{s}) = \mathbf{W}$. The two properties can then be proved as follows:

(i) This is a consequence of Hadamard's inequality, applied to the expression on the RHS of (3.3), which states that, for a matrix $\mathbf{W}$ with columns $\mathbf{w}_i$, $\sum_{i=1}^{n} \log \|\mathbf{w}_i\| \geq \log |\mathbf{W}|$; equality in Hadamard's inequality is achieved iff. the vectors $\mathbf{w}_i$ are orthogonal.

(ii) We split the proof in three parts.

a. *Invariance to left multiplication by an orthogonal matrix:* Let $\tilde{\mathbf{W}} = \mathbf{OW}$, with $\mathbf{O}$ an orthogonal matrix, i.e., $\mathbf{OO}^\top = \mathbf{I}$. Then the property follows from writing $c_{\text{IMA}}(\mathbf{f},)$ as in (B.8):

$$\frac{1}{2} D_{KL}^{\text{left}}(\tilde{\mathbf{W}}^\top \tilde{\mathbf{W}}) = \frac{1}{2} D_{KL}^{\text{left}}(\mathbf{W}^\top \mathbf{O}^\top \mathbf{OW}) = \frac{1}{2} D_{KL}^{\text{left}}(\mathbf{W}^\top \mathbf{IW}) = \frac{1}{2} D_{KL}^{\text{left}}(\mathbf{W}^\top \mathbf{W})$$

b. *Invariance to right multiplication by a permutation matrix:* Let $\tilde{\mathbf{W}} = \mathbf{WP}$, with $\mathbf{P}$ a permutation matrix. Then $\tilde{\mathbf{W}}$ is just $\mathbf{W}$ with permuted columns. Clearly, the sum of the log-column-norms does not change by changing the order of the summands. Further, $\log |\tilde{\mathbf{W}}| = \log |\mathbf{W}| + \log |\mathbf{P}| = \log |\mathbf{W}|$, because the absolute value of the determinant of a permutation matrix is one.

c. *Invariance to right multiplication by a diagonal matrix*: Let $\tilde{\mathbf{W}} = \mathbf{WD}$, with $\mathbf{D}$ a diagonal matrix. Consider the two

terms on the RHS of (3.3). For the first term, we know that the columns of $\tilde{\mathbf{W}}$ are scaled versions of the columns of $\mathbf{W}$, that is $\tilde{\mathbf{w}}_i = d_i \mathbf{w}_i$, where d_i denotes the i^{th} diagonal element of $\mathbf{D}$. Then $\|\tilde{\mathbf{w}}_i\| = |d_i| \, \|\mathbf{w}_i\|$. For the second term, we use the decomposition of the determinant:

$$\log|\tilde{\mathbf{W}}| = \log|\mathbf{W}| + \log|\mathbf{D}| = \log|\mathbf{W}| + \sum_{i=1}^{n} \log|d_i|.$$

Taken together, we obtain

$$\begin{aligned}
\sum_{i=1}^{n} \log\|\tilde{\mathbf{w}}_i\| - \log|\tilde{\mathbf{W}}| &= \sum_{i=1}^{n} \log\left(|d_i| \, \|\mathbf{w}_i\|\right) - \left(\log|\mathbf{W}| + \sum_{i=1}^{n} \log|d_i|\right) \\
&= \sum_{i=1}^{n} \log\|\mathbf{w}_i\| + \sum_{i=1}^{n} \log|d_i| - \log|\mathbf{W}| - \sum_{i=1}^{n} \log|d_i| \\
&= \sum_{i=1}^{n} \log\|\mathbf{w}_i\| - \log|\mathbf{W}|
\end{aligned}$$

□

B.2.2 Proof of Proposition 3.4.2

Proposition 3.4.2 (Properties of $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}})$) *The global IMA contrast $C_{IMA}(\mathbf{f}, p_{\mathbf{s}})$ from (3.4) satisfies:*

(i) $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) \geq 0$, *with equality iff.* $\mathbf{J_f}(\mathbf{s}) = \mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s})$ *almost surely w.r.t.* $p_{\mathbf{s}}$, *where* $\mathbf{O}(\mathbf{s}), \mathbf{D}(\mathbf{s}) \in \mathbb{R}^{n \times n}$ *are orthogonal and diagonal matrices, respectively;*

(ii) $C_{\text{IMA}}(\mathbf{f}, p_{\mathbf{s}}) = C_{\text{IMA}}(\tilde{\mathbf{f}}, p_{\tilde{\mathbf{s}}})$ *for any* $\tilde{\mathbf{f}} = \mathbf{f} \circ \mathbf{h}^{-1} \circ \mathbf{P}^{-1}$ *and* $\tilde{\mathbf{s}} = \mathbf{P}\mathbf{h}(\mathbf{s})$, *where* $\mathbf{P} \in \mathbb{R}^{n \times n}$ *is a permutation and* $\mathbf{h}(\mathbf{s}) = (h_1(s_1), \ldots, h_n(s_n))$ *an invertible element-wise function.*

Proof. The properties can be proved as follows:

(i) From property *(i)* of Proposition 3.4.1, we know that $c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) \geq 0$. Hence, $C_{\text{IMA}}(\mathbf{f}, p(\mathbf{s})) \geq 0$ follows as a direct consequence of integrating the non-negative quantity $c_{\text{IMA}}(\mathbf{f}, \mathbf{s})$.
Equality is attained iff. $c_{\text{IMA}}(\mathbf{f}, \mathbf{s}) = 0$ almost surely w.r.t. $p_{\mathbf{s}}$, which according to property *(i)* of Proposition 3.4.1 occurs iff. the columns of $\mathbf{J_f}(\mathbf{s})$ are orthogonal almost surely w.r.t. $p_{\mathbf{s}}$.
It remains to show that this is the case iff. $\mathbf{J_f}(\mathbf{s})$ can be written as $\mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s})$, with $\mathbf{O}(\mathbf{s})$ and $\mathbf{D}(\mathbf{s})$ orthogonal and diagonal matrices, respectively. (To avoid confusion, note that *orthogonal columns* need not have unit norm, whereas an *orthogonal matrix* $\mathbf{O}$ satisfies $\mathbf{O}\mathbf{O}^\top = \mathbf{I}$.)
The *if* is clear since right multiplication by a diagonal matrix merely re-scales the columns, and hence does not affect their orthogonality.
For the *only if*, let $\mathbf{J_f}(\mathbf{s})$ be any matrix with orthogonal columns $\mathbf{j}_i(\mathbf{s})$, $\mathbf{j}_i(\mathbf{s})^\top \mathbf{j}_j(\mathbf{s}) = 0, \forall i \neq j$, and denote the column norms by $d_i(\mathbf{s}) = ||\mathbf{j}_i(\mathbf{s})||$. Further denote the normalised columns of $\mathbf{J_f}(\mathbf{s})$ by $\mathbf{o}_i(\mathbf{s}) = \mathbf{j}_i(\mathbf{s})/d_i(\mathbf{s})$ and let $\mathbf{O}(\mathbf{s})$ and $\mathbf{D}(\mathbf{s})$ be the orthogonal and

diagonal matrices with columns $\mathbf{o}_i(\mathbf{s})$ and diagonal elements $d_i(\mathbf{s})$, respectively. Then $\mathbf{J}_\mathbf{f}(\mathbf{s}) = \mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s})$.

(ii) Let $\tilde{\mathbf{f}} = \mathbf{f} \circ \mathbf{h}^{-1} \circ \mathbf{P}^{-1}$ and $\tilde{\mathbf{s}} = \mathbf{P}\mathbf{h}(\mathbf{s})$, where $\mathbf{P} \in \mathbb{R}^{n \times n}$ is a permutation matrix and $\mathbf{h}(\mathbf{s}) = (h_1(s_1), \ldots, h_n(s_n))$ is an invertible element-wise function. Then

$$C_{\text{IMA}}(\tilde{\mathbf{f}}, p_\mathbf{s}t) = \int c_{\text{IMA}}(\tilde{\mathbf{f}}, \tilde{\mathbf{s}}) p_\mathbf{s}t(\tilde{\mathbf{s}}) d\tilde{\mathbf{s}} = \int c_{\text{IMA}}(\tilde{\mathbf{f}}, \tilde{\mathbf{s}}) p_\mathbf{s}(\mathbf{s}) d\mathbf{s} \quad \text{(B.9)}$$

where, for the second equality, we have used the fact that

$$p_\mathbf{s}t(\tilde{\mathbf{s}})d\tilde{\mathbf{s}} = p_\mathbf{s}(\mathbf{s})d\mathbf{s}\,.$$

since $\mathbf{P} \circ \mathbf{h}$ is an invertible tranformation (see, e.g., [493]). It thus suffices to show that

$$c_{\text{IMA}}(\tilde{\mathbf{f}}, \tilde{\mathbf{s}}) = c_{\text{IMA}}(\mathbf{f},)\,. \quad \text{(B.10)}$$

at any point $\tilde{\mathbf{s}} = \mathbf{P}\mathbf{h}(\mathbf{s})$. To show this, we write

$$\begin{aligned}
\mathbf{J}_{\tilde{\mathbf{f}}}(\tilde{\mathbf{s}}) &= \mathbf{J}_{\mathbf{f}\circ\mathbf{h}^{-1}\circ\mathbf{P}^{-1}}(\mathbf{P}\mathbf{h}(\mathbf{s})) \\
&= \mathbf{J}_{\mathbf{f}\circ\mathbf{h}^{-1}}\left(\mathbf{P}^{-1}\mathbf{P}\mathbf{h}(\mathbf{s})\right)\, \mathbf{J}_{\mathbf{P}^{-1}}\left(\mathbf{P}\mathbf{h}(\mathbf{s})\right) \\
&= \mathbf{J}_{\mathbf{f}\circ\mathbf{h}^{-1}}(\mathbf{h}(\mathbf{s}))\, \mathbf{J}_{\mathbf{P}^{-1}}(\mathbf{P}\mathbf{h}(\mathbf{s})) \\
&= \mathbf{J}_\mathbf{f}(\mathbf{h}^{-1} \circ \mathbf{h}(\mathbf{s}))\, \mathbf{J}_{\mathbf{h}^{-1}}(\mathbf{h}(\mathbf{s}))\, \mathbf{J}_{\mathbf{P}^{-1}}(\mathbf{P}\mathbf{h}(\mathbf{s})) \\
&= \mathbf{J}_\mathbf{f}(\mathbf{s})\, \mathbf{D}(\mathbf{s})\mathbf{P}^{-1} \qquad \text{(B.11)}
\end{aligned}$$

where we have repeatedly used the chain rule for Jacobians, as well as that $\mathbf{P}^{-1}\mathbf{P} = \mathbf{I}$; that permutation is a linear operation, so $\mathbf{J}_\mathbf{P}(\mathbf{s}) = \mathbf{P}$ for any $\mathbf{s}$; and that $\mathbf{h}$ (and thus $\mathbf{h}^{-1}$) is an element-wise transformation, so the Jacobian $\mathbf{J}_{\mathbf{h}^{-1}}$ is a diagonal matrix $\mathbf{D}(\mathbf{s})$.
The equality in (B.10) then follows from (B.11) by applying property *(ii)* of Proposition 3.4.1, according to which c_{IMA} is invariant to right multiplication of the Jacobian $\mathbf{J}_\mathbf{f}(\mathbf{s})$ by diagonal and permutation matrices.
Substituting (B.10) into the RHS of (B.9), we finally obtain

$$C_{\text{IMA}}(\tilde{\mathbf{f}}, p_\mathbf{s}t) = C_{\text{IMA}}(\mathbf{f}, p_\mathbf{s}).$$

□

B.2.3 Remark on a similar condition to IMA, expressed in terms of the rows of the Jacobian

We remark that the condition imposed by the IMA Principle 3.4.1 needs to be expressed in terms of the columns of the Jacobian, and would not lead to a criterion with desirable properties for BSS if it were instead expressed in terms of its rows (which correspond to gradients of the $f_i(\mathbf{s})$). One way to justify this is that, for the same condition expressed on the rows of the Jacobian, that is

$$\sum_{i=1}^{n} \log \|\nabla f_i(\mathbf{s})\| - \log |\mathbf{J}_\mathbf{f}(\mathbf{s})| = 0\,,$$

property *(ii)* of Proposition 3.4.1 would not hold (because invariance would hold w.r.t. right, not left, multiplication with a diagonal matrix). As a consequence, the resulting global contrast would not be blind to reparametrisation of the source variables by permutation and element-wise invertible transformations, thereby not being a good contrast in the context of blind source separation.

B.2.4 Proof of Thm. 3.4.3

Before proving the main theorem, we first introduce some additional details on the Jacobian of the Darmois construction [70] which will be important for the proof.

Jacobian of the Darmois construction for $n = 2$. Consider the Darmois construction for $n = 2$,

$$y_1 = g_1^{\mathrm{D}}(x_1) := F_{X_1}(x_1) = \mathbb{P}_{X_1}(X_1 \leq x_1)$$
$$y_2 = g_2^{\mathrm{D}}(y_1, x_2) := F_{X_2|Y_1=y_1}(x_2) = \mathbb{P}_{X_2|Y_1=y_1}(X_2 \leq x_2 | Y_1 = y_1)$$

Its Jacobian takes the form

$$\mathbf{J}_{\mathbf{g}^{\mathrm{D}}}(\mathbf{x}) = \begin{pmatrix} p(x_1) & 0 \\ c_{21}(\mathbf{x}) & p(x_2|x_1) \end{pmatrix}, \tag{B.12}$$

where

$$c_{21}(\mathbf{x}) = \frac{\partial}{\partial x_1} \int_{-\infty}^{x_2} p(x_2'|x_1) dx_2' .$$

Jacobian of the Darmois construction: general case. In the general case, the Jacobian of the Darmois construction will be

$$\mathbf{J}_{\mathbf{g}^{\mathrm{D}}}(\mathbf{x}) = \begin{pmatrix} p(x_1) & \cdots & 0 \\ & \ddots & \vdots \\ \mathbf{C}(\mathbf{x}) & & p(x_n|x_1, \ldots, x_{n-1}) \end{pmatrix} \tag{B.13}$$

where the components $c_{ji}(\mathbf{x}_{1:j})$ of $\mathbf{C}(\mathbf{x})$ for all $i < j$ are defined by

$$c_{ji}(\mathbf{x}_{1:j}) = \frac{\partial}{\partial x_i} \int_{-\infty}^{x_j} p(x_j'|\mathbf{x}_{1:j-1}) dx_j' .$$

It is additionally useful to introduce the following lemmas.

Lemma B.2.1 *A function $\mathbf{f}$ with triangular Jacobian has $C_{IMA}(\mathbf{f}, p_{\mathbf{s}}) = 0$ iff. its Jacobian is diagonal almost everywhere. Otherwise, $C_{IMA}(\mathbf{f}, p_{\mathbf{s}}) > 0$.*

Proof. Let $\mathbf{f}$ have lower triangular Jacobian at $\mathbf{s}$, and denote $\mathbf{J}_{\mathbf{f}}(\mathbf{s}) = \mathbf{W}$. Then we have

$$c_{\mathrm{IMA}}(\mathbf{f},) = \sum_{i=1}^{n} \log \left(\sqrt{\sum_{j=i}^{n} w_{ji}^2} \right) - \sum_{i=1}^{n} \log |w_{ii}| ,$$

where $w_{ji} = [\mathbf{W}]_{ji}$. Since the logarithm is a strictly monotonically increasing function and since

$$\sqrt{\sum_{j=1}^{n} w_{ji}^2} \geq |w_{ii}|\,,$$

with equality iff. $w_{ji} = 0, \forall j \neq i$ (i.e., iff. $\mathbf{W}$ is a diagonal matrix), we must have $c_{\text{IMA}}(\mathbf{f}_,) = 0$ iff. $\mathbf{W}$ is diagonal.

$C_{\text{IMA}}(\mathbf{f}, p_\mathbf{s})$ is therefore equal to zero iff. $\mathbf{f}$ has diagonal Jacobian almost everywhere, and it is strictly larger than zero otherwise. □

Lemma B.2.2 *A smooth function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ whose Jacobian is diagonal everywhere is an element-wise function, $\mathbf{f}(\mathbf{s}) = (f_1(s_1), \ldots, f_n(s_n))$.*

Proof. Let $\mathbf{f}$ be a smooth function with diagonal Jacobian everywhere.

Consider the function $f_i(\mathbf{s})$ for any $i \in \{1, \ldots, n\}$. Suppose *for a contradiction* that f_i depends on s_j for some $j \neq i$. Then there must be at least one point $\mathbf{s}^*$ such that $\partial f_i / \partial s_j(\mathbf{s}^*) \neq 0$. However, this contradicts the assumption that $\mathbf{J}_\mathbf{f}$ is diagonal everywhere (since $\partial f_i / \partial s_j$ is an off-diagonal element for $i \neq j$). Hence, f_i can only depend on s_i for all i, i.e., $\mathbf{f}$ is an element wise function. □

We can now restate and prove Thm. 3.4.3.

Theorem 3.4.3 *Assume the data generating process in* (2.1) *and assume that $x_i \not\perp\!\!\!\perp x_j$ for some $i \neq j$. Then any Darmois solution $(\mathbf{f}^D, p_\mathbf{u})$ based on $\mathbf{g}^D$ as defined in* (2.15) *satisfies $C_{IMA}(\mathbf{f}^D, p_\mathbf{u}) > 0$. Thus a solution satisfying $C_{IMA}(\mathbf{f}, p_\mathbf{s}) = 0$ can be distinguished from $(\mathbf{f}^D, p_\mathbf{u})$ based on the contrast C_{IMA}.*

Proof. First, the Jacobian $\mathbf{J}_{\mathbf{g}^\text{D}}(\mathbf{x})$ of the Darmois construction $\mathbf{g}^\text{D}$ is lower triangular $\forall \mathbf{x}$, see (B.13).

Because CDFs are monotonic functions (strictly monotonically increasing given our assumptions on $\mathbf{f}$ and $p_\mathbf{s}$), $\mathbf{g}^\text{D}$ is invertible.

We can thus apply the inverse function theorem (with $\mathbf{f}^\text{D} = (\mathbf{g}^\text{D})^{-1}$) to write

$$\mathbf{J}_{\mathbf{f}^\text{D}}(\mathbf{y}) = \left(\mathbf{J}_{\mathbf{g}^\text{D}}(\mathbf{x})\right)^{-1}$$

Since the inverse of a lower triangular matrix is lower triangular, we conclude that $\mathbf{J}_{\mathbf{f}^\text{D}}(\mathbf{y})$ is lower triangular for all $\mathbf{y} = \mathbf{g}^\text{D}(\mathbf{x})$.

Now, according to Lemma B.2.1, we have $C_{\text{IMA}}(\mathbf{f}^\text{D}, p_\mathbf{u}) > 0$, unless $\mathbf{J}_{\mathbf{f}^\text{D}}$ is diagonal almost everywhere.

Suppose *for a contradiction* that $\mathbf{J}_{\mathbf{f}^\text{D}}$ is diagonal almost everywhere.

Since $\mathbf{f}$ and $p_\mathbf{s}$ are smooth by assumption, so is the push-forward $p_\mathbf{x} = \mathbf{f}_* p_\mathbf{s}$, and thus also $\mathbf{g}^\text{D}$ (CDF of a smooth density) and its inverse $\mathbf{f}^\text{D}$. Hence, the partial derivatives $\partial f_i^\text{D} / \partial y_j$, i.e., the elements of $\mathbf{J}_{\mathbf{f}^\text{D}}$ are continuous.

Consider an off-diagonal element $\partial f_i^\text{D} / \partial y_j$ for $i \neq j$. Since these are zero almost everywhere, and because continuous functions which are zero

almost everywhere must be zero everywhere, we conclude that $\partial f_i^{\mathrm{D}}/\partial y_j = 0$ everywhere for $i \neq j$, i.e., the Jacobian $\mathbf{J}_{\mathbf{f}^{\mathrm{D}}}$ is *diagonal everywhere*.

Hence, we conclude from Lemma B.2.2 that $\mathbf{f}^{\mathrm{D}}$ must be an element-wise function, $\mathbf{f}^{\mathrm{D}}(\mathbf{y}) = (f_1^{\mathrm{D}}(y_1), ..., f_1^{\mathrm{D}}(y_n))$.

Since $\mathbf{y}$ has independent components by construction, it follows that $x_i = f_i^{\mathrm{D}}(y_i)$ and $x_j = f_j^{\mathrm{D}}(y_j)$ are independent for any $i \neq j$.

However, this constitutes a contradiction to the assumption that $x_i \not\perp\!\!\!\perp x_j$ for some x_j.

We conclude that $\mathbf{J}_{\mathbf{f}^{\mathrm{D}}}$ cannot be diagonal almost everywhere, and hence, by Lemma B.2.1, we must have $C_{\mathrm{IMA}}(\mathbf{f}^{\mathrm{D}}, p_{\mathbf{u}}) > 0$. □

B.2.5 Proof of Corollary 3.4.4

Corollary 3.4.4 *Under assumptions of Thm. 3.4.3, if additionally* $\mathbf{f}$ *is a conformal map, then* $(\mathbf{f}, p_{\mathbf{s}}) \in \mathcal{M}_{\mathrm{IMA}}$ *for any* $p_{\mathbf{s}} \in \mathcal{P}$ *due to Proposition 3.4.2 (i), see Defn. 3.4.3. Based on Thm. 3.4.3,* $(\mathbf{f}, p_{\mathbf{s}})$ *is thus distinguishable from Darmois solutions* $(\mathbf{f}^D, p_{\mathbf{u}})$.

Proof. The proof follows from property *(i)* of Proposition 3.4.2: by definition, the Jacobian of conformal maps at any point $\mathbf{s}$ can be written as $\mathbf{O}(\mathbf{s})\lambda(\mathbf{s})$, with $\lambda : \mathbb{R}^n \to \mathbb{R}$, which is a special case of $\mathbf{O}(\mathbf{s})\mathbf{D}(\mathbf{s})$, with $\mathbf{D}(\mathbf{s}) = \lambda(\mathbf{s})\mathbf{I}$. □

B.2.6 Proof of Corollary 3.4.5

Corollary 3.4.5 *Consider a linear ICA model,* $\mathbf{x} = \mathbf{A}\mathbf{s}$*, with* $\mathbb{E}[\mathbf{s}^\top \mathbf{s}] = \mathbf{I}$*, and* $\mathbf{A} \in O(n)$ *an orthogonal, non-trivial mixing matrix, i.e., not the product of a diagonal and a permutation matrix* $\mathbf{DP}$*. If at most one of the* s_i *is Gaussian, then* $C_{\mathrm{IMA}}(\mathbf{A}, p_{\mathbf{s}}) = 0$ *and* $C_{\mathrm{IMA}}(\mathbf{f}^D, p_{\mathbf{u}}) > 0$.

Proof. Since, by assumption, the mixing matrix is non-trivial (i.e., not the product of a diagonal and permutation matrix), and at most one of the s_i is Gaussian, according to Thm. 2.2.1 there must be at least one pair x_i, x_j, with $i \neq j$, such that $x_i \not\perp\!\!\!\perp x_j$.

We can then use the same argument as in the proof of Thm. 3.4.3 to show that the Darmois construction has nonzero C_{IMA}, whereas the linear orthogonal transformation $\mathbf{A}$ has orthogonal Jacobian, and thus $C_{\mathrm{IMA}} = 0$ by property *(i)* of Proposition 3.4.2. □

B.2.7 Proof of Thm. 3.4.6

Theorem 3.4.6 *Let* $(\mathbf{f}, p_{\mathbf{s}}) \in \mathcal{M}_{\mathrm{IMA}}$ *and assume that* $\mathbf{f}$ *is a conformal map. Given* $\mathbf{R} \in O(n)$*, assume additionally that* $\exists$ *at least one non-Gaussian* s_i *whose associated canonical basis vector* $\mathbf{e}_i$ *is not transformed by* $\mathbf{R}^{-1} = \mathbf{R}^\top$ *into another canonical basis vector* $\mathbf{e}_j$*. Then* $C_{\mathrm{IMA}}(\mathbf{f} \circ \mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}}), p_{\mathbf{s}}) > 0$.

Proof. Recall the definition

$$\mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}}) = \mathbf{F}_{\mathbf{s}}^{-1} \circ \mathbf{\Phi} \circ \mathbf{R} \circ \mathbf{\Phi}^{-1} \circ \mathbf{F}_{\mathbf{s}}.$$

For notational convenience, we denote $\sigma = \mathbf{\Phi}^{-1} \circ \mathbf{F}_{\mathbf{s}}$ and write

$$\mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}}) = \sigma^{-1} \circ \mathbf{R} \circ \sigma.$$

Note that, since both $\mathbf{F}_{\mathbf{s}}$ and $\mathbf{\Phi}$ are element-wise transformations, so is σ.

First, by using property *(ii)* of Proposition 3.4.2 (invariance of C_{IMA} to element-wise transformation), we obtain

$$C_{\text{IMA}}(\mathbf{f} \circ \mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}}), p_{\mathbf{s}}) = C_{\text{IMA}}(\mathbf{f} \circ \sigma^{-1} \circ \mathbf{R} \circ \sigma, p_{\mathbf{s}}) = C_{\text{IMA}}(\mathbf{f} \circ \sigma^{-1} \circ \mathbf{R}, p_{\mathbf{z}}) \,,$$

with $\mathbf{z} = \sigma(\mathbf{s})$ such that $p_{\mathbf{z}}$ is an isotropic Gaussian distribution.

Suppose *for a contradiction* that $C_{\text{IMA}}(\mathbf{f} \circ \sigma^{-1} \circ \mathbf{R}, p_{\mathbf{z}}) = 0$.

According to property *(i)* of Proposition 3.4.2, this entails that the matrix

$$\mathbf{J}_{\mathbf{f}\circ\sigma^{-1}\circ\mathbf{R}}(\mathbf{z})^\top \mathbf{J}_{\mathbf{f}\circ\sigma^{-1}\circ\mathbf{R}}(\mathbf{z}) = \mathbf{R}^\top \mathbf{J}_{\sigma^{-1}}(\mathbf{z})^\top \mathbf{J}_{\mathbf{f}}(\sigma^{-1}(\mathbf{z}))^\top \mathbf{J}_{\mathbf{f}}(\sigma^{-1}(\mathbf{z})) \mathbf{J}_{\sigma^{-1}}(\mathbf{z}) \mathbf{R} \tag{B.14}$$

is diagonal almost surely w.r.t. $p_{\mathbf{z}}$. Moreover, smoothness of $p_{\mathbf{s}}$ and $\mathbf{f}$ implies the matrix expression of (B.14) is a continuous function of $\mathbf{z}$. Thus (B.14) actually needs to be diagonal for all $\mathbf{z} \in \mathbb{R}^n$, i.e., *everywhere* (c.f., the argument used in the proof of Thm. 3.4.3, l.1008–1013).

Since $(\mathbf{f}, p_{\mathbf{s}}) \in \mathscr{M}_{\text{IMA}}$ by assumption, by property *(i)* of Proposition 3.4.2, the inner term on the RHS of (B.14),

$$\mathbf{J}_{\mathbf{f}}(\sigma^{-1}(\mathbf{z}))^\top \mathbf{J}_{\mathbf{f}}(\sigma^{-1}(\mathbf{z})),$$

is diagonal. Moreover, since σ is an element-wise transformation, $\mathbf{J}_{\sigma^{-1}}(\mathbf{z})^\top$ and $\mathbf{J}_{\sigma^{-1}}(\mathbf{z})$ are also diagonal. Taken together, this implies that

$$\mathbf{J}_{\sigma^{-1}}(\mathbf{z}) \mathbf{J}_{\mathbf{f}}(\sigma^{-1}(\mathbf{z}))^\top \mathbf{J}_{\mathbf{f}}(\sigma^{-1}(\mathbf{z})) \mathbf{J}_{\sigma^{-1}}(\mathbf{z}) \tag{B.15}$$

is diagonal (i.e., (B.14) is of the form $\mathbf{R}^\top \mathbf{D}(\mathbf{z}) \mathbf{R}$ for some diagonal matrix $\mathbf{D}(\mathbf{z})$).

Without loss of generality, we assume the first component s_1 of $\mathbf{s}$ is non-Gaussian and satisfies the assumptions stated relative to $\mathbf{R}$ (axis not invariant nor sent to another canonical axis).

Now, since both the Gaussian CDF $\mathbf{\Phi}$ and the CDF $\mathbf{F}_{\mathbf{s}}$ are smooth (the latter by the assumption that of $p_{\mathbf{s}}$ is a smooth density), σ is a smooth function, and thus has continuous partial derivatives.

By continuity of the partial derivative, the first diagonal element $\frac{\partial \sigma_1^{-1}}{\partial z_1}$ of $\mathbf{J}_{\sigma^{-1}}$ must be strictly monotonic in a neighborhood of some z_1^0 (otherwise σ_1 would be an affine transformation, which would contradict non-Gaussianity of s_1).

On the other hand, our assumptions relative to $\mathbf{R}$ entail that there are at least two non-vanishing coefficients in the first row of $\mathbf{R}$ (i.e., first column of $\mathbf{R}^\top$).[4] Let us call $i \neq j$ such pair of coordinates, i.e., $r_{1j} \neq 0$ and $r_{1i} \neq 0$.

4: In short, if this were not the case, this column would have a single non-vanishing coefficient, which would need to be one due to the unit norm of the rows of this orthogonal matrix. Such structure of the matrix $\mathbf{R}$ would entail that the associated canonical basis vector $\mathbf{e}_1$ is transformed by $\mathbf{R}^{-1} = \mathbf{R}^\top$ into a canonical basis vector $\mathbf{e}_j$ which contradicts the assumptions.

Now consider the off-diagonal term (i, j) of (B.14), which we assumed (for a contradiction) must be zero almost surely w.r.t. $p_{\mathbf{z}}$. Since the term in (B.15) is diagonal, this off-diagonal term is given by:

$$\sum_{k=1}^{n} \left(\frac{d\sigma_k^{-1}}{dz_k}(z_k) \right)^2 \left\| \frac{\partial \mathbf{f}}{ds_k} \circ \boldsymbol{\sigma}^{-1}(\mathbf{z}) \right\|^2 r_{ki} r_{kj} = \sum_{k=1}^{n} \left(\frac{d\sigma_k^{-1}}{dz_k}(z_k) \right)^2 \lambda(\boldsymbol{\sigma}^{-1}(\mathbf{z}))^2 r_{ki} r_{kj} = 0 \,.$$

where for the first equality we have used the fact that $\mathbf{f}$ is a conformal map with conformal factor $\lambda(\mathbf{s})$ (by assumption), and where the second equality must hold almost surely w.r.t. $p_{\mathbf{z}}$.

Since $\mathbf{f}$ is invertible, it has non vanishing Jacobian determinant. Hence, the conformal factor λ must be a strictly positive function, so

$$\lambda(\boldsymbol{\sigma}^{-1}(\mathbf{z}))^2 > 0, \; \forall \mathbf{z}.$$

Thus, for almost all $\mathbf{z}$, we must have:

$$\sum_{k=1}^{n} \left(\frac{d\sigma_k^{-1}}{dz_k}(z_k) \right)^2 r_{ki} r_{kj} = 0 \,. \tag{B.16}$$

Now consider the first term $\left(\frac{d\sigma_1^{-1}}{dz_1}(z_1) \right)^2 r_{1i} r_{1j}$ in the sum.

Recall that $r_{1i} r_{1j} \neq 0$, and that $\frac{d\sigma_1^{-1}}{dz_1}(z_1)$ is strictly monotonic on a neighborhood of z_1^0.

As a consequence, $\left(\frac{d\sigma_1^{-1}}{dz_1}(z_1) \right)^2 r_{1i} r_{1j}$ is also strictly monotonic with respect to z_1 on a neighborhood of z_1^0 (where the other variables $(z_2, \ldots, z_n)$ are left constant), while the other terms in the sum in (B.16) are left constant because σ is an element-wise transformation.

This leads to a contradiction as (B.16) (which should be satisfied for all $\mathbf{z}$) cannot stay constantly zero as z_1 varies within the neighbourhood of z_1^0.

Hence our assumption that $C_{\text{IMA}}(\mathbf{f} \circ \mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}}), p_{\mathbf{s}}) = 0$ cannot hold.

We conclude that $C_{\text{IMA}}(\mathbf{f} \circ \mathbf{a}^{\mathbf{R}}(p_{\mathbf{s}}), p_{\mathbf{s}}) > 0$. □

B.3 Worked out example

Example B.3.1 (Polar to Cartesian coordinates) Consider the following example of a nonlinear ICA model which represents a change of basis from polar to Cartesian coordinates:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathbf{f}(\mathbf{s}) = \begin{pmatrix} f_1(\mathbf{s}) \\ f_2(\mathbf{s}) \end{pmatrix} = \begin{pmatrix} r\cos(\theta) \\ r\sin(\theta) \end{pmatrix}$$

with sources

$$= \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} r \\ \theta \end{pmatrix}, \qquad r \sim U[0, R], \qquad \theta \sim U[0, 2\pi],$$

First, we consider the Jacobian of the true mixing $\mathbf{f}$ which is given by:

$$\mathbf{J}_\mathbf{f}(\mathbf{s}) = \mathbf{J}_\mathbf{f}(r,\theta) = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix},$$

and its determinant and column norms are given by

$$\begin{aligned} |\det \mathbf{J}_\mathbf{f}(\mathbf{s})| &= r\left(\cos^2(\theta) + \sin^2(\theta)\right) = r \\ \left\| \frac{\partial \mathbf{f}}{\partial s_1}(\mathbf{s}) \right\| &= \left\| \frac{\partial \mathbf{f}}{\partial r}(r,\theta) \right\| = \cos^2(\theta) + \sin^2(\theta) = 1 \\ \left\| \frac{\partial \mathbf{f}}{\partial s_2}(\mathbf{s}) \right\| &= \left\| \frac{\partial \mathbf{f}}{\partial \theta}(r,\theta) \right\| = r\left(\cos^2(\theta) + \sin^2(\theta)\right) = r \end{aligned}$$

In other words, the columns of $\mathbf{J}_\mathbf{f}(\mathbf{s})$ are orthogonal for all $\mathbf{s}$, so that $C_{\text{IMA}} = 0$ for the true solution.

Next, we apply the Darmois construction.

First, we write the joint density of (x_1, x_2) using the change of variable formula:

$$p(x_1,x_2) = |\det \mathbf{J}_\mathbf{f}(r,\theta)|^{-1} p(r,\theta) = r^{-1}\frac{1}{2\pi R} = \frac{1}{\sqrt{x_1^2+x_2^2}}\frac{1}{2\pi R}.$$

Next, we compute the marginal density $p(x_1)$. Note that the observations $\mathbf{x}$ live on the disk of radius R, $\|\mathbf{x}\| \leq R$, so $p(x_1, x_2) = 0$ whenever $x_1^2 + x_2^2 > R^2$.

$$p(x_1) = \int_{-\sqrt{R^2-x_1^2}}^{\sqrt{R^2-x_1^2}} p(x_1,x_2)dx_2 = \frac{1}{2\pi R}\int_{-\sqrt{R^2-x_1^2}}^{\sqrt{R^2-x_1^2}} \frac{dx_2}{\sqrt{x_1^2+x_2^2}} = \frac{1}{2\pi R}\int_{-\sqrt{R^2-x_1^2}}^{\sqrt{R^2-x_1^2}} \frac{dx_2}{x_1\sqrt{1+(\frac{x_2}{x_1})^2}}$$

Applying the change of variable $t = \frac{x_2}{x_1}$ with $dt = \frac{dx_2}{x_1}$, and using the integral $\int (1+t^2)^{-\frac{1}{2}} dt = \operatorname{arcsinh}(t) + C$, as well as the fact that arcsinh is an odd function, we obtain

$$p(x_1) = \frac{1}{2\pi R}\int_{-\sqrt{\left(\frac{R}{x_1}\right)^2-1}}^{\sqrt{\left(\frac{R}{x_1}\right)^2-1}} \frac{dt}{\sqrt{1+t^2}} = \frac{1}{\pi R}\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)$$

Next, we compute the conditional density $p(x_2|x_1)$:

$$p(x_2|x_1) = \frac{p(x_1,x_2)}{p(x_1)} = \frac{(2\pi R)^{-1}\left(x_1^2+x_2^2\right)^{-1}}{(\pi R)^{-1}\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)} = \left(2\sqrt{x_1^2+x_2^2}\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)\right)^{-1}$$

Finally, we compute the off-diagonal term in the general form of the

inverse Jacobian for Damois-style solutions in (B.12):

$$
\begin{aligned}
c_{21}(\mathbf{x}) &= \frac{\partial}{\partial x_1}\int_{-\infty}^{x_2} p(x_2|x_1)dx_2 = \frac{\partial}{\partial x_1}\int_{-\sqrt{R^2-x_1^2}}^{x_2} \frac{dx_2}{2\sqrt{x_1^2+x_2^2}\,\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)} \\
&= \frac{1}{2}\frac{\partial}{\partial x_1}\left(\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)^{-1}\int_{-\sqrt{R^2-x_1^2}}^{x_2}\frac{dx_2}{\sqrt{x_1^2+x_2^2}}\right) \\
&= \frac{1}{2}\frac{\partial}{\partial x_1}\left(\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)^{-1}\left(\operatorname{arcsinh}(x_2)-\operatorname{arcsinh}\left(-\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)\right)\right) \\
&= \frac{1}{2}\frac{\partial}{\partial x_1}\left(\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)^{-1}\left(\operatorname{arcsinh}(x_2)+\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)\right)\right) \\
&= \frac{1}{2}\frac{\partial}{\partial x_1}\left(1+\frac{\operatorname{arcsinh}(x_2)}{\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)}\right) \\
&= \frac{1}{2}\operatorname{arcsinh}(x_2)\frac{\partial}{\partial x_1}\left(\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)^{-1}\right) \\
&= -\frac{1}{2}\operatorname{arcsinh}(x_2)\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)^{-2}\frac{\partial}{\partial x_1}\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)
\end{aligned}
$$

Using the derivative $\frac{\partial}{\partial t}\operatorname{arcsinh}(t) = (t^2+1)^{-\frac{1}{2}}$ and repeatedly applying the chain rule, we obtain:

$$
\begin{aligned}
c_{21}(\mathbf{x}) &= -\frac{1}{2}\operatorname{arcsinh}(x_2)\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)^{-2}\frac{x_1}{R}\frac{\partial}{\partial x_1}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right) \\
&= -\frac{1}{2}\operatorname{arcsinh}(x_2)\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)^{-2}\frac{x_1}{R}\frac{1}{2}\frac{1}{\sqrt{\left(\frac{R}{x_1}\right)^2-1}}(-2)R^2x_1^{-3} \\
&= \frac{R}{2x_1\sqrt{R^2-x_1^2}}\operatorname{arcsinh}(x_2)\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2-1}\right)^{-2}
\end{aligned}
$$

Again, recall that this only holds inside the disk of radius R, otherwise $c_{12} = 0$ (as the CDF will be zero or one, irrespective of x_1).

The C_{IMA} for the Darmois solution thus takes the form:

$$
\begin{aligned}
C_{\text{IMA}}^{\text{Darmois}} &= \int \frac{1}{2}\log\left(p(x_1)^{-2} + c_{21}(\mathbf{x})^2 p(x_1,x_2)^{-2}\right) + \log\left(p(x_2|x_1)^{-1}\right) - \log\left(p(x_1,x_2)^{-1}\right) d\mathbf{s} \\
&= \int \frac{1}{2}\log\Bigg[\left(\frac{1}{\pi R}\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2 - 1}\right)\right)^{-2} \\
&\quad + \left(\frac{R}{2x_1\sqrt{R^2 - x_1^2}}\operatorname{arcsinh}(x_2)\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2 - 1}\right)^{-2}\right)^2 \left(\frac{1}{\sqrt{x_1^2 + x_2^2}}\frac{1}{2\pi R}\right)^{-2}\Bigg] \\
&\quad + \log\left(2\sqrt{x_1^2 + x_2^2}\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2 - 1}\right)\right) \\
&\quad - \log(2\pi R) - \frac{1}{2}\log\left(x_1^2 + x_2^2\right) d\mathbf{s} \\
&= \int \frac{1}{2}\log\Bigg[\pi^2 R^2 \operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2 - 1}\right)^{-2} \\
&\quad + \frac{R^2}{4x_1^2(R^2 - x_1^2)}\operatorname{arcsinh}(x_2)^2 \operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2 - 1}\right)^{-4} (x_1^2 + x_2^2) 4\pi^2 R^2\Bigg] \\
&\quad + \log(2) + \frac{1}{2}\log(x_1^2 + x_2^2) + \log\left(\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2 - 1}\right)\right) \\
&\quad - \log(2) - \log(\pi R) - \frac{1}{2}\log\left(x_1^2 + x_2^2\right) d\mathbf{s} \\
&= \int \frac{1}{2}\log\Bigg[\pi^2 R^2 \operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2 - 1}\right)^{-2} \\
&\quad + \frac{\pi^2 R^4 (x_1^2 + x_2^2)}{x_1^2(R^2 - x_1^2)}\operatorname{arcsinh}(x_2)^2 \operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2 - 1}\right)^{-4}\Bigg] \\
&\quad + \log\left(\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2 - 1}\right)\right) - \log(\pi R)\, d\mathbf{s} \\
&= \int \frac{1}{2}\log\left(1 + \frac{R^2(x_1^2 + x_2^2)\operatorname{arcsinh}(x_2)^2}{x_1^2(R^2 - x_1^2)\operatorname{arcsinh}\left(\sqrt{\left(\frac{R}{x_1}\right)^2 - 1}\right)^2}\right) d\mathbf{s} > 0
\end{aligned}
$$

where the strict inequality in the last step follows from the fact that the fraction inside the logarithm, and hence the entire integrand, is strictly positive within the disk of integration.

We have thus shown that for the example of an orthogonal coordinate transformation from polar to Cartesian coordinates, which is not a conformal map, the C_{IMA} os the true solution is zero and that of the Darmois construction is strictly greater than zero, hence the two can be distinguished based on the value of the C_{IMA} contrast.

B.4 Experiments

For our experiments we used Jax [177], Distrax [178] and Haiku [494] to implement our models; the Jacobian and C_{IMA} computation and optimisation are performed with the automatic differentiation tools provided in Jax.

B.4.1 Sampling random Möbius transformations.

In order to generate mixing functions with $C_{\text{IMA}} = 0$, we use Möbius transformations (see Appendix B.5 and in particular Thm. B.5.2, for additional details on this kind of functions) with randomly sampled parameters, as specified below. A Möbius transformation $\mathbf{f}^{\text{M}} : \mathbb{R}^n \to \mathbb{R}^n$ is given by

$$\mathbf{f}^{\text{M}}(\mathbf{s}) = \mathbf{t} + \frac{r\mathbf{A}(\mathbf{s} - \mathbf{b})}{\|\mathbf{s} - \mathbf{b}\|^{\epsilon}}, \tag{B.17}$$

with parameters $\mathbf{b}, \mathbf{t} \in \mathbb{R}^n$, $r \in \mathbb{R}$, $\mathbf{A}$ is an orthogonal matrix and $\epsilon \in \{0, 2\}$. The flow models we train have an diagonal affine layer at the top with fixed shift and scale set to the mean and standard deviation of the training data, thereby normalizing the inputs. Hence, without loss of generality, we can set the $\mathbf{t}$ parameter to zero and r to one. Since $\epsilon = 0$ corresponds to a linear transformation, we generally set $\epsilon = 2$ in our experiments unless otherwise specified. We sample the orthogonal matrix through the `ortho_group` function in `scipy.stats` [495]. To avoid singularities given by a vanishing denominator in the second term on the RHS of (B.17), which would yield observed distributions with strong outliers and therefore hard to fit for our models, we restrict $\mathbf{b}$ to lie outside the unit square $\mathbf{s}$ is sampled from. We achieve this by sampling $\mathbf{b}$ from a normal distribution and reject the sample until it is located outside of the unit square.

B.4.2 How to implement the Darmois construction

In the following, we describe how the Darmois construction can be implemented based on normalising flow models [71]. The key idea is that the components g_i^{D} of the Darmois construction (2.15) are conditional (cumulative) density functions corresponding to a given factorisation $p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|\mathbf{x}_{1:i-1})$ of the likelihood. A flow model with triangular Jacobian can be used to maximise the likelihood of the observations under a change of variable respecting said factorisation, and learning to map the observed variables onto a given (factorised) base distribution. After training, and provided that the model is expressive enough, the CDF of each component of the reconstructed sources should match that of the base distribution. By further transforming each reconstructed variable through said CDF, we achieve a global mapping of the observations onto a Uniform distribution on the n-dimensional hypercube, with a triangular Jacobian, matching the transformation operated by the Darmois construction (see also see [71], section 2.2). Note that, for the purpose of computing the C_{IMA} of the Darmois construction, this final step can be omitted due to Proposition 3.4.2, *(ii)*, stating that the contrast is blind to element-wise reparametrisations of the sources.

We remark that, while the possibility of using normalising flows to "learn" the Darmois construction is mentioned in [71, 179], where a similar construction is mentioned in a theoretical argument to prove "universal approximation capacity for densities" for normalising flow models with triangular Jacobian, it has to the best of our knowledge not been tested empirically, since autoregressive modules with triangular Jacobian are typically used in combination with permutation, shuffling or linear layers which overall lead to architectures with a non-triangular Jacobian.

Expressive normalising flow with triangular Jacobian. To obtain an expressive normalizing flow with triagular Jacobian, we modify the residual flow model [180].[5] A residual flow is a residual network which is made invertible through spectral normalization. Each layer is given by

$$\mathbf{z}' = \mathbf{z} + \mathbf{g}(\mathbf{z}), \tag{B.18}$$

where $\mathbf{z}', \mathbf{z} \in \mathbb{R}^n$ and $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^n$ is a small neural network. Due to the chain rule, for the Jacobian of the overall flow model to be triangular, a sufficient condition is that all the layers have triangular Jacobian. Since the Jacobian of $\mathbf{f}(\mathbf{z}) = \mathbf{z}$ is the identity matrix, we can restrict our attention to the neural network $\mathbf{g}$. In our experiments, this is going to be a fully connected network. If it has l layers and $h \geq n$ hidden units, it is given by

$$\mathbf{g}(\mathbf{z}) = \mathbf{b}_1 + \mathbf{W}_1\phi(\mathbf{b}_2 + \mathbf{W}_2\phi(\mathbf{b}_3 + \mathbf{W}_3 \cdots \phi(\mathbf{b}_l + \mathbf{W}_l\mathbf{z}) \cdots)), \tag{B.19}$$

where $\phi : \mathbb{R}^n \to \mathbb{R}^n$ is an element-wise nonlinearity, $\mathbf{b}_1 \in \mathbb{R}^n$, $\mathbf{b}_2, ..., \mathbf{b}_l \in \mathbb{R}^h$ are the biases, and $\mathbf{W}_1 \in \mathbb{R}^{n\times h}$, $\mathbf{W}_2, ..., \mathbf{W}_{l-1} \in \mathbb{R}^{h\times h}$, $\mathbf{W}_l \in \mathbb{R}^{h\times n}$ are the weight matrices. In order for the Jacobian of $\mathbf{g}$ to be triangular, $g_n(\mathbf{z})$ should only depend on z_n, $g_{n-1}(\mathbf{z})$ should only depend on z_n and z_{n-1}, and so on. To achieve this, we make the weight matrices block triangular as indicated in (B.20), (B.21), and (B.22).

5: We describe how to implement a function with upper triangular Jacobian, but the reasoning can be extended to implement functions whose Jacobian is lower triangular.

$$\mathbf{W}_1 = \begin{pmatrix} * & * & & * \\ \vdots & \vdots & & \vdots \\ * & * & & * \\ 0 & * & & * \\ \vdots & \vdots & & \vdots \\ 0 & * & & * \\ & & \ddots & \\ 0 & 0 & & * \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & * \end{pmatrix} \begin{matrix} \\ \\ \} h_1 \\ \\ \\ \} h_2 \\ \\ \\ \} h_n \end{matrix} \tag{B.20}$$

$$\mathbf{W}_l = \begin{pmatrix} * & \cdots & * & * & \cdots & * & & * & \cdots & * \\ 0 & \cdots & 0 & * & \cdots & * & & * & \cdots & * \\ & & & & & & \ddots & & & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & & * & \cdots & * \end{pmatrix} \tag{B.21}$$

$$\underbrace{}_{h_1} \quad \underbrace{}_{h_2} \quad \underbrace{}_{h_n}$$

$$\mathbf{W}_i = \begin{pmatrix} * & \cdots & * & * & \cdots & * & & * & \cdots & * \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ * & \cdots & * & * & \cdots & * & & * & \cdots & * \\ 0 & \cdots & 0 & * & \cdots & * & & * & \cdots & * \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & * & \cdots & * & & * & \cdots & * \\ & & & & & & \ddots & & & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & & * & \cdots & * \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & & * & \cdots & * \end{pmatrix} \begin{matrix} \} h_1 \\ \\ \\ \} h_2 \\ \\ \\ \\ \} h_n \end{matrix} \quad \text{for } i \in \{2, ..., l-1\}$$
$$\underbrace{\qquad}_{h_1} \quad \underbrace{\qquad}_{h_2} \quad \underbrace{\qquad}_{h_n} \tag{B.22}$$

Here, h_i is the number of hidden units dedicated to transforming $\mathbf{z}_i$ with the constraint $\sum_{i=1}^{n} h_i = h$. We perform an even split such that the h_i and h_j differ by at most 1 for $i, j \in \{1, ..., n\}$. The weight matrices are restricted to be block triangular during optimization by setting the respective matrix elements to zero after each iteration of the optimizer. The model can simply be made and kept invertible using the same spectral normalization as is used for dense residual flows [180]. We train our model to map onto a standard Normal base distribution.

B.4.3 Generating random MLP mixing functions

In order to generate random MLP mixing functions, we adopt the same initalisation as in [86]: we initialise the square weight matrices to be orthogonal,[6] and use the `leaky_tanh` invertible nonlinearity.

6: Note that orthogonality of the weight matrices in a MLP does not guarantee satisfying Principle 3.4.1, due to the elementwise nonlinearities between the layers, which overall lead to a Jacobian whose columns are in general not orthogonal.

B.4.4 Maximum likelihood with low C_{IMA}

The modified maximum likelihood objective described in Subsection 3.5.2 can be written as follows:[7]

7: while the objective in Subsection 3.5.2 involves an expectation over $p_{\mathbf{x}}$, we consider the loss for a single point $\mathbf{x}$ here, $\mathscr{L}(\mathbf{g}; \mathbf{x})$.

$$\begin{aligned} \mathscr{L}(\mathbf{g}; \mathbf{x}) &= \log p(\mathbf{x}) - \lambda \cdot c_{\text{IMA}}(\mathbf{g}^{-1}, p_{\mathbf{y}}) \\ &= \sum_{i=1}^{n} \log p_{y_i}(\mathbf{g}^i(\mathbf{x})) + \log |\mathbf{J}_{\mathbf{g}}(\mathbf{x})| - \lambda \cdot \left(\sum_{i=1}^{n} \log \left\| [\mathbf{J}_{\mathbf{g}^{-1}}(\mathbf{g}(\mathbf{x}))]_i \right\| - \log \left| \mathbf{J}_{\mathbf{g}^{-1}}(\mathbf{g}(\mathbf{x})) \right| \right) \\ &= \sum_{i=1}^{n} \log p_{y_i}(\mathbf{g}^i(\mathbf{x})) + \log |\mathbf{J}_{\mathbf{g}}(\mathbf{x})| - \lambda \cdot \left(\sum_{i=1}^{n} \log \left\| [\mathbf{J}_{\mathbf{g}}^{-1}(\mathbf{x})]_i \right\| + \log \left| \mathbf{J}_{\mathbf{g}}(\mathbf{x}) \right| \right) \\ &= \sum_{i=1}^{n} \log p_{y_i}(\mathbf{g}^i(\mathbf{x})) + (1-\lambda) \log |\mathbf{J}_{\mathbf{g}}(\mathbf{x})| - \lambda \sum_{i} \log \left\| [\mathbf{J}_{\mathbf{g}}^{-1}(\mathbf{x})]_i \right\| , \end{aligned} \tag{B.23}$$

where $[\mathbf{J}_{\mathbf{g}}^{-1}(\mathbf{x})]_i$ represents the i-th column of the inverse of the Jacobian of $\mathbf{g}$ computed at $\mathbf{x}$.

We use the same model as the one described in Appendix B.4.2, but without the constraint that the Jacobian should be triangular, and train with a Logistic base distribution.

Note that the computational efficiency of optimising objective (B.23) is cubic in the input size n, due to a number of operations (matrix inversion, Jacobian and determinant computation via automatic differentiation, etc.) which are $\mathcal{O}(n^3)$. However, similarly to what already observed in [78], we found that for data of moderate dimensionality computing and optimising objective (B.23) with automatic differentiation is feasible. For example, training a residual flow with 64 layers for 10^5 iterations takes roughly 5.3 hours for $n = 2$, 5.7 hours for $n = 5$, and 6.3 hours for $n = 7$ on the same hardware (see section B.4.5). An interesting direction for future work would be to find computationally efficient ways of optimising (B.23).

When computing the C_{IMA} of the Darmois solutions of randomly generated functions, we restricted ourselves to Möbius transformations, i.e. conformal maps. However, there are also nonconformal maps satisfying $C_{\text{IMA}} = 0$, e.g. the transformation of Cartesian to Polar coordinates, see Appendix B.3. To test whether the C_{IMA} of the Darmois solutions is actually bigger than 0, we gener

B.4.5 Evaluation

Mean correlation coefficient. To evaluate the performance of our method, we compute the mean correlation coefficient (MCC) between the original sources and the corresponding latents, see for example [62]. We first compute the matrix of correlation coefficients between all pairs of ground truth and reconstructed sources. Then, we solve a linear sum assignment problem (e.g. using the Hungarian algorithm) to match each reconstructed source to the ground truth one which has the highest correlation with it. The MCC matrix contains the Spearman rank-order correlations between the ground truth and reconstructed sources, a measure which is blind to nonlinear invertible reparametrisations of the sources.

Nonlinear Amari metric. While the MCC metric evaluates BSS by comparing ground truth and reconstructed sources, we propose an additional evaluation directly based on comparing the (Jacobians of the) true mixing and the learnt unmixing. We take inspiration from an evaluation metric used in the context of linear ICA, the Amari distance [181]: Given a learnt unmixing $\mathbf{W}$ and the true mixing $\mathbf{A}$, and defining the matrix $\mathbf{R} = \mathbf{AW}$, the Amari distance is defined as

$$d^{\text{Amari}}(\mathbf{R}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \frac{[\mathbf{R}]_{ij}^2}{\max_l [\mathbf{R}]_{il}^2} - 1 \right) + \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \frac{[\mathbf{R}]_{ji}^2}{\max_l [\mathbf{R}]_{lj}^2} - 1 \right), \quad \text{(B.24)}$$

and is greater than or equal to zero, canceling if and only if $\mathbf{R}$ is a scale and permutation matrix, that is when the learnt unmixing is matching the unresolvable ambiguities of linear ICA.

We extend this idea to the nonlinear setting: Given a true mixing $\mathbf{f}$ and a learnt unmixing $\mathbf{g}$, we define our nonlinear Amari distance as

$$d^{\text{n-Amari}}(\mathbf{g}, \mathbf{f}) = \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}} \left[d^{\text{Amari}} \left(\mathbf{J}_{\mathbf{g}}(\mathbf{x}) \mathbf{J}_{\mathbf{f}}(\mathbf{f}^{-1}(\mathbf{x})) \right) \right] . \quad \text{(B.25)}$$

Then, according to the definition of Amari distance (B.24), if the smooth function $\mathbf{g} \circ \mathbf{f}$ is a permutation composed with a scalar function, thus precisely matching the BSS equivalence class defined in Defn. 2.4.1, this would result in its Jacobian (that is, the product of the Jacobians $\mathbf{J_g}(\mathbf{x})\mathbf{J_f}(\mathbf{f}^{-1}(\mathbf{x}))$) equalling the product of a diagonal matrix and a permutation matrix at every point $\mathbf{x}$: the quantity $d^{\text{n-Amari}}(\mathbf{g}, \mathbf{f})$ would therefore be equal to zero.

This metric can be of independent interest and potentially useful in contexts where the reconstructed sources might be a noisy version of the true ones, but the true unmixing is nevertheless identifiable. Our implementation is based on the one for the (linear) Amari distance provided in the code for [48].

C_{IMA} of Darmois solutions for nonconformal maps satisfying the IMA principle. When computing the C_{IMA} of the Darmois solutions of randomly generated functions, we restricted ourselves to Möbius transformations which are conformal maps. However, there are also nonconformal maps satisfying $C_{\text{IMA}} = 0$, e.g., the transformation from polar to Cartesian coordinates with $n = 2$, see Appendix B.3. To test whether the C_{IMA} of the Darmois solutions is actually bigger than 0, we generate random radial transformations by imposing a random scale and shift before applying the radial transformation, compute the Darmois solution as we have done in Subsection 3.5.1, and calculate its C_{IMA} on the test set. We did 50 runs and the results are shown in Fig. B.7.

Similar to Fig. 3.5 **(a)** we can clearly see that all C_{IMA} values of the final models are larger than 0, with the smallest value being 0.01. This confirms the result we have already shown theoretically.

Additional plots for Subsection 3.5.2. We show additional plots for the quantitative experiments involving training with the objective described in (B.23), see Fig. B.3, Fig. B.4 and Fig. B.5.

For $\epsilon = 0$ (that is, ground truth mixing linear), there appears to be an almost perfect recovery of the ground truth sources (resp. unmixing function) for $\lambda \in \{0.5, 1.0\}$, as can be seen by the high (resp. low) values of the MCC (resp. nonlinear Amari distance) evaluations ; this is in stark contrast with the distribution of the MCC (resp. nonlinear Amari distance) values for models trained with $\lambda = 0$, which are typically much higher (resp. lower), indicating that the learnt solutions do not achieve blind source separation (see $n = 2$, Fig. B.3 **(g), (h)**; $n = 5$, Fig. B.4 **(g), (h)**). All models achieve a comparably good fit, reflected in the KL-divergence values ($n = 2$, Fig. B.3 **(e)**; $n = 5$, Fig. B.4 **(e)**).

The trend is confirmed when the true mixing is nonlinear ($\epsilon = 2$), with slightly lower (resp. higher) values achieved with C_{IMA} regularisation for the MCC (resp. nonlinear Amari) metrics; this possibly due to the increased difficulty of fitting observations generated by a nonlinear mixing, as can be seen from the higher values of the KL-divergence ($n = 2$, Fig. B.3 **(a)**; $n = 5$, Fig. B.4 **(a)**; $n = 7$, Fig. B.5 **(a)**);[8] still, the beneficial effect of $\lambda \in \{0.5, 1.0\}$ with respect to models trained with $\lambda = 0$ is clear, and is apparently stronger for $\lambda = 1.0$ and with higher

8: The distribution of the KL values contains outliers, and seemingly more strongly for lower values of λ.

data dimensionality n ($n = 2$, Fig. B.3 **(c), (d)**; $n = 5$, Fig. B.4 **(c), (d)**; $n = 7$, Fig. B.5 **(c), (d)**).

We additionally plot the C_{IMA} values for the all trained models, for all values of λ. It can be seen that solutions found by unregularised maximum likelihood estimation typically learn functions with relatively high values of C_{IMA}, while as expected the regularised version achieves low values ($n = 2$, Fig. B.3 **(b), (f)**; $n = 5$, Fig. B.4 **(b), (f)**; $n = 7$, Fig. B.5 **(b)**).

Finally, in figure B.6, we report the same plot as in 3.5, top row, but with a perceptually uniform colormap.

Comparison to FastICA. We compared the performance of our proposed regularised maximum likelihood procedure to a state of the art method for linear ICA, FastICA [182], in the implementation from the Scikit-learn package [312], over 50 repetitions. Our experiments show that our regularised method ($\lambda = 0.5$, and particularly $\lambda = 1.0$; $\lambda = 0.0$ provides the unregularised nonlinear baseline) is superior in learning the true unmixing and reconstructing the sources. This indicates that the linearity assumption of FastICA does not allow enough flexibility to solve blind source separation in our setting, whereas our criterion does (see Fig. B.8, Fig. B.9 and Fig. B.10).[9] While the spread in the distributions of MCC and Amari distance can be largely attributed to the brittleness of neural networks, the median values for the MCC (resp. nonlinear Amari distance) are consistently higher (resp. lower) for our regularised method than for FastICA. In contrast, the performance of FastICA is consistently better than the unregularised baseline.

9: the experimental setting and the plots for the normalising flow models correspond to those already shown in the paper, but here we modified the y-axis scale to facilitate the comparison of all methods

Details on resources used. All models were trained on compute instances with 16 Intel Xeon E5-2698 CPUs and a Nvidia Geforce GTX980 GPU. The cluster we used has 204 thereof. Training the models took between 4 and 16 hours depending mainly on the dimensionality n and number of samples in the dataset, and on the number of iterations used for training. Overall, we trained around 2000 models, amounting to roughly 18000 GPU hours.

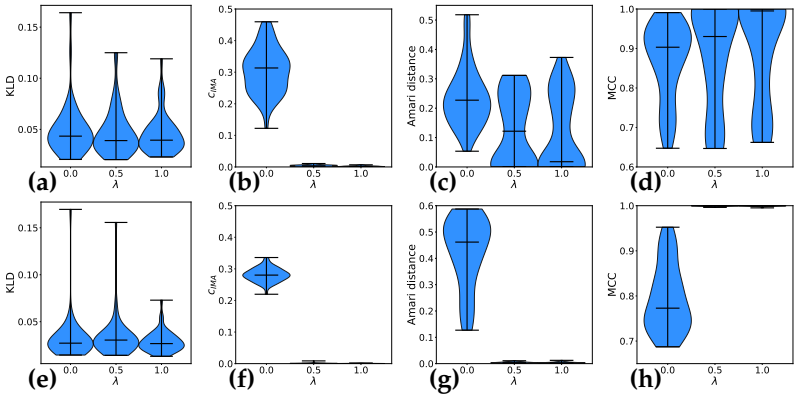


Figure B.3: BSS via C_{IMA}-regularised MLE for $n = 2$ dimensions with $\lambda \in \{0.0, 0.5, 1.0\}$. The true mixing function is a randomly generated Möbius transformation, nonlinear (with $\epsilon = 2$) in **(a)–(d)** and linear (with $\epsilon = 0$) transformation for **(e)–(h)**. For each type of transformation and λ, seeded runs are done. **(a)**, **(e)** KL-divergence between ground truth likelihood and learnt model; **(b)**, **(f)** C_{IMA} of the learnt models; **(c)**, **(g)** nonlinear Amari distance given true mixing and learnt unmixing; **(d)**, **(h)** MCC between true and reconstructed sources.

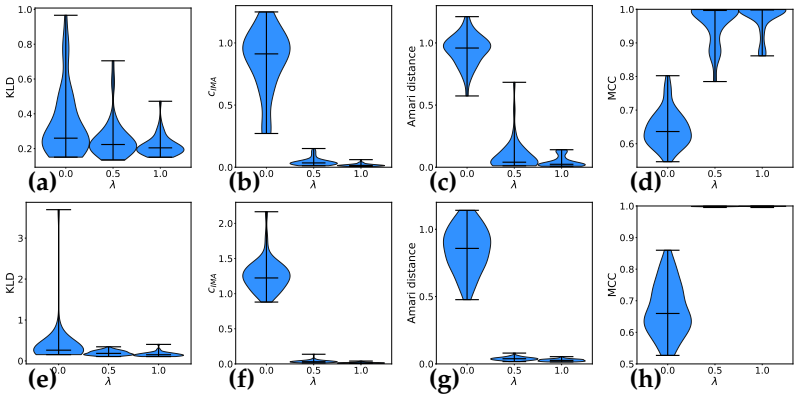


Figure B.4: BSS via C_{IMA}-regularised MLE for $n = 5$ dimensions with $\lambda \in \{0.0, 0.5, 1.0\}$. The true mixing function is a randomly generated Möbius transformation, nonlinear (with $\epsilon = 2$) in **(a)–(d)** and linear (with $\epsilon = 0$) transformation for **(e)–(h)**. For each type of transformation and λ, seeded runs are done. **(a)**, **(e)** KL-divergence between ground truth likelihood and learnt model; **(b)**, **(f)** C_{IMA} of the learnt models; **(c)**, **(g)** nonlinear Amari distance given true mixing and learnt unmixing; **(d)**, **(h)** MCC between true and reconstructed sources.

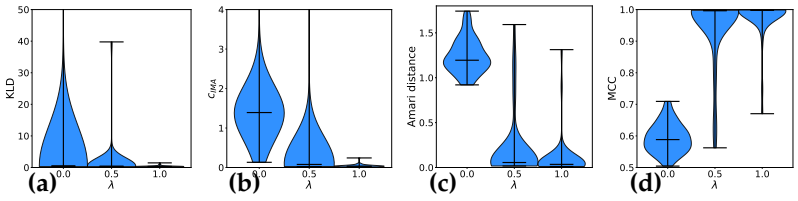


Figure B.5: BSS via C_{IMA}-regularised MLE for $n = 7$ dimensions with $\lambda \in \{0.0, 0.5, 1.0\}$. The true mixing function is a randomly generated Möbius transformation (with $\epsilon = 2$). For each λ, seeded runs are done. **(a)** KL-divergence between ground truth likelihood and learnt model; **(b)** C_{IMA} of the learnt models; **(c)** nonlinear Amari distance given true mixing and learnt unmixing; **(d)** MCC between true and reconstructed sources.

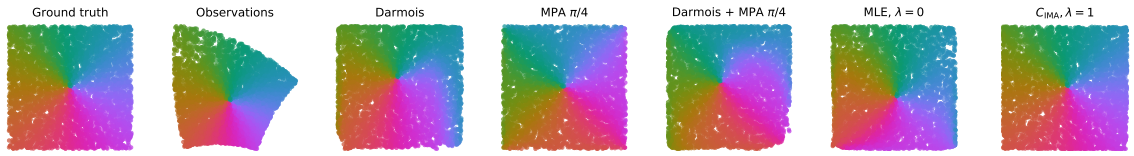


Figure B.6: Visual comparison of different nonlinear ICA solutions for $n = 2$: *(left to right)* true sources; observed mixtures; Darmois solution; true unmixing, composed with the measure preserving automorphism (MPA) from (2.16) (with rotation by $\pi/4$); Darmois solution composed with the same MPA; maximum likelihood ($\lambda = 0$); and C_{IMA}-regularised approach ($\lambda = 1$).

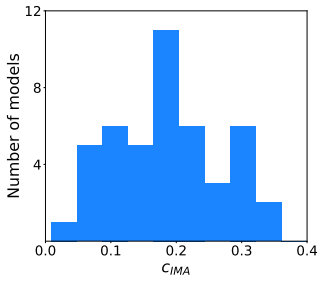
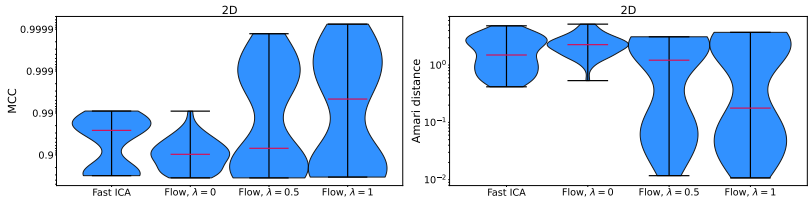
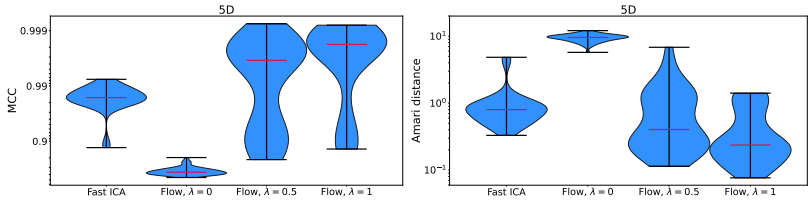


Figure B.7: Histogram of the C_{IMA} values of the Darmois solutions of 50 randomly generated radial transformations.



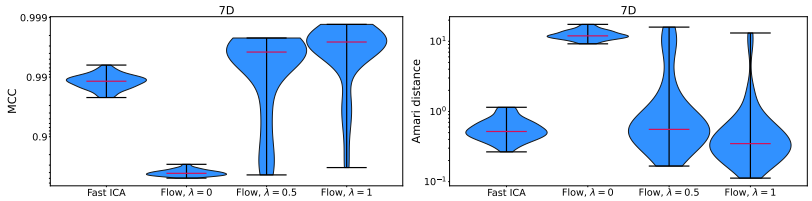
(a) (b)

Figure B.8: Comparison between FastICA and our normalising flow method with $\lambda \in \{0.0, 0.5, 1.0\}$, $n = 2$. (a) MCC; (b) Amari distance.



(a) (b)

Figure B.9: Comparison between FastICA and our normalising flow method with $\lambda \in \{0.0, 0.5, 1.0\}$, $n = 5$. (a) MCC; (b) Amari distance.



(a) (b)

Figure B.10: Comparison between FastICA and our normalising flow method with $\lambda \in \{0.0, 0.5, 1.0\}$, $n = 7$. (a) MCC; (b) Amari distance.

B.5 Additional background on conformal maps and Möbius transformations

Similarities. A *similarity* of a Euclidean space is a bijection $\mathbf{f}$ from the space onto itself that multiplies all distances by the same positive real number r, so that for any two points $\mathbf{x}$ and $\mathbf{y}$ we have

$$d(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{y})) = rd(\mathbf{x}, \mathbf{y}),$$

where $d(\mathbf{x}, \mathbf{y})$ is the Euclidean distance from $\mathbf{x}$ to $\mathbf{y}$ [496]. The scalar r is sometimes termed the ratio of similarity, the stretching factor and the similarity coefficient. When $r = 1$ a similarity is called an isometry (rigid transformation). Two sets are called similar if one is the image of the other under a similarity.

As a map $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, a similarity of ratio r takes the form

$$\mathbf{f}(\mathbf{x}) = r\mathbf{A}\mathbf{x} + \mathbf{t},$$

where $\mathbf{A}$ is a orthogonal matrixn $n \times n$ and $\mathbf{t} \in \mathbb{R}^n$ is a translation vector.

Note that such a similarity $\mathbf{f}$ has Jacobian $\mathbf{J}_\mathbf{f}(\mathbf{x}) = r\mathbf{A}$ for any $\mathbf{x}$.

Conformal maps. Conformal maps are angle preserving transformation, and in this sense, are a generalization of similarities. In short, let U be an open subset of $\mathbb{R}^n$, $\varphi : U \to \mathbb{R}^n$ is a conformal map if, for two arbitrary curves $\gamma_1(t)$ and $\gamma_2(t)$ on $\mathbb{R}^n$, where these curves intersect each other with angle θ in point $\mathbf{p} \in U$, then $\varphi \circ \gamma_1(t)$ and $\varphi \circ \gamma_2(t)$ intersect each other with the same angle θ in the point $\varphi(\mathbf{p})$.

A characterisation of conformal maps directly related to orthogonal coordinate systems is the following.

Proposition B.5.1 (See e.g. [497]) *Let U be an open subset of $\mathbb{R}^n$ with a C^1-function $\varphi : U \to \mathbb{R}^n$. Then φ is conformal iff there exists a scalar function $\lambda : U \to \mathbb{R}$ such that $\lambda(\mathbf{x})^{-1}\mathbf{J}_\varphi(\mathbf{x})$ is an orthogonal matrix for all $\mathbf{x}$ in U. We call λ the scale factor of φ.*

While it can be shown that *linear* conformal maps are similarities, an interesting class of *nonlinear* conformal maps are the unit radius sphere inversion (restriction to unit radius is only to avoid unnecessary notational complexity):

$$\begin{aligned} I_\mathbf{b} : \mathbb{R}^n \setminus \{0\} &\to \mathbb{R}^n \setminus \{0\} \\ \mathbf{x} &\mapsto \frac{\mathbf{x} - \mathbf{b}}{\|\mathbf{x} - \mathbf{b}\|^2} + \mathbf{b} \end{aligned}$$

We can notice that such transformation leaves the hypersphere of center $\mathbf{b}$ and radius 1 invariant, while the points outside of the unit ball are mapped to the interior of the unit ball, and vice-versa.

Interestingly, conformal maps in Euclidean spaces of dimension superior or equal to 3 can be restricted to two kinds according to the following result from Liouville.

Theorem B.5.2 (see e.g. [498]) *Let $f : U \to \mathbb{R}^n$ be a conformal map defined on a connected open subset of Euclidean space $\mathbb{R}^n$ of dimension $n \geq 3$. Then $f = L_{|U}$ can be written either as the restriction of a similarity L to U, or as the composition $f = I \circ L_{|U}$ of such a map with an inversion with respect to a hypersphere of unit radius, centered at the origin.*

The class of function described in Thm. B.5.2 corresponds exactly to the Möbius transformations described in (B.17). These transformation can as well be defined in dimension 2, with the specificity that they are only a subset of the class conformal maps in this dimension.

Properties of sphere inversion. We characterize the properties of the unit sphere centered at zero, that we denote I

$$\begin{aligned} I : \mathbb{R}^n \setminus \{0\} &\to \mathbb{R}^n \setminus \{0\} \\ \mathbf{x} &\mapsto \frac{\mathbf{x}}{\|\mathbf{x}\|^2} \end{aligned}$$

Now let us derive the Jacobian of I. A straightforward computation leads to

$$\mathbf{J}_I(\mathbf{x}) = \frac{1}{\|\mathbf{x}\|^2}\left(\mathbf{I}_n - 2\frac{\mathbf{x}\mathbf{x}^\top}{\|\mathbf{x}\|^2}\right)$$

where $\mathbf{I}_n$ denote the identity matrix.

By noticing that $\frac{\mathbf{x}\mathbf{x}^\top}{\|\mathbf{x}\|^2}$ is rank one symmetric with eigenvalue 1 associated with unit norm eigenvector $\frac{\mathbf{x}}{\|\mathbf{x}\|}$, we can diagonalize this matrix in any (space dependent) orthogonal basis that has $\frac{\mathbf{x}}{\|\mathbf{x}\|}$ as the first basis vector.

Let us thus pick the unit vectors associated to the hyperspherical coordinates (which satisfy this condition by definition), and consider the orthogonal matrix $\mathbf{B}(\frac{\mathbf{x}}{\|\mathbf{x}\|})$ gathering these basis vectors as its columns (it is parameterized by the unit vector $\frac{\mathbf{x}}{\|\mathbf{x}\|}$, as this basis is radially invariant. Then we can write

$$\frac{\mathbf{x}\mathbf{x}^\top}{\|\mathbf{x}\|^2} = \mathbf{B}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)\mathbf{D}\mathbf{B}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)^\top$$

and thus

$$\mathbf{J}_I(\mathbf{x}) = \frac{1}{\|\mathbf{x}\|^2}\left(\mathbf{I}_n - 2\mathbf{B}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)\mathbf{D}\mathbf{B}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)^\top\right) = \frac{1}{\|\mathbf{x}\|^2}\mathbf{B}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)(\mathbf{I}_n - 2\mathbf{D})\,\mathbf{B}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)^\top$$

with $\mathbf{D}$ a diagonal matrix with diagonal elements $[1, 0, \ldots, 0]$. This leads to

$$\mathbf{J}_I(\mathbf{x}) = \frac{1}{\|\mathbf{x}\|^2}\mathbf{B}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)\mathbf{D}_I\mathbf{B}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)^\top$$

with $\mathbf{D}_I = \mathbf{I}_n - 2\mathbf{D}$ a diagonal matrix with diagonal elements $[-1, 1, \ldots, 1]$. The Jacobian thus takes the form predicted by the above proposition for conformal maps

$$\mathbf{J}_I(\mathbf{x}) = \lambda(\mathbf{x})\mathbf{O}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)$$

with scale factor $\lambda(\mathbf{x}) = \frac{1}{\|\mathbf{x}\|^2}$ and $\mathbf{O}(\frac{\mathbf{x}}{\|\mathbf{x}\|}) = \mathbf{B}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)\mathbf{D}_I\mathbf{B}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)^\top$ a space dependent orthogonal matrix, which has the additional property to be radially invariant for the specific case of sphere inversions.

C

Additional Material on Chapter 4

C.1 Complementary notes

C.1.1 ELBO decompositions

Connection between (4.1) and (4.2). Here we show how the two decompositions of the ELBO objective in (4.1) and (4.2) can be connected. We start from equation (4.2):

$$\text{ELBO}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \log p_\theta(\mathbf{x}) - \text{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right].$$

By definition of KL-divergence, and applying Bayes rule, we get

$$\begin{aligned}\text{ELBO}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) &= \log p_\theta(\mathbf{x}) - \int q_\phi(\mathbf{s}|\mathbf{x}) \left(\log q_\phi(\mathbf{s}|\mathbf{x}) - \log p_\theta(\mathbf{s}|\mathbf{x})\right) d\mathbf{s} \\ &= \log p_\theta(\mathbf{x}) - \int q_\phi(\mathbf{s}|\mathbf{x}) \left(\log q_\phi(\mathbf{s}|\mathbf{x}) - \log \left(p_\theta(\mathbf{x}|\mathbf{s}) \frac{p_\mathbf{s}(\mathbf{s})}{p_\theta(\mathbf{x})}\right)\right) d\mathbf{s}\,.\end{aligned}$$

We observe that the two terms involving $p_\theta(\mathbf{x})$ cancel, resulting in

$$\text{ELBO}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = -\int q_\phi(\mathbf{s}|\mathbf{x}) \left(\log q_\phi(\mathbf{s}|\mathbf{x}) - \log \left(p_\theta(\mathbf{x}|\mathbf{s}) p_\mathbf{s}(\mathbf{s})\right)\right) d\mathbf{s},$$

which leads to (4.1) by rearranging the terms:

$$\text{ELBO}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{q_\phi(\mathbf{s}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{s})\right] - \text{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\mathbf{s}(\mathbf{s})\right].$$

Expressions for the two terms in equation (4.1) under Assumption 4.3.1. The above two terms take the following form in our setting. For the second ("KL") term, we get

$$\begin{aligned}-\text{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\mathbf{s}(\mathbf{s})\right] &= \int q_\phi(\mathbf{s}|\mathbf{x}) \log p_\mathbf{s}(\mathbf{s}) d\mathbf{s} - \int q_\phi(\mathbf{s}|\mathbf{x}) \log q_\phi(\mathbf{s}|\mathbf{x}) d\mathbf{s} \\ &= \mathbb{E}_{q_\phi(\mathbf{s}|\mathbf{x})}[\log(p_\mathbf{s}(\mathbf{s}))] + H(q_\phi(\mathbf{s}|\mathbf{x}))\,,\end{aligned}$$

where H denotes the entropy. Writing the expression for the entropy of univariate Gaussian variables ($1/2 \log(2\pi\sigma^2) + 1/2$), we have under Assumption 4.3.1

$$H(q_\phi(\mathbf{s}|\mathbf{x})) = \frac{n}{2}\left(\log(2\pi) + 1\right) + \frac{1}{2}\sum_{k=1}^{n} \log \sigma_k^\phi(\mathbf{x})^2 = \kappa_n + \frac{1}{2}\sum_{k=1}^{n} \log \sigma_k^\phi(\mathbf{x})^2,$$

where we introduce the dimension dependent constant $\kappa_n = \frac{n}{2}(\log(2\pi)+1)$. This leads to

$$-\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_{\mathbf{s}}(\mathbf{s})\right] = \mathbb{E}_{q_\phi(\mathbf{s}|\mathbf{x})}[\log(p_{\mathbf{s}}(\mathbf{s}))] + \frac{1}{2}\sum_{k=1}^{n}\log\sigma_k^\phi(\mathbf{x})^2 + \kappa_n\,. \quad \text{(C.1)}$$

The first ("reconstruction") term, under the isotropic Gaussian decoder of Assumption 4.3.1, takes the form

$$\mathbb{E}_{q_\phi(\mathbf{s}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{s})\right] = -\frac{\gamma^2}{2}\mathbb{E}_{q_\phi(\mathbf{s}|\mathbf{x})}\left[\|\mathbf{x}-\mathbf{f}^\theta(\mathbf{s})\|^2\right] + n\log\gamma - \frac{n}{2}\log(2\pi)\,. \quad \text{(C.2)}$$

Expression for the gap between ELBO and log-likelihood Let us now write the KL divergence between variational and true posteriors, which is the gap appearing in (4.2).

$$\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] = -\int q_\phi(\mathbf{s}|\mathbf{x})\log p_\theta(\mathbf{x}|\mathbf{s})d\mathbf{s} - H(q_\phi(\mathbf{s}|\mathbf{x}))$$

Using again the expression of the entropy of Gaussian variables, this leads to

$$\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] = -\int q_\phi(\mathbf{s}|\mathbf{x})\log p_\theta(\mathbf{x}|\mathbf{s})d\mathbf{s} - \sum_{k=1}^{n}\log\sigma_k^\phi(x) - \frac{n}{2}(\log(2\pi)+1)\,,$$

such that, using the Bayes formula for the true posterior and Assum. 4.3.1, we get

$$\begin{aligned}\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] &= -\sum_{k=1}^{n}\log\sigma_k^\phi(\mathbf{x}) + c(\mathbf{x},\gamma) \\ &\quad + \frac{1}{2}\mathbb{E}_{z\sim q_\phi(\cdot|\mathbf{x})}\left[\|\mathbf{x}-\mathbf{f}^\theta(\mathbf{s})\|^2\gamma^2 - 2\sum_{k=1}^{n}\log d(z_k)\right], \quad \text{(C.3)}\end{aligned}$$

with additive constant $c(\mathbf{x},\gamma) = -\frac{n}{2}\left(\log(\gamma^2)+1\right) + \log p_\theta(\mathbf{x})$. Note the $\log(2\pi)$ term in the previous expression cancels with the one coming from the true log posterior.

The analysis of the optima of (C.3) is non-trivial due to the second term which involves taking expectations of functions of $\mathbf{s}$ w.r.t. its posterior distribution q_ϕ parameterized by $\boldsymbol{\mu}^\phi$ and $\boldsymbol{\sigma}^\phi$. Much of the derivations to obtain our results will revolve around constructing bounds that no longer involve such expectations, but instead only depend on $\boldsymbol{\mu}^\phi$ and $\boldsymbol{\sigma}^\phi$.

C.1.2 Justification of the intuition

We add here more qualitative details to the statement of subsection 4.3.2 that the true posterior density is approximately the pushforward of $p_\theta(\mathbf{x}|\mathbf{s}=\mathbf{s}_0)$. Note that they are not meant to replace a rigorous treatment, which is deferred to Appendix C.2.

As the decoder becomes deterministic, the marginal observed density becomes the pushforward[1] of the latent prior by $\mathbf{f}^\theta$ such that

1: because the conditional distribution of the decoder tends to a Dirac measure at $\mathbf{f}^\theta$

$$p_{\boldsymbol{\theta}}(\mathbf{x}) \approx p_0\left(\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})\right)\left|\mathbf{J}_{\mathbf{g}^{\boldsymbol{\theta}}}(\mathbf{x})\right|.$$

The true posterior is therefore approximately

$$p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})p_{\mathbf{s}}(\mathbf{s})/p_{\boldsymbol{\theta}}(\mathbf{x}) \approx p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})p_{\mathbf{s}}(\mathbf{s})/p_0\left(\mathbf{g}^{\boldsymbol{\theta}}(x)\right)\left|\mathbf{J}_{\mathbf{g}^{\boldsymbol{\theta}}}(\mathbf{x})\right|^{-1}.$$

Conditioning on a given observation $\mathbf{x} = \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0)$, we get

$$\begin{aligned} p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x} = \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0)) &= p_{\boldsymbol{\theta}}(\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0)|\mathbf{s})p_{\mathbf{s}}(\mathbf{s})/p_{\boldsymbol{\theta}}(\mathbf{x} = \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0)) \\ &\approx p_{\boldsymbol{\theta}}(\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0)|\mathbf{s})p_{\mathbf{s}}(\mathbf{s})/p_0\left(\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0))\right)\left|\mathbf{J}_{\mathbf{g}^{\boldsymbol{\theta}}}(\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0))\right|^{-1} \\ &\approx p_{\boldsymbol{\theta}}(\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0)|\mathbf{s})p_{\mathbf{s}}(\mathbf{s})/p_0(\mathbf{s}_0)\left|\mathbf{J}_{\mathbf{g}^{\boldsymbol{\theta}}}(\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0))\right|^{-1} \end{aligned}$$

Neglecting the variations of the prior relative to those of the posterior (due to near-determinism), we make the approximation $p_{\mathbf{s}}(\mathbf{s}) \approx p_0(\mathbf{s}_0)$ such that the above approximation becomes

$$p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x} = \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0)) \approx p_{\boldsymbol{\theta}}(\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0)|\mathbf{s})\left|\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}}(\mathbf{s}_0)\right|.$$

Using the isotropic Gaussian decoder assumption, we get

$$p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x} = \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0)) \approx \frac{\gamma^n}{\sqrt{2\pi}^n}\exp\left(-\frac{\gamma^2}{2}\left\|\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0) - \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})\right\|^2\right)\left|\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}}(\mathbf{s}_0)\right|.$$

In the near-deterministic regime, this posterior distribution should be concentrated in the region where $\mathbf{s}$ is close to $\mathbf{s}_0$, we can then further approximate this density using a Taylor formula

$$\begin{aligned} p_{\boldsymbol{\theta}}(z|\mathbf{x} = \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0)) &\approx \frac{\gamma^n}{\sqrt{2\pi}^n}\exp\left(-\frac{\gamma^2}{2}\left\|\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}}(\mathbf{s}_0)(\mathbf{s}_0 - \mathbf{s})\right\|^2\right)\left|\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}}(\mathbf{s}_0)\right| \\ &= \frac{\sqrt{2\pi}^{-n}\gamma^n}{\sqrt{\left|\mathbf{G}\mathbf{G}^T\right|}}\exp\left(-\frac{1}{\gamma^2}(\mathbf{s}_0 - \mathbf{s})^T\left(\mathbf{G}\mathbf{G}^T\right)^{-1}(\mathbf{s}_0 - \mathbf{s})\right), \end{aligned}$$

with $\mathbf{G} = \mathbf{J}_{\mathbf{g}^{\boldsymbol{\theta}}}(\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0)) = \mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}}(\mathbf{s}_0)^{-1}$, which is also matching the expression of the pushforward of the Gaussian density $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s} = \mathbf{s}_0)$ by the linearization of $\mathbf{g}^{\boldsymbol{\theta}}$ around $\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}_0)$ (i.e. replacing the mapping by its Jacobian at that point, $\mathbf{G}$).

C.1.3 A connection between the β parameter of β-VAEs and the decoder precision γ^2

In the context of disentanglement, a commonly used variant of standard VAEs [191] is the β-VAE [194, 197, 198, 202, 203]. In this model, an additional parameter β is added to modify the weight of the KL term in (4.1), whereas the decoder precision γ^2 is typically set to one [202, 203, 499, 500]. The β-VAE objective [194] can be written as

$$\mathscr{L}_\beta(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})\right] - \beta KL\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})\|p_{\mathbf{s}}(\mathbf{s})\right]. \tag{C.4}$$

The influence of the decoder precision γ^2 and the β parameters on

the objective have been related in the literature, see for example [193, § 2.4.3]—and similar observations can be found in [218, § 3.1]. Under the assumption of a Gaussian decoder, the ELBO from Equation 4.1 can be written as

$$\begin{aligned}\text{ELBO}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\phi}) &= -\text{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})||p_{\mathbf{s}}(\mathbf{s})\right] + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})\right] \\ &= -\text{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})||p_{\mathbf{s}}(\mathbf{s})\right] - \frac{\gamma^2}{2}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})}\left[\left\|\mathbf{x}-\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})\right\|^2\right] + n\log\gamma - \frac{n}{2}\log(2\pi) \\ &= \gamma^2\left[-\frac{1}{\gamma^2}\text{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})||p_{\mathbf{s}}(\mathbf{s})\right] - \frac{1}{2}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})}\left[\left\|\mathbf{x}-\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})\right\|^2\right] + c(\gamma)\right]; \\ c(\gamma) &:= \frac{n}{\gamma^2}\log\gamma - \frac{n}{2\gamma^2}\log(2\pi)\end{aligned} \tag{C.5}$$

Given that usually optimization is performed with a fixed value for γ for the ELBO (and with fixed β for $\mathcal{L}_\beta$), this suggests that β and $1/\gamma^2$, play a similar role in (C.4) and (C.5)—since the γ^2 outside parenthesis only changes the objective and its gradients by a global scaling factor.

Why is proving self-consistency for β-VAEs harder? The connection of β and γ^2 above makes the following statement counterintuitive: proving self-consistency for β-VAEs is harder. The reason is that β is a parameter in (C.4) (*i.e.*, (4.1) is modified), whereas the proof for self-consistency uses the ELBO-decompoisition with the log-likelihood and the KL between variational and true posteriors (4.2). As we cannot express (C.4) in a form with the gap as in (4.2) and it is not necessarily a lower bound (β can be smaller than 1), proving self-consistency is more complicated.

C.2 Main Theoretical Results

C.2.1 Proof of Proposition 4.3.1

We proceed in two steps: first we prove the existence of variational parameters that achieve a global minimum of the ELBO gap, then we characterize its near-deterministic properties. We then combine these results, which rely on specific assumptions, to obtain our main text result under Assumption 4.3.1.

We initially use the following milder assumptions than in main text to prove intermediate results.

Assumption C.2.1 (Gaussian Encoder-Gaussian Decoder VAE, minimal properties) *We are given a fixed latent prior and three parameterized classes of $\mathbb{R}^n \to \mathbb{R}^n$ mappings: the mean decoder class $\boldsymbol{\theta} \mapsto \mathbf{f}^{\boldsymbol{\theta}}$, and the mean and standard deviation encoder classes, $\boldsymbol{\phi} \mapsto \boldsymbol{\mu}^{\boldsymbol{\phi}}$ and $\boldsymbol{\phi} \mapsto \boldsymbol{\sigma}^{\boldsymbol{\phi}}$ such that*

(i) the latent prior has a factorized independent and identically distributed (i.i.d.) density $p_{\mathbf{s}}(\mathbf{s}) \sim \prod_k d(s_k)$, with d smooth fully supported on $\mathbb{R}$, with concave $\log d$,

(ii) conditional on the latent, the decoder has a factorized Gaussian density

$p_{\boldsymbol{\theta}}$ *with mean* $\mathbf{f}^{\boldsymbol{\theta}}$ *such that*

$$\mathbf{x}|\mathbf{s} \sim \mathcal{N}\left(\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}), \gamma^{-2}\mathbf{I}_n\right) \tag{C.6}$$

(iii) the encoder is factorized Gaussian with posterior mean and variance maps $\mu_k^{\boldsymbol{\phi}}(\mathbf{x})$, $\sigma_k^{\boldsymbol{\phi}}(\mathbf{x})^2$ *for each component* k*, leading to the factorized posterior density* $q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$ *such that*

$$s_k|\mathbf{x} \sim \mathcal{N}(\mu_k^{\boldsymbol{\phi}}(\mathbf{x}), \sigma_k^{\boldsymbol{\phi}}(\mathbf{x})^2) \tag{C.7}$$

(iv) the mean and variance encoders classes can fit any function,
(v) for all possible $\boldsymbol{\theta}$, $\mathbf{f}^{\boldsymbol{\theta}}$ *is a diffeomorphism of* $\mathbb{R}^n$ *with inverse* $\mathbf{g}^{\boldsymbol{\theta}}$.

Existence of at least one global minimizer of the gap between true and variational posterior is given by the following proposition.

Proposition C.2.1 (Existence of global minimum) *Under Assumption C.2.1. For a fixed* $\boldsymbol{\theta}$ *assume additionally that* $\mathbf{g}^{\boldsymbol{\theta}}$ *is Lipschitz continuous with Lipschitz constant* $B > 0$*, in the sense that*

$$\forall \mathbf{x}, \boldsymbol{y} \in \mathbb{R}^n : \qquad \left\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) - \mathbf{g}^{\boldsymbol{\theta}}(\boldsymbol{y})\right\|_2 \leq B\|\mathbf{x} - \boldsymbol{y}\|_2 .$$

Then there exists at least one choice $(\boldsymbol{\mu}^{\boldsymbol{\phi}} \in \mathbb{R}^n, \boldsymbol{\sigma}^{\boldsymbol{\phi}} \in \mathbb{R}^n_{>0})$ *that achieves the minimum of* $KL\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})\right]$.

Proof. Using Prop. C.3.1, we have the lower bound

$$\begin{aligned}\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})\right] \geq &-\sum_{k=1}^{n}\left[\log\sigma_k^{\boldsymbol{\phi}}(\mathbf{x}) + \log d(\mu_k^{\boldsymbol{\phi}})\right] + c(\mathbf{x},\gamma) \\ &+ \frac{\gamma^2}{2}B^{-2}\left[\left\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\mathbf{x})\right\|^2 + \sum_{k=1}^{n}\sigma_k^{\boldsymbol{\phi}}(\mathbf{x})^2\right] .\end{aligned} \tag{C.8}$$

We then notice (see lemma C.3.4) that for all k,

$$\sigma_k^{\boldsymbol{\phi}}(\mathbf{x}) \longrightarrow -\log\sigma_k^{\boldsymbol{\phi}}(\mathbf{x}) + \frac{\gamma^2}{2}B^{-2}\sigma_k^{\boldsymbol{\phi}}(\mathbf{x})^2$$

achieves a global minimum $m(B,\gamma) = -\log(B/\gamma) + 1/2$ at $\sigma_k^{\boldsymbol{\phi}}(\mathbf{x}) = B/\gamma$.

For arbitrary k_0, we now 1) lower bound the $k \neq k_0$ terms by $m(B,\gamma)$; 2) lower bound and all the $\log d$ terms by their global maximum, which exists by Assum. 1i (log-concave prior); and 3) drop the non-negative squared norm term, leading to the following weaker lower bound:

$$\begin{aligned}\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})\right] \geq &\,(n-1)m(B,\gamma) - \log\sigma_{k_0}^{\boldsymbol{\phi}}(\mathbf{x}) \\ &- n\max_t(\log m(t)) + c(\mathbf{x},\gamma) + \frac{\gamma^2}{2}B^{-2}\left[\sigma_{k_0}^{\boldsymbol{\phi}}(\mathbf{x})^2\right] .\end{aligned} \tag{C.9}$$

The KL divergence is well-defined and finite for any choice of parameters in their domain, therefore it achieves a particular value $K_0 \geq 0$ at one arbitrary selected point of the domain. Since for all k, the lower bound tends to $+\infty$ for both $\sigma_k^{\boldsymbol{\phi}} \to +\infty$ (as the quadratic term dominates the

$-\log$ term) and $\sigma_k^\phi \to 0^+$, there exist $a > b > 0$ (possibly dependent on $(\gamma, \mathbf{x})$) such that $\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] > K_0$ for any $\sigma_k^\phi < b$ or $\sigma_k^\phi > a$.

Moreover, starting again from the lower bound from Prop. C.3.1,

$$\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] \geq -\sum_{k=1}^{n}\left[\log \sigma_k^\phi(\mathbf{x}) + \log d(\mu_k^\phi)\right] + c(\mathbf{x},\gamma) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\mathbf{g}^\theta(\mathbf{x}) - \boldsymbol{\mu}^\phi(\mathbf{x})\right\|^2 + \sum_{k=1}^{n}\sigma_k^\phi(\mathbf{x})^2\right], \quad \text{(C.10)}$$

we now focus on $\boldsymbol{\mu}^\phi$ and lower bound all $\boldsymbol{\sigma}^\phi$ terms. With this, we get the following weaker lower bound in terms of $\boldsymbol{\mu}^\phi$:

$$\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] \geq nm(B,\gamma) - n\max_t(\log m(t)) + c(\mathbf{x},\gamma) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\mathbf{g}^\theta(\mathbf{x}) - \boldsymbol{\mu}^\phi(\mathbf{x})\right\|^2\right]. \quad \text{(C.11)}$$

The lower bound also tends to $+\infty$ for $\|\boldsymbol{\mu}^\phi\| \to +\infty$, so there exists a radius $R > 0$ (possibly dependent on $(\gamma, \mathbf{x})$) such that $\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] > K_0$ if $\|\boldsymbol{\mu}^\phi\| > R$.

As a consequence, the infimum ($\leq K_0$) of the minimization problem (4.6) cannot be achieved outside the compact set $(\boldsymbol{\mu}^\phi, \boldsymbol{\sigma}^\phi) \in \{\boldsymbol{\mu}^\phi \in \mathbb{R}^n : \|\boldsymbol{\mu}^\phi\| \leq R\} \times [a,b]^n$. Since the divergence is continuous in $(\boldsymbol{\mu}^\phi, \boldsymbol{\sigma}^\phi)$, there exists a value $(\boldsymbol{\mu}^{\widehat{\phi}}, \boldsymbol{\sigma}^{\widehat{\phi}})$ in this compact set achieving the minimum of the KL over the whole parameter domain, and all values achieving this minimum are in this compact set. □

For given $\mathbf{x}$, $\boldsymbol{\theta}$ and $\gamma > 0$, the variational posterior KL divergence mapping

$$(\boldsymbol{\mu}^\phi(\mathbf{x}), \boldsymbol{\sigma}^\phi(\mathbf{x})) \to \mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right]$$

thus has a minimum, and by smoothness of this mapping, this minimum can be characterized by the vanishing gradient of the KL divergence with respect to the parameters. Now, let us try to characterize how this minimum behaves for large γ.

Proposition C.2.2 (Self-consistency of the encoder in the deterministic limit) *Under Assumption C.2.1, assume additionally $\mathbf{f}^\theta$ and $\mathbf{g}^\theta$ are Lipschitz continuous with respective Lipschitz constants $C, B > 0$, in the sense that*

$$\forall \mathbf{s}, \boldsymbol{w} \in \mathbb{R}^n : \quad \left\|\mathbf{f}^\theta(\mathbf{s}) - \mathbf{f}^\theta(\boldsymbol{w})\right\|_2 \leq C\|\mathbf{s} - \boldsymbol{w}\|_2, \quad \text{(C.12)}$$

$$\forall \mathbf{x}, \boldsymbol{y} \in \mathbb{R}^n : \quad \left\|\mathbf{g}^\theta(\mathbf{x}) - \mathbf{g}^\theta(\boldsymbol{y})\right\|_2 \leq B\|\mathbf{x} - \boldsymbol{y}\|_2. \quad \text{(C.13)}$$

Assume additionally that $-\log d$ is quadratically dominated, in the sense that

$$\exists D > 0, E > 0 : \quad -\log d(u) \leq D|u|^2 + E, \quad \forall u \in \mathbb{R}.$$

Then for all $\mathbf{x}, \boldsymbol{\theta}$, *as* $\gamma \to +\infty$, *any global minimum of (4.6) satisfies*

$$\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x}) = \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) + O(1/\gamma) \tag{C.14}$$

$$\sigma^{\widehat{\phi}}(\mathbf{x})^2 = O(1/\gamma^2)\,. \tag{C.15}$$

More precisely, for all $\mathbf{x} \in \mathbb{R}^n$, $\gamma > 0$

$$\left\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})\right\|^2 \leq B^2 \frac{2n}{\gamma^2}\left(\frac{1}{2}(C^2-1) + E + D\left[\frac{\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})\|^2}{n} + \frac{1}{\gamma^2}\right] + M + \frac{1}{2}\log(B^2)\right).$$

and

$$\sum_{k=1}^{n} \sigma_k^{\widehat{\phi}}(\mathbf{x})^2 \leq B^2 \frac{4n}{\gamma^2}\left(\frac{1}{2}(C^2-1) + E + D\left[\frac{\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})\|^2}{n} + \frac{1}{\gamma^2}\right] + M + \frac{1}{2}\left(\log(2B^2)\right)\right).$$

Proof. We start from the lower bound expression of Prop. C.3.1

$$\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] \geq -\sum_{k=1}^{n}\left[\log \sigma_k^{\phi}(\mathbf{x}) + \log d(\mu_k^{\phi})\right] + c(\mathbf{x},\gamma) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\mu}^{\phi}\right\|^2 + \sum_{k=1}^{n}\sigma_k^{\phi}(\mathbf{x})^2\right],$$

with $c(\mathbf{x},\gamma) = -\frac{n}{2}\left(\log(\gamma^2)+1\right) + \log p_\theta(\mathbf{x})$. For any $\nu \in (0,1]$, we can thus write

$$\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] \geq \sum_{k=1}^{n}\left[-\log \sigma_k^{\phi}(\mathbf{x}) + \nu\gamma^2 B^{-2}\frac{\sigma_k^{\phi}(\mathbf{x})^2}{2} - \log d(\mu_k^{\phi})\right] + c(\mathbf{x},\gamma) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\mu}^{\phi}\right\|^2 + (1-\nu)\sum_{k=1}^{n}\sigma_k^{\phi}(\mathbf{x})^2\right].$$

Now, from lemma C.3.4 we get

$$\forall u > 0: \qquad -\log u + \alpha u^2/2 \geq \frac{1}{2}\log(\alpha) + \frac{1}{2}\,.$$

We exploit this lower bound to obtain

$$\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] \geq \frac{n}{2}\left(\log(\nu\gamma^2B^{-2})+1\right) - \sum_{k=1}^{n}\left[\log d(\mu_k^{\phi})\right] + c(\mathbf{x},\gamma) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\mu}^{\phi}\right\|^2 + (1-\nu)\sum_{k=1}^{n}\sigma_k^{\phi}(\mathbf{x})^2\right].$$

Using the expression of $c(\mathbf{x}, \gamma)$ we get

$$\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] \geq \frac{n}{2}\left(\log(\nu B^{-2}) + \log\gamma^2 + 1\right) - \sum_{k=1}^{n}\left[\log d(\mu_k^\phi)\right] - \frac{n}{2}\left(\log\gamma^2 + 1\right) + \log p_\theta(\mathbf{x}) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\mathbf{g}^\theta(\mathbf{x}) - \boldsymbol{\mu}^\phi\right\|^2 + (1-\nu)\sum_{k=1}^{n}\sigma_k^\phi(\mathbf{x})^2\right].$$

and both the "$n \log \gamma$" as well as "$n/2$" terms cancel out such that

$$\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] \geq \frac{n}{2}\left(\log(\nu B^{-2})\right) - \sum_{k=1}^{n}\left[\log d(\mu_k^\phi)\right] + \log p_\theta(\mathbf{x}) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\mathbf{g}^\theta(\mathbf{x}) - \boldsymbol{\mu}^\phi\right\|^2 + (1-\nu)\sum_{k=1}^{n}\sigma_k^\phi(\mathbf{x})^2\right].$$

Finally, using Prop. C.3.2, the above right hand side is bounded from above by a constant as $\gamma \to +\infty$, and as a consequence, the positive factor of the γ^2 term must vanish (by continuity assumption and its limits note $-\log d$ is bounded from below)

$$\left\|\mathbf{g}^\theta(\mathbf{x}) - \boldsymbol{\mu}^\phi\right\|^2 + (1-\nu)\sum_{k=1}^{n}\sigma_k^\phi(\mathbf{x})^2 \to 0$$

This entails that both positive terms it comprises must vanish too.

More precisely, we get the inequality between lower and upper bounds at the optimal solution

$$\begin{aligned}\frac{n}{2}\left(\log(\nu B^{-2})\right) - \sum_{k=1}^{n}\left[\log d(\mu_k^{\widehat{\phi}})\right] + \log p_\theta(\mathbf{x}) \\ + \frac{\gamma^2}{2}B^{-2}\left[\left\|\mathbf{g}^\theta(\mathbf{x}) - \boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})\right\|^2 + (1-\nu)\sum_{k=1}^{n}\sigma_k^{\widehat{\phi}}(\mathbf{x})^2\right] \\ \leq n\left(\frac{1}{2}C^2 + E + D\left[\frac{\|\mathbf{g}^\theta(\mathbf{x})\|^2}{n} + \frac{1}{\gamma^2}\right]\right) - \frac{n}{2} + \log p_\theta(\mathbf{x}),\end{aligned}$$

which simplifies to

$$\begin{aligned}\frac{n}{2}\left(\log(\nu B^{-2})\right) - \sum_{k=1}^{n}\left[\log d(\mu_k^{\widehat{\phi}})\right] + \frac{\gamma^2}{2}B^{-2}\left[\left\|\mathbf{g}^\theta(\mathbf{x}) - \boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})\right\|^2 + (1-\nu)\sum_{k=1}^{n}\sigma_k^{\widehat{\phi}}(\mathbf{x})^2\right] \\ \leq n\left(\frac{1}{2}C^2 + E + D\left[\frac{\|\mathbf{g}^\theta(\mathbf{x})\|^2}{n} + \frac{1}{\gamma^2}\right]\right) - \frac{n}{2}.\end{aligned}$$

Moreover by continuity assumption and its limits, $-\log d$ is bounded from below by $-M = -\max_t \log d(t)$, yielding

$$\begin{aligned}\frac{n}{2}\left(\log(\nu B^{-2}) - 2M\right) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\mathbf{g}^\theta(\mathbf{x}) - \boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})\right\|^2 + (1-\nu)\sum_{k=1}^{n}\sigma_k^{\widehat{\phi}}(\mathbf{x})^2\right] \\ \leq n\left(\frac{1}{2}(C^2-1) + E + D\left[\frac{\|\mathbf{g}^\theta(\mathbf{x})\|^2}{n} + \frac{1}{\gamma^2}\right]\right)\end{aligned}$$

such that

$$\frac{\gamma^2}{2}B^{-2}\left[\left\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})-\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})\right\|^2+(1-\nu)\sum_{k=1}^{n}\sigma_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})^2\right]$$
$$\leq n\left(\frac{1}{2}(C^2-1)+E+D\left[\frac{\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})\|^2}{n}+\frac{1}{\gamma^2}\right]-\frac{1}{2}\left(\log(\nu B^{-2})-2M\right)\right)$$

and finally

$$B^{-2}\left[\left\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})-\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})\right\|^2+(1-\nu)\sum_{k=1}^{n}\sigma_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})^2\right]$$
$$\leq \frac{2n}{\gamma^2}\left(\frac{1}{2}(C^2-1)+E+D\left[\frac{\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})\|^2}{n}+\frac{1}{\gamma^2}\right]+M+\frac{1}{2}\log(B^2/\nu)\right) \tag{C.16}$$

Taking $\nu = 1$ in (C.16) we get the first intended inequality

$$\left\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})-\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})\right\|^2 \leq B^2\frac{2n}{\gamma^2}\left(\frac{1}{2}(C^2-1)+E+D\left[\frac{\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})\|^2}{n}+\frac{1}{\gamma^2}\right]+M+\frac{1}{2}\log(B^2)\right).$$

Alternatively, (C.16) implies

$$(1-\nu)\sum_{k=1}^{n}\sigma_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})^2 \leq B^2\frac{2n}{\gamma^2}\left(\frac{1}{2}(C^2-1)+E+D\left[\frac{\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})\|^2}{n}+\frac{1}{\gamma^2}\right]+M+\frac{1}{2}\left(\log(B^2/\nu)\right)\right)$$

Taking a fixed value of ν, say $1/2$, we get the second intended inequality

$$\sum_{k=1}^{n}\sigma_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})^2 \leq B^2\frac{4n}{\gamma^2}\left(\frac{1}{2}(C^2-1)+E+D\left[\frac{\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})\|^2}{n}+\frac{1}{\gamma^2}\right]+M+\frac{1}{2}\left(\log(2B^2)\right)\right).$$

□

We now restate the main text proposition and provide the proof.

Proposition 4.3.1 *[Self-consistency of near-deterministic VAEs] Under Assumption 4.3.1, for all* $\mathbf{x}$, $\boldsymbol{\theta}$, *as* $\gamma \to +\infty$, *there exists at least one global minimum solution of (4.6). These solutions satisfy*

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x}) = \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) + O(1/\gamma) \quad \text{and} \quad \sigma_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})^2 = O(1/\gamma^2)\,, \text{ for all } k\,. \tag{4.8}$$

Proof. We only have to check that Assumption 4.3.1 allow fulfilling the following requirements of Prop. C.2.2:

- the Lipschitz continuity requirements in Prop. C.2.2 results from the boundedness of the first order derivatives of the decoder mean and of its inverse (by using the multivariate Taylor theorem),

- concavity of $\log d$, required by Assumption C.2.1, is a direct consequence of non-positivity of the second-order logarithmic derivative of m in Assumption 4.3.1i,
- quadratic domination of $-\log d$ comes from the boundedness of the second-order logarithmic derivative of m (by integrating twice).

Then Prop. C.2.2 follows and the $O(1/\gamma)$ convergence of the variational posterior mean of the inverse, as well as the $O(1/\gamma^2)$ convergence of the variational posterior variance. □

Finer approximation of parameter values We now derive a finer result for the convergence of the mean, that we will exploit in Thm. 4.3.2. This relies on the existence of an optimum shown by Prop. C.2.1.

At such optimum $\widehat{\boldsymbol{\phi}}$ we thus have for all k

$$\frac{\partial}{\partial \mu_k^{\boldsymbol{\phi}}} \left[\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x}) || p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x}) \right] \right]_{|\widehat{\boldsymbol{\phi}}} = 0\,,$$

and

$$\frac{\partial}{\partial \sigma_k^{\boldsymbol{\phi}}} \left[\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x}) || p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x}) \right] \right]_{|\widehat{\boldsymbol{\phi}}} = 0\,.$$

We derive the constraints entailed by the first expression:

$$\begin{aligned}
\frac{\partial}{\partial \mu_k^{\boldsymbol{\phi}}} \left[\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x}) || p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x}) \right] \right]_{|\widehat{\boldsymbol{\phi}}} &= \frac{1}{2} \int \frac{\partial}{\partial \mu_k^{\boldsymbol{\phi}}} q_{\boldsymbol{\phi}}(\mathbf{s}) \left[\|\mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})\|^2 \gamma^2 - 2 \sum_{k=1}^{n} \log d(s_k) \right] d\mathbf{s} \\
&= \frac{1}{2} \int \prod_{j \neq k} q_{\boldsymbol{\phi}}^j(z_j) \frac{\partial q_{\boldsymbol{\phi}}^k(s_k)}{\partial \mu_k^{\boldsymbol{\phi}}} \left[\|\mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})\|^2 \gamma^2 - 2 \sum_{k=1}^{n} \log d(s_k) \right] d\mathbf{s}
\end{aligned}$$

with

$$\frac{\partial q_{\boldsymbol{\phi}}^k(s_k)}{\partial \mu_k^{\boldsymbol{\phi}}} = \frac{\mu_k^{\boldsymbol{\phi}} - s_k}{{\sigma_k^{\boldsymbol{\phi}}}^2} q_{\boldsymbol{\phi}}^k(s_k),$$

which leads to a set of constraints at optimum

$$\begin{aligned}
&\int q_{\widehat{\boldsymbol{\phi}}}(\mathbf{s}) \mu_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x}) \left[\|\mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})\|^2 \gamma^2 - 2 \sum_{k=1}^{n} \log d(s_k) \right] d\mathbf{s} \\
&\quad = \int q_{\widehat{\boldsymbol{\phi}}}(\mathbf{s}) s_k \left[\|\mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})\|^2 \gamma^2 - 2 \sum_{k=1}^{n} \log d(s_k) \right] d\mathbf{s}\,, \ \forall k \qquad \text{(C.17)}
\end{aligned}$$

Based on this expression we derive the following result.

Proposition C.2.3 *Under Assumption 4.3.1, as* $\gamma \longrightarrow +\infty$

$$\mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})) = \mathbf{x} + \frac{1}{\gamma^2} \mathbf{J}^{-T}_{\mathbf{f}^{\boldsymbol{\theta}} | \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})} n'(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})) + O(1/\gamma^3). \qquad \text{(C.18)}$$

and

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x}) = \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) + \frac{1}{\gamma^2}\mathbf{J}^{-1}_{\mathbf{f}^{\boldsymbol{\theta}}|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})}\mathbf{J}^{-T}_{\mathbf{f}^{\boldsymbol{\theta}}|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})}n'(\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})) + O(1/\gamma^3) \tag{C.19}$$

Proof. We start from the constraints of (C.17) that we rewrite

$$\begin{aligned}\int q_{\widehat{\boldsymbol{\phi}}}(\mathbf{s})\left(s_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})\right)\left[\|\mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})\|^2\gamma^2\right]d\mathbf{s} \\ = \int q_{\widehat{\boldsymbol{\phi}}}(\mathbf{s})\left(s_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})\right)\left[2\sum_{k=1}^{n}\log d(s_k)\right]d\mathbf{s}\end{aligned}$$

We then proceed to approximate the left hand side using a Taylor formula. Assuming bounded Hessian components, we can upper and lower bound using third order centered absolute moments of the Gaussian as

$$\gamma^2\int q_{\widehat{\boldsymbol{\phi}}}(\mathbf{s})\left(s_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})\right)\left[\|\mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})) - \mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})}(\mathbf{s} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x}))\|^2\right]d\mathbf{s} + O(1/\gamma),$$

which we can rewrite (by 1) expanding the norm of the sum; 2) removing constants in the bracket, which lead to zeros after multiplying the zero mean variable and taking the expectation; 3) using Gaussianity, all centered third order terms vanish.)

$$\begin{aligned}&\gamma^2\int q_{\widehat{\boldsymbol{\phi}}}(\mathbf{s})\left(s_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})\right)\Big[\|\mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x}))\|^2 + \|\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})}(\mathbf{s} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x}))\|^2 \\ &\qquad -2\left\langle \mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})), \mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})}(\mathbf{s} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x}))\right\rangle\Big]d\mathbf{s} + O(1/\gamma) \\ &\qquad = \gamma^2\int q_{\widehat{\boldsymbol{\phi}}}(\mathbf{s})\left(s_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})\right)\Big[\|\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})}(\mathbf{s} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x}))\|^2 \\ &\qquad -2\left\langle \mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})), \mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})}(\mathbf{s} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x}))\right\rangle\Big]d\mathbf{s} + O(1/\gamma) \\ &= \gamma^2\int q_{\widehat{\boldsymbol{\phi}}}(\mathbf{s})\left(s_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})\right)\Big[(\mathbf{s} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x}))^T\mathbf{J}^T_{\mathbf{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})}\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})}(\mathbf{s} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})) \\ &\qquad -2\left\langle \mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})), \mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})}(\mathbf{s} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x}))\right\rangle\Big]d\mathbf{s} + O(1/\gamma) \\ &= \gamma^2\int q_{\widehat{\boldsymbol{\phi}}}(\mathbf{s})\left(s_k - \mu_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})\right)\Big[-2\left\langle \mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})), \mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})}(\mathbf{s} - \boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x}))\right\rangle\Big]d\mathbf{s} + O(1/\gamma)\end{aligned}$$

Finally computing this integral we get the left hand side as

$$-2\gamma^2\sigma_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})^2\left\langle \mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})), [\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})}]_{.k}\right\rangle + O(1/\gamma)$$

For the right hand side we get using a Taylor expansion (with notation

$n : \mathbf{s} \longrightarrow \log(d(\mathbf{s}))$

$$\int q_{\widehat{\phi}}(\mathbf{s}) \left(s_k - \mu_k^{\widehat{\phi}}(\mathbf{x})) \right) \left[2 \sum_{k=1}^{n} \log d(s_k) \right] d\mathbf{s}$$

$$= \int q_{\widehat{\phi}}(\mathbf{s}) \left(s_k - \mu_k^{\widehat{\phi}}(\mathbf{x})) \right) \left[2 \sum_{k=1}^{n} \log d(\mu_k^{\widehat{\phi}}(\mathbf{x})) + n'(\mu_k^{\widehat{\phi}}(\mathbf{x}))(s_k - \mu_k^{\widehat{\phi}}(\mathbf{x})) \right] d\mathbf{s} + O(1/\gamma^2)$$

$$= 2\sigma_k^{\widehat{\phi}}(\mathbf{x})^2 n'(\mu_k^{\widehat{\phi}}(\mathbf{x})) + O(1/\gamma^2).$$

Equating the non-negligible terms of the left and right-hand sides we get for each k

$$\gamma^2 \left\langle \mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})),\ [\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})}]_{.k} \right\rangle = -n'(\mu_k^{\widehat{\phi}}(\mathbf{x})) + O(1/\gamma)$$

such that

$$(\mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})))^T \mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})} = -\frac{1}{\gamma^2} n'(\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})) + O(1/\gamma^3),$$

where n' is applied component-wise. Because the Jacobian is everywhere invertible (implicit consequence of Lipschitz assumptions), we can solve for this equations and get

$$\mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})) = \mathbf{x} + \frac{1}{\gamma^2} \mathbf{J}^{-T}_{\mathbf{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})} n'(\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})) + O(1/\gamma^3). \tag{C.20}$$

Using again a similar Taylor approximation we get

$$\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x}) = \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) + \frac{1}{\gamma^2} \mathbf{J}^{-1}_{\mathbf{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})} \mathbf{J}^{-T}_{\mathbf{f}^{\boldsymbol{\theta}}|\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})} n'(\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})) + O(1/\gamma^3).$$

This equation has the shortcoming of still referring to the posterior mean on both sides. To fix this, we first note that it implies, by boundedness of the Jacobian, that

$$|\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x}) - \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})| \leq \frac{1}{\gamma^2} K |n'(\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x}))| + O(1/\gamma^3).$$

By bounding the second-order derivative of the log prior, we get

$$|\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x}) - \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})| \leq \frac{1}{\gamma^2} K |n'(\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})) + O(\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x}) - \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}))| + O(1/\gamma^3),$$

which implies

$$\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x}) = \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) + O(1/\gamma^2)\,,$$

i.e., we obtain an improved convergence rate. Using this rate and Taylor theorem, we obtain the final equation by replacing the variational posterior mean by the inverse decoder in (C.20)

$$\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x}) = \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) + \frac{1}{\gamma^2} \mathbf{J}^{-1}_{\mathbf{f}^{\boldsymbol{\theta}}|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})} \mathbf{J}^{-T}_{\mathbf{f}^{\boldsymbol{\theta}}|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})} n'(\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})) + O(1/\gamma^3)$$

□

C.2.2 Proof of Theorem 4.3.2

This will be a corollary of the following result, that uses as a key assumption a rate of $O(1/\gamma^2)$ in the convergence of the self-consistency equation of the variational mean.

Proposition C.2.4 (VAEs with log-concave factorized prior and close–to-deterministic decoder approximate the IMA objective) *Under Assumption 4.3.1, if additionally the VAE satisfies the following self-consistency in the deterministic limit*

$$\left\| \boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x}) - \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) \right\| = O_{\gamma \to +\infty}(1/\gamma^2)\,, \tag{C.21}$$

$$\left\| \boldsymbol{\sigma}^{\widehat{\phi}}(\mathbf{x})^2 \right\|^2 = O_{\gamma \to +\infty}(1/\gamma^2)\,. \tag{C.22}$$

then

$$\sigma_k^{\widehat{\phi}}(\mathbf{x})^2 = \left(-\frac{n^2 \log p_0}{ds_k^2}(g_k^{\boldsymbol{\theta}}(\mathbf{x})) + \gamma^2 \left\| \left[\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}} \left(\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) \right) \right]_{:k} \right\|^2 \right)^{-1} + O(1/\gamma^3)\,, \tag{C.23}$$

and the self-consistent ELBO (4.7) *approximates the* IMA*-regularized log-likelihood* (3.5)*:*

$$\mathrm{ELBO}^*(\mathbf{x};\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\mathbf{x}) - c_{\mathrm{IMA}}(\mathbf{f}^{\boldsymbol{\theta}}, \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})) + O_{\gamma \to \infty}(1/\gamma^2)\,. \tag{C.24}$$

Proof. We start from the self-consistent ELBO decomposition as "reconstruction error plus posterior regularization" terms:

$$\mathrm{ELBO}^*(\mathbf{x};\boldsymbol{\theta}) = -\mathrm{KL}\left[q_{\widehat{\phi}}(\mathbf{s}|\mathbf{x}) || p_{\mathbf{s}}(\mathbf{s}) \right] + \mathbb{E}_{q_{\widehat{\phi}}(\mathbf{s}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s}) \right], \tag{C.25}$$

and continue with reformulating both terms, based on Assumption 4.3.1. That is, p_0 is factorized with components i.i.d. distributed according to a fully supported **log-concave** density $s_k \sim d$.

Posterior regularization term Assumption 4.3.1 gives us the formula of (C.1) for this term in the ELBO. Taking optimal encoder parameters, we get the posterior regularization term for the ELBO*

$$-\mathrm{KL}\left[q_{\widehat{\phi}}(\mathbf{s}|\mathbf{x}) || p_{\mathbf{s}}(\mathbf{s}) \right] = \mathbb{E}_{q_{\widehat{\phi}}(\mathbf{s}|\mathbf{x})}[\log(p_{\mathbf{s}}(\mathbf{s}))] + \frac{1}{2}\sum_{k=1}^{n}\left[\log \sigma_k^{\widehat{\phi}}(\mathbf{x})^2 \right] + \kappa_n\,,$$

with $\kappa_n = \frac{n}{2}(\log(2\pi) + 1)$. Using the factorized Gaussian encoder and i.i.d. prior assumptions we get

$$-\mathrm{KL}\left[q_{\widehat{\phi}}(\mathbf{s}|\mathbf{x}) || p_{\mathbf{s}}(\mathbf{s}) \right] = \sum_{k=1}^{n} \mathbb{E}_{s_k \sim \mathcal{N}(\mu_k^{\widehat{\phi}}(\mathbf{x}), \sigma_k^{\widehat{\phi}}(\mathbf{x})^2)}[\log(d(s_k))] + \frac{1}{2}\sum_{k=1}^{n}\left[\log \sigma_k^{\widehat{\phi}}(\mathbf{x})^2 \right] + \kappa_n\,,$$

where we rewrote the distribution p_0 as $p_0 = \prod_k m(s_k)$.

Based on the Taylor theorem, with a residual in Lagrange form of $n = \log d$, we have that for all k and u there exists $\xi \in [\mu_k^{\phi}(\mathbf{x}), u]$ if

$u \geq \mu_k^{\phi}(\mathbf{x})$, or $\xi \in [u, \mu_k^{\phi}(\mathbf{x})]$ if $u \leq \mu_k^{\phi}(\mathbf{x})$ such that

$$n(u) = \log(d(u)) = \log(d(\mu_k^{\widehat{\phi}}(\mathbf{x}))) + n'(\mu_k^{\widehat{\phi}}(\mathbf{x}))(u - \mu_k^{\widehat{\phi}}(\mathbf{x})) + \frac{1}{2}n''(\mu_k^{\widehat{\phi}}(\mathbf{x}))(u - \mu_k^{\widehat{\phi}}(\mathbf{x}))^2 + \frac{1}{3!}n^{(3)}(\xi)(u - \mu_k^{\widehat{\phi}}(\mathbf{x}))^3$$

We assumed that $|n^{(3)}|$ is bounded over $\mathbb{R}$ by F, such that

$$-F\left|u - \mu_k^{\widehat{\phi}}(\mathbf{x})\right|^3 \leq \log(d(u)) - \log(d(\mu_k^{\widehat{\phi}}(\mathbf{x}))) - n'(\mu_k^{\widehat{\phi}}(\mathbf{x}))(u - \mu_k^{\widehat{\phi}}(\mathbf{x})) - \frac{1}{2}n''(\mu_k^{\widehat{\phi}}(\mathbf{x}))(u - \mu_k^{\widehat{\phi}}(\mathbf{x}))^2 \leq F\left|u - \mu_k^{\widehat{\phi}}(\mathbf{x})\right|^3 .$$

Taking the expectation and using the expression of centered Gaussian absolute moments[2]

2: see e.g. https://arxiv.org/pdf/1209.4340

$$\left|\mathbb{E}_{s_k \sim \mathcal{N}(\mu_k^{\widehat{\phi}}(\mathbf{x}), \sigma_k^{\widehat{\phi}}(\mathbf{x})^2)}[\log(d(s_k))] - \log(d(\mu_k^{\widehat{\phi}}(\mathbf{x}))) - \frac{1}{2}n''(\mu_k^{\widehat{\phi}}(\mathbf{x}))\sigma_k^{\widehat{\phi}}(\mathbf{x})^2\right| \leq F\mathbb{E}\left[\left|u - \mu_k^{\widehat{\phi}}(\mathbf{x})\right|^3\right] = F\sigma_k^{\widehat{\phi}}(\mathbf{x})^3\frac{2^{3/2}}{\sqrt{\pi}} . \quad \text{(C.26)}$$

As the assumptions entail that optimal posterior variances $\sigma_k^{\widehat{\phi}}(\mathbf{x})^2$ get small for γ large (cf. (C.22)), this implies the near-deterministic approximation

$$\mathbb{E}_{s_k \sim \mathcal{N}(\mu_k^{\widehat{\phi}}(\mathbf{x}), \sigma_k(\mathbf{x})^2)}[\log(d(s_k))] = \log(d(\mu_k^{\widehat{\phi}}(\mathbf{x}))) + \frac{1}{2}n''(\mu_k^{\widehat{\phi}}(\mathbf{x}))\sigma_k^{\widehat{\phi}}(\mathbf{x})^2 + O_{\gamma \to +\infty}(1/\gamma^3) .$$

In addition, using again a Taylor formula and the self-consistency assumption for the mean

$$\log(d(\mu_k^{\widehat{\phi}}(\mathbf{x}))) = \log(d(g_k^{\boldsymbol{\theta}}(\mathbf{x}))) + n'(g_k^{\boldsymbol{\theta}}(\mathbf{x}))(\mu_k^{\widehat{\phi}}(\mathbf{x}) - g_k^{\boldsymbol{\theta}}(\mathbf{x})) + O_{\gamma \to +\infty}(1/\gamma^2) = \log(d(g_k^{\boldsymbol{\theta}}(\mathbf{x}))) + O_{\gamma \to +\infty}(1/\gamma^2).$$

Moreover, using again a Taylor formula for n'' under boundedness of $n^{(3)}$ and again using the self-consistency assumption for the mean yields

$$n''(\mu_k^{\widehat{\phi}}(\mathbf{x})) = n''(g_k^{\boldsymbol{\theta}}(\mathbf{x})) + O(\mu_k^{\widehat{\phi}}(\mathbf{x}) - g_k^{\boldsymbol{\theta}}(\mathbf{x})) = n''(g_k^{\boldsymbol{\theta}}(\mathbf{x})) + O_{\gamma \to +\infty}(1/\gamma^2) .$$

Overall this leads to the approximation of the posterior regularization term

$$-\mathrm{KL}\left[q_{\widehat{\phi}}(\mathbf{s}|\mathbf{x})||p_{\mathbf{s}}(\mathbf{s})\right] = \sum_{k=1}^{n} \log(d(g_k^{\boldsymbol{\theta}}(\mathbf{x}))) + \frac{1}{2}n''(g_k^{\boldsymbol{\theta}}(\mathbf{x}))\sigma_k^{\widehat{\phi}}(\mathbf{x})^2 + \frac{1}{2}\log\sigma_k^{\widehat{\phi}}(\mathbf{x})^2 + \kappa_n + O_{\gamma \to +\infty}(1/\gamma^2) . \quad \text{(C.27)}$$

Reconstruction term Now switching to the first (reconstruction) term of the ELBO*, adapting the decomposition of (C.2) by using optimal

encoder parameters we get

$$\mathbb{E}_{q_{\widehat{\phi}}(\mathbf{s}|\mathbf{x})}\left[\log p_{\theta}(\mathbf{x}|\mathbf{s})\right] = -\frac{\gamma^2}{2}\mathbb{E}_{q_{\widehat{\phi}}(\mathbf{s}|\mathbf{x})}\left[\|\mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})\|^2\right] + n\log\gamma - \frac{n}{2}\log(2\pi).$$

Then in the small encoder noise limit $\sigma_k(\mathbf{x})^2 \ll 1, \forall k$ (justified by Proposition 4.3.1), we rely on a Taylor approximation around the posterior mean $\mathbf{s}^o = \boldsymbol{\mu}^{\phi}(\mathbf{x})$ based on Lemma C.3.3, which bounds this approximation as follows

$$\mathbb{E}_{q_{\phi}(\mathbf{s}|\mathbf{x})}\left[\left\|\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}) - \mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x})) - \sum_{k=1}^{n}\frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial s_k}_{|\mathbf{s}^o}(s_k - \mu_k^{\widehat{\phi}}(\mathbf{x}))\right\|^2\right] \leq \frac{n^3}{4}3K^2\sum_i \sigma_i^{\widehat{\phi}}(\mathbf{x})^4\,. \tag{C.28}$$

The linear term in this approximation is easily computed using successively Lemma C.3.1 and Lemma C.3.2 to get an expression with the squared column norms of the partial derivatives scaled by the standard deviations $\frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\phi}(\mathbf{x})}$. We get

$$\begin{aligned}\mathbb{E}_{q_{\phi}(\mathbf{s}|\mathbf{x})}\left[\left\|\sum_{k=1}^{n}\frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial s_k}_{|z^o}(s_k - \mu_k^{\phi}(\mathbf{x}))\right\|^2\right] &= \operatorname{trace}\left[\operatorname{Cov}\left[\sum_{k=1}^{n}\frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\phi}(\mathbf{x})}(s_k - \mu_k^{\phi}(\mathbf{x}))\right]\right] \\ &= \sum_{k=1}^{n}\left[\left\|\frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\phi}(\mathbf{x})}\right\|^2 \sigma_k^{\phi}(\mathbf{x})^2\right]. \quad \text{(C.29)}\end{aligned}$$

This term can be used as an approximation for the expectation term in the reconstruction loss thanks to the following reverse triangle inequality

$$\begin{aligned}&\left|\mathbb{E}_{q_{\phi}(\mathbf{s}|\mathbf{x})}\left[\|\mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})\|^2\right] - \mathbb{E}_{q_{\phi}(\mathbf{s}|\mathbf{x})}\left[\left\|\sum_{k=1}^{n}\frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial s_k}_{|z^o}(s_k - \mu_k^{\phi}(\mathbf{x}))\right\|^2\right]\right| \\ &\quad= \left|\mathbb{E}_{q_{\phi}(\mathbf{s}|\mathbf{x})}\left[\|\mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})\|^2\right] - \sum_{k=1}^{n}\left[\left\|\frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\phi}(\mathbf{x})}\right\|^2 \sigma_k^{\phi}(\mathbf{x})^2\right]\right| \\ &\quad\leq \mathbb{E}_{q_{\phi}(\mathbf{s}|\mathbf{x})}\left[\left\|\mathbf{x} - \left(\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}) - \sum_{k=1}^{n}\frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial s_k}_{|z^o}(s_k - \mu_k^{\phi}(\mathbf{x}))\right)\right\|^2\right],\end{aligned}$$

such that the resulting upper bound can be itself bounded as follows

$$\begin{aligned}&\mathbb{E}_{q_{\phi}(\mathbf{s}|\mathbf{x})}\left[\left\|\mathbf{x} - \left(\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}) - \sum_{k=1}^{n}\frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial s_k}_{|\mathbf{s}^o}(s_k - \mu_k^{\phi}(\mathbf{x}))\right)\right\|^2\right] \\ &\leq \mathbb{E}_{q_{\phi}(\mathbf{s}|\mathbf{x})}\left[\left\|\mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\phi}(\mathbf{x}))\right\|^2\right] + \mathbb{E}_{q_{\phi}(\mathbf{s}|\mathbf{x})}\left[\left\|\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}) - \mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\phi}(\mathbf{x})) - \sum_{k=1}^{n}\frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial s_k}_{|\boldsymbol{\mu}^{\phi}(\mathbf{x})}(s_k - \mu_k^{\phi}(\mathbf{x}))\right\|^2\right].\end{aligned}$$

Each term of the upper bound can be bounded for the optimum encoder parameters: using from left to right the assumption of (C.21) and (C.28),

respectively, leading to

$$\left| \mathbb{E}_{q_{\widehat{\phi}}(\mathbf{s}|\mathbf{x})} \left[\|\mathbf{x} - \mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})\|^2 \right] - \sum_{k=1}^{n} \left[\left\| \frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\widehat{\phi}}(\mathbf{x})} \right\|^2 \sigma_k^{\widehat{\phi}}(\mathbf{x})^2 \right] \right| \leq O_{\gamma \to +\infty}(1/\gamma^4) + \frac{n^3}{4} 3K^2 \sum_i \sigma_i^{\widehat{\phi}}(\mathbf{x})^4 .$$

Getting back to the whole reconstruction term, using additionally the variance self-consistency assumption (C.22), the above shows that we can make the approximation

$$\mathbb{E}_{q_{\widehat{\phi}}(\mathbf{s}|\mathbf{x})} [\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})] = -\frac{\gamma^2}{2} \sum_{k=1}^{n} \left[\left\| \frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\widehat{\phi}}(\mathbf{x})} \right\|^2 \sigma_k^{\widehat{\phi}}(\mathbf{x})^2 \right] + n \log \gamma - \frac{n}{2} \log(2\pi) + O_{\gamma \to +\infty}(1/\gamma^2)$$

We can further replace the dependency of the derivatives on the encoder mean using a Taylor formula for the derivative

$$\frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu^{\widehat{\phi}}(\mathbf{x})} = \frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})} + O(\boldsymbol{\mu}^{\widehat{\phi}}(\mathbf{x}) - \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})) = \frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})} + O(1/\gamma^2)$$

such that

$$\mathbb{E}_{q_{\widehat{\phi}}(\mathbf{s}|\mathbf{x})} [\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{s})] = -\frac{\gamma^2}{2} \sum_{k=1}^{n} \left[\left\| \frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})} \right\|^2 \sigma_k^{\widehat{\phi}}(\mathbf{x})^2 \right] + n \log \gamma - \frac{n}{2} \log(2\pi) + O_{\gamma \to +\infty}(1/\gamma^2) \quad \text{(C.30)}$$

ELBO* approximation As a consequence of (C.27) and (C.30) the ELBO* becomes

$$\begin{aligned}
\text{ELBO}^*(\mathbf{x}; \boldsymbol{\theta}) &= -\frac{1}{2} \sum_{k=1}^{n} \left[\log \frac{1}{\sigma_k^{\widehat{\phi}}(\mathbf{x})^2} + \sigma_k^{\widehat{\phi}}(\mathbf{x})^2 \left(-n''(g^{\boldsymbol{\theta}}{}_k(\mathbf{x})) + \gamma^2 \left\| \frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|g^{\boldsymbol{\theta}}{}_k} \right\|^2 \right) \right. \\
&\quad \left. - 2 \log(d(g^{\boldsymbol{\theta}}{}_k(\mathbf{x}))) \right] + n \log \gamma + \kappa_n - \frac{n}{2} \log(2\pi) + O_{\gamma \to \infty}(1/\gamma^2) \\
&= -\frac{1}{2} \sum_{k=1}^{n} \left[\log \frac{1}{\sigma_k^{\widehat{\phi}}(\mathbf{x})^2} - 1 + \sigma_k^{\widehat{\phi}}(\mathbf{x})^2 \left(-n''(g^{\boldsymbol{\theta}}{}_k(\mathbf{x})) + \gamma^2 \left\| \frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})} \right\|^2 \right) \right. \\
&\quad \left. - 2 \log(d(g^{\boldsymbol{\theta}}{}_k(\mathbf{x}))) \right] + n \log \gamma + O_{\gamma \to \infty}(1/\gamma^2) \\
&= \widehat{\text{ELBO}}(\sigma^{\widehat{\phi}}(\mathbf{x})^2; \mathbf{x}, \boldsymbol{\theta}, \widehat{\boldsymbol{\phi}}) + \sum_{k=1}^{n} \log(d(g^{\boldsymbol{\theta}}{}_k(\mathbf{x}))) + O_{\gamma \to \infty}(1/\gamma^2) ,
\end{aligned}$$

where we isolated the terms that depend on parameters $\sigma_k^{\widehat{\phi}}(\mathbf{x})^2$ and γ in the approximate objective $\widehat{\text{ELBO}}(\boldsymbol{\sigma}^2 = \sigma^{\widehat{\phi}}(\mathbf{x})^2; \mathbf{x}, \boldsymbol{\theta}, \widehat{\boldsymbol{\phi}})$ that we define for

arbitrary $\boldsymbol{\sigma}^2$.

$$\widehat{\mathrm{ELBO}}(\boldsymbol{\sigma}^2;\mathbf{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}) = -\frac{1}{2}\sum_{k=1}^{n}\left[\log\frac{1}{\gamma^2\sigma_k^2} - 1 + \sigma_k^2\left(-n''({g^\theta}_k(\mathbf{x})) + \gamma^2\left\|\frac{\partial \mathbf{f}^\theta}{\partial z_k}_{|\mathbf{g}^\theta(\mathbf{x})}\right\|^2\right)\right]$$
$$= \sum_{k=1}^{n}\widehat{\mathrm{ELBO}}_k(\sigma_k^2;\mathbf{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}})$$

Where we further break this objective in n components $\widehat{\mathrm{ELBO}}_k(\sigma_k^{\widehat{\phi}}(\mathbf{x})^2;\mathbf{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}})$ according to the terms of the sum as follows

$$\widehat{\mathrm{ELBO}}_k(\sigma_k^2;\mathbf{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}) = -\frac{1}{2}\left[\log\frac{1}{\gamma^2\sigma_k^2} - 1 + \gamma^2\sigma_k^2\left(-\frac{1}{\gamma^2}n''({g^\theta}_k(\mathbf{x})) + \left\|\frac{\partial \mathbf{f}^\theta}{\partial z_k}_{|\mathbf{g}^\theta(\mathbf{x})}\right\|^2\right)\right]$$

and where we note that $-n'' \geq 0$ due to the log-concavity assumption.

Solving term in k $\widehat{\mathrm{ELBO}}_k(\sigma_k^2)$ for optimal $\gamma^2\sigma_k^*$ we get (see lemma C.3.4):

$$\gamma^2\sigma_k^{*2} = \left(-\frac{1}{\gamma^2}n''({g^\theta}_k(\mathbf{x})) + \left\|\frac{\partial \mathbf{f}^\theta}{\partial z_k}_{|{g^\theta}_k(\mathbf{x})}\right\|^2\right)^{-1} \tag{C.31}$$

and the resulting optimal value $\widehat{\mathrm{ELBO}}_k^*(\mathbf{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}) = \widehat{\mathrm{ELBO}}_k(\sigma_k^{*2};\mathbf{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}})$ is

$$\widehat{\mathrm{ELBO}}_k^*(\mathbf{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}})^* = -\frac{1}{2}\log\left(-\frac{1}{\gamma^2}n''({g^\theta}_k(\mathbf{x})) + \left\|\frac{\partial \mathbf{f}^\theta}{\partial z_k}_{|{g^\theta}_k(\mathbf{x})}\right\|^2\right)$$

A Taylor formula around this optimum leads, for some value $\xi_\gamma(\mathbf{x})$ lying between σ_k^{*2} and σ_k^2 to (note the first order derivative vanishes, and the second order derivative is upper bounded hence the second line)

$$\widehat{\mathrm{ELBO}}_k(\sigma_k^2;\mathbf{x},\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}) = \widehat{\mathrm{ELBO}}_k^*(\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}) + \frac{n\widehat{\mathrm{ELBO}}_k(\mathbf{x};\boldsymbol{\theta},\widehat{\boldsymbol{\phi}})}{d\gamma^2\sigma_k^2}_{|\sigma_k^{*2}}(\gamma^2\sigma_k^2 - \gamma^2\sigma_k^{*2})$$
$$+ \frac{n^2\widehat{\mathrm{ELBO}}_k(\mathbf{x};\boldsymbol{\theta},\widehat{\boldsymbol{\phi}})}{d(\gamma^2\sigma_k^2)^2}_{|\xi_\gamma(\mathbf{x})}(\gamma^2\sigma_k^2 - \gamma^2\sigma_k^{*2})^2$$
$$\leq \widehat{\mathrm{ELBO}}_k^*(\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}) - \frac{1}{2}\left\|\frac{\partial \mathbf{f}^\theta}{\partial z_k}_{|\mathbf{g}^\theta(\mathbf{x})}\right\|^2(\gamma^2\sigma_k^2 - \gamma^2\sigma_k^{*2})^2$$

as a consequence the non-approximate solution for the true optimal ELBO^*, as γ grows, must achieve a value below this quadratic function, up to a term in $O(1/\gamma^2)$, and at the same time above $\widehat{\mathrm{ELBO}}^*$, also up to a term in $O(1/\gamma^2)$. This entails that it is restricted to a smaller and smaller domain near the approximate solution and we get

$$\sigma_k^{\widehat{\phi}}(\mathbf{x})^2 = \sigma_k^{*2} + O(1/\gamma^3) = \left(-n''({g^\theta}_k(\mathbf{x})) + \gamma^2\left\|\frac{\partial \mathbf{f}^\theta}{\partial z_k}_{|{g^\theta}_k(\mathbf{x})}\right\|^2\right)^{-1} + O(1/\gamma^3). \tag{C.32}$$

Leading to the approximation of the true objective

$$\text{ELBO}^*(\mathbf{x};\boldsymbol{\theta}) = -\frac{1}{2}\sum_{k=1}^{n}\left[\log\left(-\frac{1}{\gamma^2}n''(\mu_k^{\boldsymbol{\phi}}(\mathbf{x})) + \left\|\frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial z_k}_{|\mu_k^{\boldsymbol{\phi}}(\mathbf{x})}\right\|^2\right) - 2\log(d(\mu_k^{\boldsymbol{\phi}}(\mathbf{x})))\right] + O(1/\gamma^2),$$

which reduces to

$$\text{ELBO}^*(\mathbf{x};\boldsymbol{\theta}) = \log p_0(\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})) - \frac{1}{2}\sum_{k=1}^{n}\left[\log\left\|\left[\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}}(\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}))\right]_{:k}\right\|^2\right] + O(1/\gamma^2),$$

which is the IMA objective.

□

We now restate the main text theorem and provide its proof.

Theorem 4.3.2 *[VAEs with a near-deterministic decoder approximate the* IMA *objective] Under Assumption 4.3.1, the variational posterior satisfies*

$$\sigma_k^{\widehat{\boldsymbol{\phi}}}(\mathbf{x})^2 = \left(-\frac{d^2 \log p_0}{ds_k^2}(g_k^{\boldsymbol{\theta}}(\mathbf{x})) + \gamma^2\left\|\left[\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}}\left(\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})\right)\right]_{:k}\right\|^2\right)^{-1} + O(1/\gamma^3)\,, \tag{4.11}$$

and the self-consistent ELBO (4.7) *approximates the* IMA*-regularized log-likelihood* (3.5)*:*

$$\text{ELBO}^*(\mathbf{x};\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\mathbf{x}) - c_{\text{IMA}}(\mathbf{f}^{\boldsymbol{\theta}}, \mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})) + O_{\gamma\to\infty}(1/\gamma^2)\,. \tag{4.12}$$

Proof. This is just a corollary of Proposition C.2.4 because Proposition C.2.3 entails through (C.19) the required $O(1/\gamma^2)$ rate of convergence for the optimal variational mean in (C.21), while (C.22) is fulfilled through Proposition 4.3.1. □

C.3 Auxiliary results

C.3.1 Squared norm statistics

Lemma C.3.1 (Squared norm variance decomposition) *For multivariate RV X with mean m*

$$\mathbb{E}\left[\|X\|^2\right] = \mathit{trace}\left[\mathit{Cov}(X)\right] + \|m\|^2$$

Proof.
$\mathbb{E}\|X-m\|^2 = \mathbb{E}\,\langle X-m,\, X-m\rangle = \mathbb{E}\left[\langle X,\, X\rangle - 2\mathbb{E}\,\langle m,\, X\rangle + \langle m,\, m\rangle\right]$

hence

$$\mathbb{E}\|X-m\|^2 = \mathbb{E}\left[\|X\|^2\right] - \|m\|^2$$

This leads to (using that the trace of a scalar is the scalar itself)

$$\mathbb{E}\left[\|X\|^2\right] = \mathbb{E}\left[\text{trace}\left[\|X-m\|^2\right]\right] + \|m\|^2 = \text{trace}\left[\mathbb{E}\left[(X-m)^T(X-m)\right]\right] + \|m\|^2$$

because trace$[AB]$ = trace$[BA]$ we get

$$\mathbb{E}\left[\|X\|^2\right] = \text{trace}\left[\mathbb{E}\left[(X-m)(X-m)^T\right]\right] + \|m\|^2 = \text{trace}\left[\text{Cov}(X)\right] + \|m\|^2$$

□

Lemma C.3.2 (Trace of transformed unit covariance) *When the covariance matrix $Cov(\boldsymbol{\epsilon})$ is the identity, then*

$$trace[Cov(A\boldsymbol{\epsilon})] = \sum_k \|[A]_{.k}\|^2 ,$$

Proof. For arbitrary matrix A, $Cov(A\boldsymbol{\epsilon}) = ACov(\boldsymbol{\epsilon})A^T$ and thus

$$\text{trace}[Cov(A\boldsymbol{\epsilon})] = \text{trace}[ACov(\boldsymbol{\epsilon})A^T] = \text{trace}[A^T ACov(\boldsymbol{\epsilon})] .$$

Moreover, in our case $Cov(\boldsymbol{\epsilon})$ is the identity such that

$$\text{trace}[Cov(A\boldsymbol{\epsilon})] = \text{trace}[A^T A] = \sum_k \|[A]_{.k}\|^2 ,$$

□

C.3.2 KL divergence bounds

Proposition C.3.1 (Lipschtiz continuity-based lower bound) *Assume $\mathbf{g}^{\boldsymbol{\theta}}$ is Lipschitz continuous with Lipschitz constant $B > 0$, in the sense*

$$\forall \mathbf{x}, \boldsymbol{y} \in \mathbb{R}^n, \left\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) - \mathbf{g}^{\boldsymbol{\theta}}(\boldsymbol{y})\right\|_2 \leq B\|\mathbf{x}-\boldsymbol{y}\|_2 .$$

Then for any encoder parameter choice

$$KL\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})\right] \geq -\sum_{k=1}^{n}\left[\log\sigma_k^{\boldsymbol{\phi}}(\mathbf{x}) + \log d(\mu_k^{\boldsymbol{\phi}})\right] + c(\mathbf{x},\gamma) + \frac{\gamma^2}{2}B^{-2}\left[\left\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\mathbf{x})\right\|^2 + \sum_{k=1}^{n}\sigma_k^{\boldsymbol{\phi}}(\mathbf{x})^2\right] , \quad \text{(C.33)}$$

with $c(\mathbf{x},\gamma) = -\frac{n}{2}\left(\log(\gamma^2)+1\right) + \log p_{\boldsymbol{\theta}}(\mathbf{x})$.

Proof. Starting from the KL divergence expression (C.3),

$$\text{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})\right] = -\sum_{k=1}^{n}\log\sigma_k^{\boldsymbol{\phi}}(\mathbf{x}) + \frac{1}{2}\mathbb{E}_{\mathbf{s}\sim q_{\boldsymbol{\phi}}}\left[\|\mathbf{x}-\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s})\|^2\gamma^2 - 2\sum_{k=1}^{n}\log d(s_k)\right] + c(\mathbf{x},\gamma)$$

with additive constant $c(\mathbf{x},\gamma) = -\frac{n}{2}\left(\log(\gamma^2)+1\right) + \log p_{\boldsymbol{\theta}}(\mathbf{x})$. By Lipschitz continuity

$$\text{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})\right] \geq -\sum_{k=1}^{n}\log\sigma_k^{\boldsymbol{\phi}}(\mathbf{x}) + \frac{1}{2}\mathbb{E}_{\mathbf{s}\sim q_{\boldsymbol{\phi}}}\left[B^{-2}\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) - \mathbf{s}\|^2\gamma^2 - 2\sum_{k=1}^{n}\log d(s_k)\right] + c(\mathbf{x},\gamma) .$$

using Lemma C.3.1 applied to $\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) - \mathbf{s}$, $\mathbf{s} \sim q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})$ we get

$$\begin{aligned}
\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})\right] &\geq -\sum_{k=1}^{n} \log \sigma_k^{\boldsymbol{\phi}}(\mathbf{x}) + \frac{\gamma^2}{2} B^{-2} \left[\left\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\mathbf{x})\right\|^2 + \mathrm{trace}\left[Cov\left[\mathbf{s}\right]\right]\right] \\
&\quad - \mathbb{E}_{\mathbf{s}\sim q_{\boldsymbol{\phi}}}\left[\sum_{k=1}^{n} \log d(s_k)\right] + c(\mathbf{x}, \gamma) \\
&\geq -\sum_{k=1}^{n} \log \sigma_k^{\boldsymbol{\phi}}(\mathbf{x}) + \frac{\gamma^2}{2} B^{-2} \left[\left\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\mathbf{x})\right\|^2 + \sum_{k=1}^{n} \sigma_k^{\boldsymbol{\phi}}(\mathbf{x})^2\right] \\
&\quad - \mathbb{E}_{\mathbf{s}\sim q_{\boldsymbol{\phi}}}\left[\sum_{k=1}^{n} \log d(s_k)\right] + c(\mathbf{x}, \gamma)\,.
\end{aligned}$$

Using Jensen's inequality for $-\log d$ (convex by Assum. 4.3.1(i)), we get

$$\begin{aligned}
\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})\right] &\geq -\sum_{k=1}^{n} \left[\log \sigma_k^{\boldsymbol{\phi}}(\mathbf{x})\right] + \frac{\gamma^2}{2} B^{-2} \left[\left\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\mathbf{x})\right\|^2 + \sum_{k=1}^{n} \sigma_k^{\boldsymbol{\phi}}(\mathbf{x})^2\right] \\
&\quad - \sum_{k=1}^{n} \left[\log d(\mu_k^{\boldsymbol{\phi}})\right] + c(\mathbf{x}, \gamma)
\end{aligned}$$

by reordering the terms we finally get

$$\begin{aligned}
\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\mathbf{s}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})\right] &\geq -\sum_{k=1}^{n} \left[\log \sigma_k^{\boldsymbol{\phi}}(\mathbf{x}) + \log d(\mu_k^{\boldsymbol{\phi}})\right] + c(\mathbf{x}, \gamma) \\
&\quad + \frac{\gamma^2}{2} B^{-2} \left[\left\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\mu}^{\boldsymbol{\phi}}(\mathbf{x})\right\|^2 + \sum_{k=1}^{n} \sigma_k^{\boldsymbol{\phi}}(\mathbf{x})^2\right]
\end{aligned}$$

which is the stated KL lower bound. □

Proposition C.3.2 (Optimal encoder KL divergence upper bound) *Assume $\mathbf{f}^{\boldsymbol{\theta}}$ is Lipschitz continuous with Lipschitz constant $C > 0$, in the sense that*

$$\forall \mathbf{s}, \boldsymbol{w} \in \mathbb{R}^n : \qquad \left\|\mathbf{f}^{\boldsymbol{\theta}}(\mathbf{s}) - \mathbf{f}^{\boldsymbol{\theta}}(\boldsymbol{w})\right\|_2 \leq C\|\mathbf{s} - \boldsymbol{w}\|_2\,.$$

Assume, $-\log d$ is quadratically dominated, in the sense that

$$\exists D > 0, E > 0, \forall u \in \mathbb{R}, -\log d(u) \leq D|u|^2 + E\,.$$

Then for the optimal encoder solution of (4.6)

$$KL\left[q_{\widehat{\boldsymbol{\phi}}}(\mathbf{s}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})\right] \leq n\left(\frac{1}{2}C^2 + E + D\left[\frac{\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})\|^2}{n} + \frac{1}{\gamma^2}\right]\right) - \frac{n}{2} + \log p_{\boldsymbol{\theta}}(\mathbf{x})\,, \quad \text{(C.34)}$$

and

$$\limsup_{\gamma \to +\infty} KL\left[q_{\widehat{\boldsymbol{\phi}}}(\mathbf{s}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{s}|\mathbf{x})\right] \leq n\left(\frac{1}{2}C^2 + E\right) + D\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})\|^2 - \frac{n}{2} - \log|\mathbf{J}_{\mathbf{f}^{\boldsymbol{\theta}}}(\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}))| + \log(p_0(\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}))) \quad \text{(C.35)}$$

Proof. Starting from the KL divergence expression (C.3),

$$\mathrm{KL}\left[q_{\phi}(\mathbf{s}|\mathbf{x})||p_{\theta}(\mathbf{s}|\mathbf{x})\right] = -\sum_{k=1}^{n}\log\sigma_k^{\phi}(\mathbf{x}) + \frac{1}{2}\mathbb{E}_{\mathbf{s}\sim q_{\phi}}\left[\|\mathbf{x}-\mathbf{f}^{\theta}(\mathbf{s})\|^2\gamma^2 - 2\sum_{k=1}^{n}\log d(s_k)\right] + c(\mathbf{x},\gamma)$$

with additive constant $c(\mathbf{x},\gamma) = -\frac{n}{2}\left(\log(\gamma^2)+1\right) + \log p_{\theta}(\mathbf{x})$.

Let us choose the following posterior (by universal approximation capabilities of the encoder):

$$\boldsymbol{\mu}^{\phi^*}(\mathbf{x}) = \mathbf{g}^{\theta}(\mathbf{x}) \tag{C.36}$$

$$\boldsymbol{\sigma}^{\phi^*}(\mathbf{x}) = \frac{1}{\gamma} \tag{C.37}$$

Using Lipschitz continuity we get

$$\mathrm{KL}\left[q_{\phi^*}(\mathbf{s}|\mathbf{x})||p_{\theta}(\mathbf{s}|\mathbf{x})\right] \leq -\sum_{k=1}^{n}\log\sigma_k^{\phi^*}(\mathbf{x}) + \frac{1}{2}\mathbb{E}_{\mathbf{s}\sim q_{\phi^*}}\left[C^2\|\boldsymbol{\mu}^{\phi^*}(\mathbf{x})-\mathbf{s}\|^2\gamma^2 - 2\sum_{k=1}^{n}\log d(s_k)\right] + c(\mathbf{x},\gamma)$$

then, using

$$\mathbb{E}_{\mathbf{s}\sim q_{\phi^*}}\left[\|\boldsymbol{\mu}^{\phi^*}(\mathbf{x})-\mathbf{s}\|^2\right] = \sum_{k=1}^{n}\mathbb{E}_{s_k\sim\mathcal{N}(\mu_k^{\phi^*}(\mathbf{x}),\sigma_k^{\phi^*}(\mathbf{x})^2)}\left[|\mu_k^{\phi^*}(\mathbf{x})-s_k|^2\right] = \sum_{k=1}^{n}\sigma_k^{\phi^*}(\mathbf{x})^2\,,$$

we get

$$\mathrm{KL}\left[q_{\phi^*}(\mathbf{s}|\mathbf{x})||p_{\theta}(\mathbf{s}|\mathbf{x})\right] \leq \sum_{k=1}^{n}\left(-\log\sigma_k^{\phi^*}(\mathbf{x}) + \frac{1}{2}C^2\sigma_k^{\phi^*}(\mathbf{x})^2\gamma^2\right) - \mathbb{E}_{\mathbf{s}\sim q_{\phi^*}}\left[\sum_{k=1}^{n}\log d(s_k)\right] + c(\mathbf{x},\gamma)$$

using quadratic domination

$$\begin{aligned}\mathrm{KL}\left[q_{\phi^*}(\mathbf{s}|\mathbf{x})||p_{\theta}(\mathbf{s}|\mathbf{x})\right] &\leq \sum_{k=1}^{n}\left(-\log\sigma_k^{\phi^*}(\mathbf{x}) + \frac{1}{2}C^2\sigma_k^{\phi^*}(\mathbf{x})^2\gamma^2\right) \\ &\quad + \mathbb{E}_{\mathbf{s}\sim q_{\phi^*}}\left[nE + \sum_{k=1}^{n}D|s_k|^2\right] + c(\mathbf{x},\gamma) \\ &\leq \sum_{k=1}^{n}\left(-\log\sigma_k^{\phi^*}(\mathbf{x}) + \frac{1}{2}C^2\sigma_k^{\phi^*}(\mathbf{x})^2\gamma^2\right) \\ &\quad + nE + D\mathbb{E}_{\mathbf{s}\sim q_{\phi^*}}\left[|s_k|^2\right] + c(\mathbf{x},\gamma)\end{aligned}$$

Using Lemma C.3.1 we get

$$\begin{aligned}\mathrm{KL}\left[(q_{\phi^*}(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] &\leq \sum_{k=1}^{n}\left(-\log\sigma_k^{\phi^*}(\mathbf{x})+\frac{1}{2}C^2\sigma_k^{\phi^*}(\mathbf{x})^2\gamma^2\right)\\ &\quad + nE + D\left[\|\boldsymbol{\mu}^{\phi^*}(\mathbf{x})\|^2+\|\boldsymbol{\sigma}^{\phi^*}(\mathbf{x})\|^2\right]+c(\mathbf{x},\gamma)\\ \leq n\left(\log\gamma+\frac{1}{2}C^2\right)&+nE+D\left[\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})\|^2+\frac{n}{\gamma^2}\right]-\frac{n}{2}\left(\log(\gamma^2)+1\right)+\log p_\theta(\mathbf{x})\end{aligned}$$

hence for a parameter $\widehat{\boldsymbol{\phi}}$ achieving the minimum divergence we get

$$\begin{aligned}\mathrm{KL}\left[q_{\widehat{\boldsymbol{\phi}}}(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] &\leq \mathrm{KL}\left[q_{\phi^*}(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] \leq n\left(\log\gamma+\frac{1}{2}C^2\right)\\ &\quad + nE + D\left[\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})\|^2+\frac{n}{\gamma^2}\right]-\frac{n}{2}\left(\log(\gamma^2)+1\right)+\log p_\theta(\mathbf{x})\\ &\leq n\left(\frac{1}{2}C^2+E+D\left[\frac{\|\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x})\|^2}{n}+\frac{1}{\gamma^2}\right]\right)-\frac{n}{2}+\log p_\theta(\mathbf{x})\end{aligned}$$

As $\gamma\to+\infty$, $\log p_\theta(\mathbf{x})\to|\mathbf{J}_{\mathbf{f}^\theta}(\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}))|^{-1}p_0(\mathbf{g}^{\boldsymbol{\theta}}(\mathbf{x}))$ such that the KL divergence for the optimal solutions is upper bounded by a finite number.

□

C.3.3 Taylor formula-based approximations

Lemma C.3.3 (Bound on expectation of multivariate Taylor expansion) *Assume $\boldsymbol{f}:\mathbb{R}^n\to\mathbb{R}$ is C^2 and assume $\mathbf{s}$ is a multivariate RV on $\mathbb{R}^n$ with indepedent Gaussian components such that*

$$z_k\sim\mathcal{N}(\mu_k^{\phi}(\mathbf{x}),\sigma_k^{\phi}(\mathbf{x})^2)$$

then for all $\mathbf{s}_o\in\mathbb{R}^n$

$$\mathbb{E}_{\mathbf{s}}\left[\left\|\boldsymbol{f}(\mathbf{s})-\boldsymbol{f}(\mathbf{s}_o)-\sum_k\frac{\partial\boldsymbol{f}}{\partial z_k}_{|\mathbf{s}_o}(z_k-z_k^o)\right\|^2\right]\leq\frac{n^3}{4}3K^2\sum_i\left(\sigma_i^{\phi}\right)^4 \tag{C.38}$$

Proof. As described in [501, p. 162], for the l-th component of the function

$$\begin{aligned}f_l(\mathbf{s}) &= f_l(\mathbf{s}_o)+\sum_k\frac{\partial f_l}{\partial z_k}_{|\mathbf{s}_o}(z_k-z_k^o)+\frac{1}{2!}\sum_{i,j}\frac{\partial f_l}{\partial z_i\partial z_j}_{|\mathbf{s}_o+t_{ij}(\mathbf{s}-\mathbf{s}_o)}(z_i-z_i^o)(z_j-z_j^o)\,,\ t_{ij}\in(0;1)\ .\\ &= f_l(\mathbf{s}_o)+\sum_k\frac{\partial f_l}{\partial z_k}_{|\mathbf{s}_o}(z_k-z_k^o)+\frac{1}{2!}\sum_{i,j}(\mathbf{s}-\mathbf{s}_o)^T\mathscr{H}_k(\mathbf{s}-\mathbf{s}_o)\,,\end{aligned} \tag{C.39}$$

where the second line puts 1/2 of the partial derivatives in matrix form (note it is not exactly the Hessian as derivatives are taken at different

points). As a consequence

$$\left(f_l(\mathbf{s}) - f_l(\mathbf{s}_o) - \sum_k \frac{\partial f_l}{\partial z_k}_{|\mathbf{s}_o} (z_k - z_k^o)\right)^2 = \left((\mathbf{s}-\mathbf{s}_o)^T \mathscr{H}_k (\mathbf{s}-\mathbf{s}_o)\right)^2,$$
$$\leq \|\mathscr{H}_k\|_2^2 \|\mathbf{s}-\mathbf{s}_o\|^4$$
$$\leq \|\mathscr{H}_k\|_F^2 \|\mathbf{s}-\mathbf{s}_o\|^4$$

where $\|\mathscr{H}_k\|_2$ is the spectral norm of the matrix and $\|\mathscr{H}_k\|_F$ is the Frobenious norm [3] leading to the bound

$$\left(f_l(\mathbf{s}) - f_l(\mathbf{s}_o) - \sum_k \frac{\partial f_l}{\partial z_k}_{|\mathbf{s}_o} (z_k - z_k^o)\right)^2 \leq \frac{n^2}{4} K^2 \|\mathbf{s}-\mathbf{s}_o\|^4,$$

3: first inequality comes from Cauchy-Schwartz: $< x, Ax > \leq \|x\|\|Ax\| \leq \|x\|\|A\|_2\|x\|$, second is a classical inequality between norms

where K is an upper bound on the absolute second order derivatives. We have $(z_k - z_k^o) = \sigma_k^{\phi}(x)\epsilon_k$, with ϵ multivariate normal, so taking the expectation of the above simplifies to:

$$\mathbb{E}_{\mathbf{Z}}\left(f_l(\mathbf{s}) - f_l(\mathbf{s}_o) - \sum_k \frac{\partial f_l}{\partial z_k}_{|\mathbf{s}_o} (z_k - z_k^o)\right)^2 \leq \frac{n^2}{4} K^2 \mathbb{E}_{\mathbf{Z}} \|\mathbf{s}-\mathbf{s}_o\|^4,$$
$$= \frac{n^2}{4} K^2 \mathbb{E}_{\mathbf{Z}} \sum_{i,j} \left\|z_i - z_j^o\right\|^2 \left\|z_i - z_j^o\right\|^2$$
$$= \frac{n^2}{4} K^2 \sum_i \mathbb{E}_{\mathbf{Z}} \left\|z_i - z_i^o\right\|^4$$
$$= \frac{n^2}{4} 3K^2 \sum_i \left(\sigma_i^{\phi}\right)^4.$$

Now gathering all components f_l to get the squared norm yields:

$$\mathbb{E}_{\mathbf{Z}}\left[\left\|\boldsymbol{f}(\mathbf{s}) - \boldsymbol{f}(\mathbf{s}_o) - \sum_k \frac{\partial \boldsymbol{f}}{\partial z_k}_{|\mathbf{s}_o} (z_k - z_k^o)\right\|^2\right] \leq \frac{n^3}{4} 3K^2 \sum_i \left(\sigma_i^{\phi}\right)^4.$$

□

C.3.4 Variational posterior variance optimization problem

Lemma C.3.4 *For $\alpha > 0$, the function*

$$h_\alpha : \mathbb{R}_{>0} \to \mathbb{R}$$
$$u \mapsto -\log u - \frac{1}{2} + \alpha u^2/2 = \frac{1}{2}\log\frac{1}{u^2} - \frac{1}{2} + \alpha u^2/2$$

is strictly convex and achieves its global minimum $\min h_\alpha = \frac{1}{2}\log\alpha$ for $u^* = \frac{1}{\sqrt{\alpha}}$.

Proof. Function h_α is stricly convex as a sum of two stricly convex

functions. Its derivative,

$$\frac{dh_\alpha}{du}(u) = -\frac{1}{u} + \alpha u,$$

thus vanishes only at the minimum for $u^* = \frac{1}{\sqrt{\alpha}}$. We then get that

$$\min h_\alpha = h_\alpha(u^*) = \frac{1}{2}\log \alpha\,.$$

□

C.4 Related work

C.4.1 Implicit inductive biases in the ELBO

[202] reason about the connection to Principal Component Analysis (PCA) in the context of nonlinear Gaussian VAEs with an isotropic prior and assume that the variational posterior has *diagonal covariance with distinct singular values*. The authors make it explicit that they investigate the consequences of optimizing the ELBO. They locally linearize the decoder to show the inductive bias in VAEs that promotes decoder orthogonality. Their results hold for β-VAEs, where β should be in the range of satisfying the polarized regime assumption (*i.e.*, when the VAE is close to partial posterior collapse). The validity of the assumptions (polarized regime and distinct singular values in $\mathbf{\Sigma}^{\phi}_{\mathbf{s}|\mathbf{x}}$) are only experimentally investigated. The same authors extend their work in [205], completing the connection to PCA for *linear* models. Their experiments, inspired by the connection to PCA for linear models, show that local perturbations in the data prohibit disentanglement for non-linear models.

[207] prove that *linear Gaussian* VAEs with an isotropic prior give rise to a *column-orthogonal decoder* and therefore uniquely recover the PCA coordinate axes (not just the correct subspace, as Probabilistic Principal Component Analysis (PPCA) [502] does), yielding identifiability for Gaussian models—but only when the eigenvalues of the data covariance are distinct. In their work, the decoder variance is shown to be small when avoiding posterior collapse. More interestingly, the authors derive a formula for the ELBO gap in the linear case that is remarkably similar to the IMA objective. We show in Appendix C.5.1 that in the limit of a deterministic decoder linear Gaussian VAEs optimize the IMA objective with $\lambda = 1$.

[203] generalizes [202], as it admits a variational posterior $q_\phi(\mathbf{s}|\mathbf{x})$ with *block-diagonal covariance* with a uniqueness result for diagonal $\mathbf{\Sigma}^{\phi}_{\mathbf{s}|\mathbf{x}}$. The authors derive a formula for the optimal $\mathbf{\Sigma}^{\phi}_{\mathbf{s}|\mathbf{x}}$ [203, Eq. 12], showing that when the decoder Hessian $\mathcal{H}$ is diagonal, the decoder Jacobian will be column-orthogonal even for *non-Gaussian* decoders. Their analysis relies on a "concentrated" $q_\phi(\mathbf{s}|\mathbf{x})$ (*i.e.*, they work in what we term the near-deterministic regime) and sufficiently small values of β—this relationship can be read off from [203, Eq. 12]. Interestingly, the authors also show that rotations of the latents can be ruled out, though they do not connect the decoder structure (especially, column-orthogonality of its Jacobian)

to any specific generative model for the data, or to considerations on identifiability of the ground truth sources.

C.4.2 (Near)-deterministic VAEs

Recent work was inspired by the normalizing flow literature and the shortcomings of the stochastic VAE architecture to propose designs that are (near)-deterministic. Arguments for this regime range from avoiding posterior collapse (as demonstrated in [207]) to avoiding sampling for the reconstruction loss term [203]. Several papers argued for a similar setting: [202] refer to the *polarized regime* (a property of which is that encoder variances are small, cf. [202, Definition 1]), [203] argue for "concentrated" variational posteriors. [500] substitute stochasticity with a regularizer on the decoder Jacobian from an intuitive, whereas [503] motivate these results from an injective flow perspective. [206] also take a normalizing flow perspective to connect VAEs to deterministic models. Besides benefits of avoiding posterior collapse or possible improvements during optimization, this regime serves as a potential connection to the identifiability literature.

C.5 Further remarks on the the IMA–VAE connection

In this section, we elaborate on the connection between VAEs and IMA, by showing that previous work on linear VAEs can be directly connected to optimizing $\mathcal{L}_{\text{IMA}}$. Our intent with this analysis is to provide additional insights about the role of γ in a simpler setting.

C.5.1 Linear VAE from [207]

We restate the linear VAE model of [207]:

$$p_\theta(\mathbf{x}|\mathbf{s}) = \mathcal{N}\left(\mathbf{W}\mathbf{s} + \boldsymbol{\mu}; \frac{1}{\gamma^2}\mathbf{I}_n\right) \tag{C.40}$$

$$q_\phi(\mathbf{s}|\mathbf{x}) = \mathcal{N}\left(\mathbf{V}\left(\mathbf{x} - \boldsymbol{\mu}\right); \mathbf{D}\right), \tag{C.41}$$

where $\mathbf{D}$ is a diagonal matrix, $\mathbf{W}$ the decoder and $\mathbf{V}$ the encoder weights, $\boldsymbol{\mu}$ the mean latent representation.

The authors show that in stationary points, the optimal value for $\mathbf{D}$ is

$$\mathbf{D}^* = \frac{1}{\gamma^2}\left(diag\,\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_n\right)^{-1} \tag{C.42}$$

If we substitute this expression into the ELBO gap (*i.e.*, the KL between the variational and true posteriors), we get a similar expression to c_{IMA}—as formalized in Prop. C.5.1.

Proposition C.5.1 (The ELBO converges to $\mathcal{L}_{\text{IMA}}$ for linear Gaussian

VAEs if $\gamma \to +\infty$) *For linear Gaussian VAEs, in the limit of* $\gamma \to \infty$*, the* ELBO *equals the* IMA*-regularized log-likelihood in stationary points with* $\lambda = 1$.

Proof. In [207, Appendix C.2], it is shown that the gap between exact log-likelihood and ELBO for linear Gaussian VAEs in stationary points reduces to

$$\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] = \frac{1}{2}\left(\log\det\tilde{\mathbf{M}} - \log\det\mathbf{M}\right) \tag{C.43}$$

$$\mathbf{M} = \mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_n \tag{C.44}$$

$$\tilde{\mathbf{M}} = diag\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_n, \tag{C.45}$$

where $\mathbf{W}$ is the decoder weight matrix. Reformulating the above expression, we arrive at :

$$\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] = \log\frac{\left|diag\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_n\right|}{\left|\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_n\right|} \tag{C.46}$$

$$= \log\frac{\left|diag\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_n\right|}{\left|\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_n\right|} \tag{C.47}$$

Noting that $\mathbf{W}^T\mathbf{W}$ is symmetric with a Singular Value Decomposition (SVD) of $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ ($\mathbf{U}$ is orthogonal, $\mathbf{\Lambda}_{ii} = \|[\mathbf{W}]_{:k}\|^2$), and $\mathbf{I}_n = \mathbf{U}\mathbf{U}^T$; thus:

$$\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_n = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T + \frac{1}{\gamma^2}\mathbf{U}\mathbf{U}^T = \mathbf{U}\left[\mathbf{\Lambda} + \frac{1}{\gamma^2}\mathbf{I}_n\right]\mathbf{U}^T$$

Therefore, (C.47) can be reformulated as the left KL-measure of diagonality [172] of the matrix $\mathbf{U}\left[\mathbf{\Lambda} + {}^1\!/\!_{\gamma^2}\mathbf{I}_n\right]\mathbf{U}^T$:

$$\mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] = \log\frac{\left|diag\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_n\right|}{\left|\mathbf{W}^T\mathbf{W} + \frac{1}{\gamma^2}\mathbf{I}_n\right|} \tag{C.48}$$

$$= \log\frac{\left|diag\mathbf{U}\left[\mathbf{\Lambda} + \frac{1}{\gamma^2}\mathbf{I}_n\right]\mathbf{U}^T\right|}{\left|\mathbf{U}\left[\mathbf{\Lambda} + \frac{1}{\gamma^2}\mathbf{I}_n\right]\mathbf{U}^T\right|}, \tag{C.49}$$

which is by definition the local IMA contrast c_{IMA} (cf. [63, Appendix C.1]). When $\gamma \to +\infty$, the above expression converges to the left KL-measure of diagonality for $\mathbf{W}^T\mathbf{W}$, *i.e.*, the local IMA contrast for the decoder.

$\gamma \to +\infty$ thus means that the ELBO converges to the IMA regularized log-likelihood $\mathscr{L}_{\mathrm{IMA}}$ with $\lambda = 1$:

$$\begin{aligned}\mathrm{ELBO} &= \log p_\theta(\mathbf{x}) - \mathrm{KL}\left[q_\phi(\mathbf{s}|\mathbf{x})||p_\theta(\mathbf{s}|\mathbf{x})\right] \\ &= \log p_\theta(\mathbf{x}) - c_{\mathrm{IMA}}(\mathbf{W}, \mathbf{s}),\end{aligned}$$

which concludes the proof. □

Prop. C.5.1, especially (C.49), gives us intuitive understanding on why and how γ influences how much the orthogonality of $\mathbf{W}$ is enforced.

1. Small γ (high observation noise) means that there is no reason to promote the orthogonality of the decoder, as the high noise level (*i.e.*, low-quality fit of $\mathbf{x}$) will drive (C.49) towards diagonality via $1/\gamma^2$.
2. On the other hand, when $\gamma \to +\infty$, then the orthogonality of the decoder is promoted. That is, the decoder precision γ^2 acts akin to a weighting factor influencing how strong the IMA principle should be enforced.

We can observe that the ELBO recovers the exact log-likelihood for column-orthogonal $\mathbf{W}$:

Corollary C.5.1 (For column-orthogonal $\mathbf{W}$ the ELBO equals the exact log-likelihood) *When* $\mathbf{W}$ *is in the form* $\mathbf{W} = \mathbf{OD}$, *then* $\mathrm{diag}\mathbf{W}^T\mathbf{W} = \mathbf{W}^T\mathbf{W} = \mathbf{D}\mathbf{O}^T\mathbf{O}\mathbf{D} = \mathbf{D}^2$, *i.e. the* ELBO *corresponds to the exact log-likelihood since* (C.49) *is zero.*

Corollary C.5.1 also implies that γ does not affect the gap between ELBO and exact log-likelihood for column-orthogonal $\mathbf{W}$.

C.6 Experimental details

C.6.1 The relationship of weight matrix structures and the IMA function class

During the experiments we have used different weight matrices either to *ensure* that the mixing is within or to *exclude* it from the IMA function class. Here we summarize our choices also including the *depth* of the network as it can affect the mixing's place w.r.t. the IMA function class.

When we use *orthogonal* weight matrices (§ 4.4.1,§ 4.4.2), then a single-layer network is within the IMA class, but adding more layers with elements-wise nonlinearities will move the MLP outside the function class. When using *triangular* MLPs (§ 4.4.2), the network is also outside the IMA class (triangular matrices are orthogonal when they are *diagonal*). Thus, we would not need to use triangular weights to design a model outside the IMA class (we could do this with orthogonal matrices). However, as we need to analytically calculate the inverse, we choose triangular weights.

Notably, Möbius transforms [176] are conformal maps (thus, they are in the IMA class) irrespective of the structure of the weight matrix used (cf. Appendix C.6.4 for details).

C.6.2 Self-consistency in practical conditions (§ 4.4.1)

For the self-consistency experiments, the mixing is a 3-layer MLP with smooth Leaky ReLU nonlinearities (see Chapter 5)and orthogonal weight matrices—which intentionally does not belong to the IMA class, since our self-consistency result is not constrained to the IMA class. The 60,000

source samples are drawn from a standard normal distribution and fed into a VAE composed of a 3-layer MLP encoder and decoder with a Gaussian prior. We use 20 seeds for each $\gamma^2 \in \{1e1; 1e2; 1e3; 1e4; 1e5\}$. Additional parameters are described in Tab. C.1. Training is continued until the ELBO* improves on the *validation set* (we use early stopping [504]), then all metrics are reported for the maximum ELBO* (Fig. 4.2).

Table C.1: Hyperparameters for the self-consistency experiments (§ 4.4.1)

Parameter	Values
Encoder	3-layer MLP
Decoder	3-layer MLP
Activation	smooth Leaky ReLU [86]
Batch size	64
# Samples (train-val-test)	$42 - 12 - 6k$
Learning rate	1e−3
n	3
Ground truth	Gaussian
$p_{\mathbf{s}}(\mathbf{s})$	Gaussian
$\mathbf{\Sigma}^{\phi}_{\mathbf{s}\|\mathbf{x}}$	Diagonal
γ^2	$\{1e1; 1e2; 1e3; 1e4; 1e5\}$
# Seeds	20

C.6.3 Relationship between ELBO*, IMA-regularized, and unregularized log-likelihoods (§ 4.4.2)

Table C.2: Hyperparameters for the *triangular MLP* (*not* from the IMA class) ELBO*–$\mathcal{L}_{\text{IMA}}$–log-likelihood experiments (§ 4.4.2)

Parameter	Values
Encoder	3-layer MLP
Decoder	2-layer triangular MLP (ground truth)
Activation	Sigmoid
Batch size	64
# Samples (train-val-test)	$100 - 30 - 15k$
Learning rate	1e−4
n	2
Ground truth	Gaussian
$p_{\mathbf{s}}(\mathbf{s})$	Gaussian
$\mathbf{\Sigma}^{\phi}_{\mathbf{s}\|\mathbf{x}}$	Diagonal
γ^2	[1e1; 1e5]
# Seeds	5
C_{IMA} (mixing)	7.072

For the experiments comparing the ELBO*, IMA-regularized, and unregularized log-likelihoods, data is generated by mixing points from a standard Gaussian prior using an invertible neural network. When the mixing is not in the IMA-class (Tab. C.2), we use a two-layer neural network with sigmoid nonlinearites and triangular weight matrices. When the mixing is from the IMA-class (Tab. C.3), we use a one-layer neural network with orthogonal weight matrices. The data dimensionality in both cases is two.

Training is carried out using a VAE with a decoder fixed to the ground-truth and separate encoder models for the means and variances of the

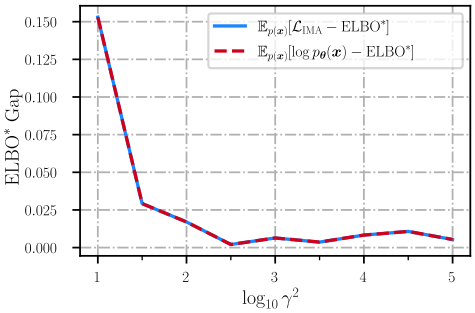


Figure C.1: Comparison of the ELBO*, the IMA-regularized and unregularized log-likelihoods over different γ^2 with an IMA-class mixing

approximate posterior. The encoder comprises two three-layer neural networks with ReLU non-linearities and a hidden layer size of 50. Due to training instabilities when using a large γ, we train the model by first fixing the mean encoder to the ground-truth inverse of the mixing for the first 30 epochs; thus, only training the variances. We then train both for the remaining epochs. Training is stopped after the ELBO* plateaus on the *validation set*. A training set of 100,000 samples is used, with a validation set and test set of 30,000 and 15,000 samples, respectively. The learning rate is 1e−4 and the batch size 64.

We provide additional results when the mixing is from the IMA class (Tab. C.3): as C_{IMA} is zero, we expect that both $\mathcal{L}_{\text{IMA}}$ and the unregularized log-likelihood match. Indeed, this is what Fig. C.1 demonstrates.

Table C.3: Hyperparameters for the *orthogonal MLP* (from the IMA class) ELBO*–$\mathcal{L}_{\text{IMA}}$–log-likelihood experiments (§ 4.4.2)

Parameter	Values
Encoder	3-layer MLP
Decoder	1-layer orthogonal MLP (ground truth)
Activation	Sigmoid
Batch size	64
# Samples (train-val-test)	100 − 30 − 15k
Learning rate	1e−4
n	2
Ground truth	Uniform
$p_{\mathbf{s}}(\mathbf{s})$	Uniform
$\mathbf{\Sigma}^{\phi}_{\mathbf{s}\|\mathbf{x}}$	Diagonal
γ^2	[1e1; 1e5]
C_{IMA} (mixing)	0

C.6.4 Connecting the IMA principle, γ^2, and disentanglement (§ 4.4.3)

Synthetic data (Möbius transform) We use 3-dimensional conformal mixings (*i.e.*, the Möbius transform [176]) from the IMA class with the functional form:

$$\mathbf{x} = \boldsymbol{t} + \alpha \frac{\mathbf{W}(\mathbf{s} - \boldsymbol{b})}{\|\mathbf{s} - \boldsymbol{b}\|^{\epsilon}},$$

where $\boldsymbol{t}, \boldsymbol{b} \in \mathbb{R}^n$, $\mathbf{W} \in \mathbb{R}^{n \times n}$, $\alpha \in \mathbb{R}$, and $\epsilon = 2$ (to ensure nonlinearity) with $n = 3$. Both ground-truth and prior distributions are *uniform* to avoid the singularity when $\mathbf{s} = \boldsymbol{b}$.

To determine whether a mixing from the IMA class is beneficial for disentanglement, we apply a volume-preserving linear map after the Möbius transform (using 100 seeds) to construct a mixing outside of the IMA class. We fix γ^2 = 1e5 and report further parameters in Tab. C.4. Training is continued until the ELBO* improves on the *validation set* (we use early stopping [504]), then all metrics are reported for the maximum ELBO* (Fig. 4.3).

Table C.4: Hyperparameters for the *synthetic (Möbius)* IMA–disentanglement experiments (§ 4.4.3) with a linear map

Parameter	Values
Encoder	3-layer MLP
Decoder	3-layer MLP
Activation	smooth Leaky ReLU [86]
Batch size	64
# Samples (train-val-test)	$42 - 12 - 6$k
Learning rate	1e−3
n	3
Ground truth	Uniform
$p_{\mathbf{s}}(\mathbf{s})$	Uniform
$\mathbf{\Sigma}^{\phi}_{\mathbf{s}\|\mathbf{x}}$	Diagonal
γ^2	1e5
# Seeds	100
C_{IMA} (mixing)	$[0.398; 6.761]$

Image data (Sprites) We train a VAE (not β-VAE) with a factorized Gaussian posterior and Beta prior on a Sprites image dataset generated using the spriteworld renderer [214] with a Beta ground truth distribution. Similar to [215], we use four latent factors, namely, *x- and y-position, color and size*, and omit factors that can be problematic, such as shape (as it is discrete) and rotation (due to symmetries) [202, 213]. Our choice is motivated by [201, 216] showing that the data-generating process presumably is in the IMA class. The architecture both for encoder and decoder consists of four convolutional and three linear layers with ReLU nonlinearities (Tab. C.5). Training is continued until the ELBO* improves on the *validation set* (we use early stopping [504]), then all metrics are reported for the maximum ELBO*.

Table C.5: Hyperparameters for the *image (Sprites)* IMA–disentanglement experiments (§ 4.4.3)

Parameter	Values
Encoder	4-layer Conv2D + 3-layer MLP
Decoder	4-layer Conv2D + 3-layer MLP
Activation	ReLU
Batch size	64
# Samples (train-val-test)	$42 - 12 - 6$k
Learning rate	1e−5
n	3
Ground truth	Beta
$p_{\mathbf{s}}(\mathbf{s})$	Beta
$\mathbf{\Sigma}^{\phi}_{\mathbf{s}\|\mathbf{x}}$	Diagonal
γ^2	1e0
# Seeds	10

C.7 Computational resources

The self-consistency (§ 4.4.1), the likelihood comparison (§ 4.4.2), and the synthetic experiments with the Möbius transform (§ 4.4.3, particularly Fig. 4.3) were ran on a MacBook Pro with a Quad-Core Intel Core i5 CPU and required approximately nine days. The Sprites experiments (§ 4.4.3, particularly Fig. 4.5) required approximately four and a half days on an Nvidia RTX 2080 GPU.

D

Additional Material on Chapter 5

D.1 Backpropagation in neural networks

We will follow [505], Chapter 7, section 7.3.3 for the notation. Let us define a two-layer neural network

$$\mathbf{g}_\theta(\mathbf{x}) = \boldsymbol{\sigma}\left(\mathbf{W}_2\boldsymbol{\sigma}\left(\mathbf{W}_1\mathbf{x}\right)\right) \tag{D.1}$$

where we also define

$$\begin{aligned}\mathbf{z}_2 &= \boldsymbol{\sigma}\left(\mathbf{W}_2\mathbf{z}_1\right)\\ \mathbf{z}_1 &= \boldsymbol{\sigma}\left(\mathbf{W}_1\mathbf{x}\right)\ .\end{aligned}$$

and

$$\begin{aligned}\mathbf{u}_2 &= \boldsymbol{\sigma}'\left(\mathbf{W}_2\mathbf{z}_1\right)\\ \mathbf{u}_1 &= \boldsymbol{\sigma}'(\mathbf{W}_1\mathbf{x})\end{aligned}$$

and

$$\begin{aligned}\mathbf{y}_2 &= \mathbf{W}_2\mathbf{z}_1\\ \mathbf{y}_1 &= \mathbf{W}_1\mathbf{x}\end{aligned}$$

We need to consider the contributions to the objective function due to the terms $\mathcal{L}_p$ and $\mathcal{L}_J^1$ (the contribution due to $\mathcal{L}_J^2$ will be dealt with separately). For $\mathcal{L}_p$, we define

$$e(x) = \frac{\partial}{\partial x}\log p(x')|_{x'=x}$$

and

$$\mathbf{e} = \begin{pmatrix} e(z_2^1)\\ e(z_2^2)\\ \vdots\\ e(z_2^n)\end{pmatrix}$$

To deal with the terms in $\mathcal{L}_J^1$, we define

$$\begin{aligned}h(x) &= \frac{\partial}{\partial x}\log x'|_{x'=x} &\text{(D.2)}\\ &= \frac{1}{x} &\text{(D.3)}\end{aligned}$$

and

$$\mathbf{h}_k = \begin{pmatrix} h(u_k^1) \\ h(u_k^2) \\ \vdots \\ h(u_k^D) \end{pmatrix}$$

for $k = 1, 2$. During forward propagation, we store the $\mathbf{D}_k = \operatorname{diag}(\sigma'(\mathbf{y}_k))$ for $k = 1, 2$,

$$\mathbf{D}_k = \begin{pmatrix} \sigma'(y_k^1) & 0 & \cdots & 0 \\ 0 & \sigma'(y_k^2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma'(y_k^n) \end{pmatrix}$$

and the $\mathbf{G}_k = \operatorname{diag}(\sigma''(\mathbf{y}_k))$ for $k = 1, 2$,

$$\mathbf{G}_k = \begin{pmatrix} \sigma''(y_k^1) & 0 & \cdots & 0 \\ 0 & \sigma''(y_k^2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma''(y_k^n) \end{pmatrix}$$

for example, if the nonlinearity were a sigmoid function $\sigma(x) = (1 + e^{-x})^{-1}$, the second derivative would be $\sigma''(x) = \sigma(x)(1-\sigma(x))\,(1 - 2\sigma(x))$. Then

$$\boldsymbol{\delta}_2 = \mathbf{D}_2\mathbf{e} + \mathbf{G}_2\mathbf{h}_2$$

and

$$\boldsymbol{\delta}_1 = \mathbf{D}_1\mathbf{W}_2\boldsymbol{\delta}_2 + \mathbf{G}_1\mathbf{h}_1$$

In general, the following recursive relationship holds

$$\boldsymbol{\delta}_k = \mathbf{D}_k\mathbf{W}_{k+1}\boldsymbol{\delta}_{k+1} + \mathbf{G}_k\mathbf{h}_k \tag{D.4}$$

Which results in the update rule

$$\Delta\mathbf{W}_k = -\mu\mathbf{z}_{k-1}\boldsymbol{\delta}_k^\top ,$$

where $\mathbf{z}_0 = \mathbf{x}$. Notice that the only necessary operations are vector-matrix, matrix-vector and vector-vector multiplications.

D.1.1 Relative gradient

Now if we want to use the relative/natural gradient trick each of these terms needs to be multiplied by $\mathbf{W}_k^\top\mathbf{W}_k$ from the right.

$$\Delta\mathbf{W}_k = -\mu\mathbf{z}_{k-1}\boldsymbol{\delta}_k^\top\mathbf{W}_k^\top\mathbf{W}_k .$$

Terms in $\mathcal{L}_J^2$ The terms in $\mathcal{L}_J^2$, consisting of $\log|\mathbf{W}_k|$ give as gradient $\left(\mathbf{W}_k^\top\right)^{-1}$. This requires a $n \times n$ matrix inversion for each of the matrices. Our strategy to avoid it is to substitute the ordinary gradient with a relative gradient, where we multiply the gradient (with respect to the whole objective but for each layer separately) by $\mathbf{W}_k^\top\mathbf{W}_k$ from the right. In this case, the updates for the $\mathbf{W}_k$ terms simply become proportional to

the $\mathbf{W}_k$ themselves. Therefore, the update rule becomes

$$\Delta \mathbf{W}_k = -\mu(\mathbf{z}_{k-1}\boldsymbol{\delta}_k^\top \mathbf{W}_k^\top \mathbf{W}_k + \mathbf{W}_k)\,. \tag{D.5}$$

As we already noted, the operations involved in these updates can be performed in a way such that no matrix-matrix multiplication needs to be performed – only matrix-vector and vector-vector multiplication. This is more apparent when the update rules are rewritten as below

$$\Delta \mathbf{W}_k = -\mu\left(\mathbf{z}_{k-1}\left(\left(\boldsymbol{\delta}_k^\top \mathbf{W}_k^\top\right)\mathbf{W}_k\right) + \mathbf{W}_k\right)\,. \tag{D.6}$$

D.2 Related work

In the following, we present a review of related work in tractable deep density estimation and invertible neural networks.

Normalizing flows The modern conception of normalizing flows was introduced in [225], which discussed density estimation through the composition of simple maps. In [226], it was then proposed that deep density models implemented through neural networks could be used in order to construct bijective maps to a representation space and obtain normalized probability density estimates. Since then, the focus mainly shifted to scalability; [97, 227] introduced scalable architectures, further refined in [223] to make them more scalable and suitable for practical applications; [210] applied the results to variational inference. Comprehensive reviews on normalizing flows can be found in [71, 98].

Autoregressive flows Autoregressive flows are among the most used in practice. They involve maps which can be written as $z_i' = \tau(z_i; \boldsymbol{h}_i)$, with $\boldsymbol{h}_i = c_i\left(\mathbf{z}_{<i}\right)$. τ is termed the *transformer* and is a strictly monotonic function of z_i, and c_i is termed the i-th *conditioner*. Its constraint is that the $i-$th conditioner can only take variables with dimension indices less than i as an input. This results in an overall transformation with a triangular Jacobian; the determinant is therefore tractable and can be computed in $\mathcal{O}(n)$ time. Autoregressive flows differ in the way the transformer and conditioner are implemented; most commonly used are affine autoregressive flows [97, 223, 227, 228, 247] and non-affine neural transformers [179].

Linear flows A strict generalization of autoregressive flows, where the Jacobian is not constrained to be triangular, is given by linear flows, which are essentially transformations of the form $\mathbf{z}' = \mathbf{W}\mathbf{z}$, where $\mathbf{W}$ is a $n \times n$ invertible matrix. The Jacobian of the trasformation is simply $\mathbf{W}$ and both computing and optimizing its determinant takes time $O(n^3)$ in general. To obtain a better scaling behaviour, [97] and [506] proposed to parameterize the invertible $\mathbf{W}$ matrix via matrix decomposition. One possibility is to compute the **PLU** decomposition of $\mathbf{W}$ and optimize the $\mathbf{L}$ and $\mathbf{U}$ triangular transformations. The drawback in this approach is that the permutation matrix $\mathbf{P}$ cannot be learnt. A more flexible alternative is to consider the **QR** decomposition of $\mathbf{W}$, where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{R}$ is upper triangular. However computing $\mathbf{Q}$ in full generality requires $O(n^3)$ operations, matching the complexity of the naive optimization of linear flows. [507] showed that we can apply the $\mathbf{Q}$ transformation as a sequence of at most n symmetry transformations

each taking linear time, effectively making it possible to compute and optimize the **QR** parameterization of **W** in $O(n^2)$ time; note however that the sequential nature of the computation makes the method unsuitable for optimization on hardware accelerators. An experimental comparison of the performance of the **PLU** and **QR** decompositions against the direct optimization of **W** is found in [506].

Flows based on residual transformations Another class of normalizing flows is based on invertible transformations of the form $\mathbf{z}' = \mathbf{z} + g_\phi(\mathbf{z})$; this kind of flows are termed *residual flows*. Two main approaches can be applied to build invertible residual flows: the first exploits the matrix determinant lemma and also results in determinants with $\mathcal{O}(n)$ computation time; however, there is no analytical way of computing their inverse. Examples of these approaches are Sylvester flows [508], planar flows [210] and radial flows [210, 225]. The second approach is that of contractive flows [230]: in this case, the determinant can not be computed simply; likelihood-based training of these models therefore needs to rely on a Hutchkinson's trace based approximation to the exact log-likelihood.

Continuous time flows A separate line of work focuses on building *continuous flows*; in these approaches, the flow's infinitesimal dynamics is parametrized in continuous time, and the corresponding transformation is then found by integration [224, 231]; Hamiltonian Flows [210] can also be regarded as such kind of flows.

Other works Recently, many works have proposed ways of incorporating convolutional modules in normalizing flows, for example see [223, 506, 509]. In particular, [510] presents a formalization of the problem which bears some similarities to ours, while focusing on convolutional layers instead of fully connected ones. Other work has been dedicated to constructing invertible neural networks, see for example [31, 239, 511].

D.3 Complexity of mathematical operations involved in gradient computation

We want to characterize the complexity of computing

$$\nabla_{\boldsymbol{\theta}} \log |\det \mathbf{J}_{\mathbf{g}_{\boldsymbol{\theta}}}(\mathbf{x})| \,, \tag{D.7}$$

where $\mathbf{g}_{\boldsymbol{\theta}}$ is a neural network.

We will first recapitulate the computational complexity of the main mathematical operations we employ (see e.g. [512]). Then we'll recapitulate the complexity of forward evaluation and backpropagation in neural networks. Finally, we'll discuss the implications on the complexity of computing the term in equation (D.7) with the three methods discussed in the paper — namely, based on automatic differentiation, the standard computation described in section 5.3 and the relative gradient based computation.

D.3.1 Matrix operations

Matrix-vector and vector-vector multiplication The multiplication of a $n \times n$ matrix with a $n \times 1$ vector scales as $\mathcal{O}(n^2)$. Same for the outer product between two vectors of dimension $n \times 1$.

Matrix-matrix multiplication For the multiplication of two square matrices of size $n \times n$

- An implementation of the Bareiss algorithm would scale as $\mathcal{O}(n^3)$;
- An implementation of the Strassen algorithm would scale as $\mathcal{O}(n^{2.807\ldots})$;
- An implementation of the Coppersmith-Winograd algorithm would scale as $\mathcal{O}(n^{2.373\ldots})$.

In practice, what is usually implemented in linear algebra libraries is some flavor of the Strassen algorithm (this is because the Coppersmith-Winograd algorithm, while having a more favorable asymptotic behaviour, is effectively slower if n is not extremely high).

Matrix inversion To find the inverse of a matrix of size $n \times n$

- An implementation of Gauss-Jordan elimination would scale as $\mathcal{O}(n^3)$;
- An implementation of the Strassen algorithm would scale as $\mathcal{O}(n^{2.807\ldots})$;
- An implementation of the Coppersmith-Winograd algorithm would scale as $\mathcal{O}(n^{2.373\ldots})$.

Determinant To find the determinant of a matrix of size $n \times n$

- An implementation of the Bareiss algorithm would scale as $\mathcal{O}(n^3)$;
- Algorithms based on fast matrix multiplication scale as $\mathcal{O}(n^{2.373\ldots})$.

For simplicity, in most of our considerations on complexity we assume that the computation of the determinant, the computation of the inverse and the multiplication of two square matrices have cubic cost. Notice that the cost of these operations always dominates over that of matrix-vector and vector-vector multiplication.

D.3.2 Other operations involved in the Jacobian term computation

Other operations turn out to be ininfluent on the overall computational complexity. Namely logarithms, absolute values, sums have no relevant effect in terms of asymptotic scaling, since their computational cost is dominated by that of the most expensive matrix operations listed above.

D.3.3 Complexity of neural network operations

Forward pass in a neural network The complexity of the forward pass in a neural network depends on the neural network structure. For simplicity, we will consider fully connected Neural Networks, which due to their dense structure will provide an upper bound for the complexity of most of the nets used in practice. Given an input vector, the forward pass is comprised of a sequential series of matrix-vector operations, plus elementwise operations on the resulting vector. The matrix-vector operations dominate the complexity; for an L layer neural network, there are L such operations. Therefore, for data of dimensionality n, the complexity of a forward pass in a Neural Network for a single data sample is $\mathcal{O}(Ln^2)$.

Minibatching The objectives should, in principle, be optimized on the full batch. Stochastic optimization [513] relies on the idea that the update steps in the optimization process can be performed on subsets of the whole training data, called minibatches. In practice these objectives will always be computed on minibatches, so the expected value must be substituted with its empirical estimate over a single minibatch. The minibatch size should in principle be considered when considering how the algorithm scales. In the remainder, however, we will neglect this term, as minibatches used in practice are usually quite small.

Gradient computation On top of this, we also need to consider the gradient computation. Since the gradient is taken over the scalar loss function, this implies (through backpropagation or reverse mode differentiation) no increase in the asymptotic computational cost. We further elaborate on this in the next section.

D.3.4 Computing the Jacobian with automatic differentiation

Jacobian through automatic differentiation Automatic differentiation [240] includes two main operational modes: the forward mode and the backward mode. Consider the computation of the Jacobian of a function $\mathbf{g}_\theta : \mathbb{R}^n \to \mathbb{R}^d$. The complexity of computing the Jacobian will depend on whether we use forward or reverse mode AD. This changes the complexity of the operation:

- forward mode requires $n\, c\, \text{ops}(\mathbf{g}_\theta)$ operations, where n is the dimensionality of the data and c is a constant, $c < 6$ and typically $c \in [2, 3]$ (see [514]);
- reverse mode requires $d\, c\, \text{ops}(\mathbf{g}_\theta)$ operations.

In the case of dimensionality reduction, reverse mode differentiation (of which backpropagation represents an instance) is clearly more efficient. This is the case when the output of the function is scalar ($d = 1$); thus, this explains our claim that gradients computation with backpropagation implies no increase in the asymptotic computational cost with respect to the forward pass alone.

For neural networks where all layers (including input and output) have the same size, both methods result in the same complexity. So in that case neither is better in terms of computational complexity — though in practice it is known that reverse mode performs better [515]. For such neural networks (including those we consider) therefore, given that $\text{ops}(\mathbf{g}_\theta)$ is $\mathcal{O}(Ln^2)$, the overall complexity of the Jacobian computation via automatic differentiation is $\mathcal{O}(Ln^3)$.

The gradient of the objective can then be computed via backpropagation; however, the forward evaluation is what dominates the overall complexity.

Standard and relative gradient computations The evaluation of the two terms $\mathcal{L}_p$ and $\mathcal{L}_J^1$ requires a forward pass of the neural networks, thus scaling as $\mathcal{O}(Ln^2)$. As we discussed, backpropagation to compute the gradient does not increase the overall cost. For $\mathcal{L}_J^2$, as we have shown, the gradient can be computed without need to actually evaluate the corresponding term (that is, side-stepping the determinant computation). However, the standard computation of the gradient still requires computing inverses of all the weight matrices, resulting in a cubic cost operation for each layer — thus utimately in $\mathcal{O}(Ln^3)$ cost.

When using the relative gradient, this inversion can be avoided, and computing the gradients of $\mathcal{L}_J^2$ implies *no additional costs*. The overall cost of the gradient computation is therefore simply $\mathcal{O}(Ln^2)$.

D.4 Implementation details

To efficiently optimize our objective (e.g. equation (5.3) in the main paper) we need to implement a variant of the backpropagation algorithm as detailed in appendix D.1. In particular, we need to compute the updates (equation (5.15) in the main paper) avoiding expensive matrix-matrix multiplications. This section is devoted to the description of an implementation strategy that takes advantage of Automatic Differentiation (AD), in order to have full flexibility in the definition of our model architectures and loss functions.

Although all modern deep learning frameworks include automatic differentiation libraries, they implement the standard backpropagation algorithm. To implement our variant, we have two straightforward alternatives:

- tweak some existing AD libraries to let us access the extra terms we need;
- implement our own AD library with the extra functionality we need.

The second alternative is easily excluded as we don't want to reinvent the wheel and the development effort would be too much. The first alternative is somewhat viable, but not future proof; we would be faced with the need to support our own modifications on top of the AD library we use.

We obviate to these problems with a little trick: we introduce in our architectures some dummy layers to accumulate the partial results that the standard backpropagation computes in the backward pass. This approach solves the previous problems by being:

- universal: it can be easily implemented on top of whatever AD library that computes reverse-mode AD, without tweaking the internals of the library;
- efficient: the dummy layer operations are $\mathcal{O}(1)$.

D.4.1 The Accumulator layer

To obtain the gradient updates (D.5) we need to compute the $\boldsymbol{\delta}$ terms (D.4). To better understand what these terms represent, we can consider a simple 2-layers "scalar" network, i.e. a network in which inputs, outputs and weights are scalar values:

$$\begin{aligned} f(x; \boldsymbol{w}) &= w_2 \sigma(w_1 x) \\ &= w_2 \sigma(y_1) \\ &= w_2 z_1 \\ &= y_2 \end{aligned} \tag{D.8}$$

where $\boldsymbol{w}$ is the vector of scalar parameters, σ is the activation function of choice and

$$y_1 = w_1 x, \quad y_2 = w_2 z_1, \quad z_1 = \sigma(y_1)\,.$$

Given a loss function $\mathcal{L}$, the gradient of $\mathcal{L}$ with respect to w_1 is easily computed with application of the chain rule

$$\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial y_2} \frac{\partial y_2}{\partial z_1} \frac{\partial z_1}{\partial y_1} \frac{\partial y_1}{\partial w_1} \tag{D.9}$$

In this simple case, it is easy to isolate δ in the gradient equation:

$$\frac{\partial \mathcal{L}}{\partial w_1} = \delta_1 \frac{\partial y_1}{\partial w_1} \tag{D.10}$$

Reverse mode AD libraries necessarily compute all the partial derivatives in (D.9) and thus the δ_1 term we need. Unfortunately, the partial results are usually not accessible by the users. To access such terms without dealing with the internals of the AD libraries, we can introduce a parameterized function

$$a(x; \lambda) = x + \lambda$$

and redefine our scalar network as

$$f(x; \boldsymbol{w}) = w_2 \sigma(a(y_1)) \tag{D.11}$$

The gradient with respect to w_1 becomes

$$\frac{\partial \mathcal{L}}{\partial w_1} = \frac{\partial \mathcal{L}}{\partial y_2}\frac{\partial y_2}{\partial z_1}\frac{\partial z_1}{\partial a}\frac{\partial a}{\partial y_1}\frac{\partial y_1}{\partial w_1} \tag{D.12}$$

The introduction of a is only a trick; in order not to modify the gradients nor the behaviour of the scalar network, we require

$$\begin{aligned} a(y_1) &= y_1 \\ \frac{\partial z_1}{\partial a} &= \frac{\partial z_1}{\partial y_1} \\ \frac{\partial a}{\partial y_1} &= 1 \end{aligned} \tag{D.13}$$

which is easily achieved by setting $\lambda = 0$.

The benefit of introducing this accumulator layer a is that now we can ask the AD library to compute the gradients with respect to the dummy parameter λ; it is easy to verify that

$$\frac{\partial a}{\partial \lambda} = \delta_1 \tag{D.14}$$

thus making it possible to obtain the δ terms we need to compute (D.5).

D.5 Universal approximation capacity in normalizing flows

Universal approximation for densities is a property often discussed in the context of autoregressive normalizing flows. It can be shown, based on the proof of existence and non-uniqueness of solutions to the nonlinear ICA problem [70], that any distribution can be mapped onto a factorized base distribution by an invertible function with triangular Jacobian, provided that the function class used for this mapping is large enough. Normalizing flows with triangular Jacobians and a high number of parameters therefore have this approximation capacity (see e.g. [179]). However, they can obviously not represent all possible *functions* — but only those with triangular Jacobians. They can therefore not be used to learn proper inverse functions and perform useful feature extraction.

A more general notion of universal approximation is the one usually discussed in the neural network literature, that is — universal approximation for functions. It has been shown that standard multilayer feedforward networks can approximate any continuous function to any degree of accuracy. For example, [516] proved that a standard multilayer feedforward network with a locally bounded piecewise continuous activation function can approximate any continuous function to any degree of accuracy if and only if the network's activation function is not a polynomial. Biases also play a crucial role in this proof, as universal approximation capacity wouldn't be possible without.

While the proof above does not directly apply to our case, since it requires hidden layers with arbitrary width, we discuss how to incorporate biases in our training procedure in appendix D.6, in order to increase the expressivity of our model. We describe the nonlinearities we employed in appendix D.8.

D.6 Relative gradient for the augmented matrix

In order to allow for the training of neural networks with biases, we present a heuristic based on the fact that affine transformations involving vector-matrix products plus biases can be represented as a single matrix operation through the formalism of the augmented matrix (see e.g. [505]).

Linear affine operations of the form $\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$ can be represented via an augmented matrix as follows

$$\begin{bmatrix} \mathbf{y} \\ 1 \end{bmatrix} = \left[\begin{array}{ccc|c} & \mathbf{W} & & \mathbf{b} \\ 0 & \dots & 0 & 1 \end{array}\right] \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix} = \overline{\mathbf{W}} \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}, \tag{D.15}$$

where we refer to the matrix $\overline{\mathbf{W}}$ as *augmented matrix*.

The question is whether the relative gradient trick can be applied to the augmented matrix. The main issue is that we would like, throughout our optimization procedure, to remain on the manifold of augmented matrices; that is, we do not want to change the last row of $\overline{\mathbf{W}}_k$. Therefore, the problem becomes a constrained optimization problem.

The $\mathscr{L}_J^2$ term It is easy to see that $\det \overline{\mathbf{W}}_k = \det \mathbf{W}_k$. The ordinary gradient for all terms in the last column and row of $\overline{\mathbf{W}}_k$ will therefore be equal to zero, and this will not be changed by the relative gradient trick; therefore, the contribution of this term will not lead us away from the manifold of augmented matrices.

The $\mathscr{L}_p$ and $\mathscr{L}_J^1$ terms Both the $\mathbf{y}_k$ and $\mathbf{z}_k$ terms will however be influenced by the presence of biases, so the gradients on the first n elements of the last column (that is $\mathbf{b}_k$) will be nonzero. Through the multiplication with $\overline{\mathbf{W}}_k^\top \overline{\mathbf{W}}_k$, the updates given by the relative gradient on the last row of $\overline{\mathbf{W}}_k$ will therefore in general be nonzero, thus implying moving outside of the manifold we are interested in.

To solve this issue, we use a projected gradient algorithm, enforcing that the update for the last row of $\overline{\mathbf{W}}_k$ at each step is equal to zero. We call this algorithm *projected relative gradient descent*.

In practice, we can use the augmented matrix formalism to apply the relative trick to the full parameters and then extract only the updates for the parameters of interest $\mathbf{W}$, $\mathbf{b}$ disregarding the dummy row in (D.15). Denoting by $\mathbf{G}$ the gradients of $\mathbf{W}$ and by $\mathbf{g}_b$ the gradients of $\mathbf{b}$, we can compute the relative gradients as

$$\left[\begin{array}{c|c} \mathbf{G} & \mathbf{g}_b \\ \hline \mathbf{g} & g \end{array}\right] \overline{\mathbf{W}}^\top \overline{\mathbf{W}} = \left[\begin{array}{c|c} \mathbf{G}\mathbf{W}^\top\mathbf{W} + \mathbf{g}_b\mathbf{b}^\top\mathbf{W} & \mathbf{G}\mathbf{W}^\top\mathbf{b} + \mathbf{g}_b\mathbf{b}^\top\mathbf{b} + \mathbf{g}_b \\ \hline \dots & \dots \end{array}\right] \tag{D.16}$$

The relative gradient updates we need are then given by

$$\Delta \mathbf{W} \rightarrow \mathbf{G}\mathbf{W}^\top \mathbf{W} + \mathbf{g}_b \left(\mathbf{b}^\top \mathbf{W}\right) \tag{D.17}$$

$$\Delta \mathbf{b} \rightarrow \mathbf{G}\left(\mathbf{W}^\top \mathbf{b}\right) + \mathbf{g}_b(1 + \mathbf{b}^\top \mathbf{b}) \tag{D.18}$$

Note that $\mathbf{G}$ is nothing more then the standard backpropagation update (5.6), thus we can efficiently compute $\Delta \mathbf{W}$ by avoiding matrix-matrix multiplications as in (5.15). For $\Delta \mathbf{b}$ we can directly avoid matrix-matrix multiplications by taking some care in the evaluation of (D.18).

D.7 Convolutions

The convolutional neural network [517] is composed of modules whose main components are: (i) a convolution layer; (ii) a pooling layer; (iii) a nonlinearity.

The convolution operation We follow the same notation as in [517]. Typically, inputs to the convolution layers are order 3 tensors with size $H^l \times W^l \times D^l$. A convolution kernel is also an order 3 tensor with size $H \times W^l \times D^l$. If D convolutions are used, this results in a order 4 tensor $\mathbb{R}^{H \times W^l \times D^l \times D}$ of parameters. If the input is $H \times W^l \times D^l$ and the kernel size is $H \times W^l \times D^l \times D$, the convolution result has size $(H^l - H + 1) \times (W^l - W + 1) \times D$. In our setting, note that the number of channels which can be used in practice is constrained, due to the formula in equation (5.3), which requires the input and output dimensionalities to be equal.

Are convolutional neural networks invertible? The convolution operation was shown to be invertible under some mild conditions. See [246] and [510], section 3.3, describing how Gaussian (or Uniform) sampled $c \times c \times r \times r$ parameter tensors will yield invertible convolutional layers with probability 1.

The pooling layer can be substituted with an invertible counterpart (see [31], section 3; or [510], figure 3), which basically becomes a tensorial extension of the permutation operation. As usual, an invertible nonlinearity can be chosen.

Relative gradient for the convolution For a convolution layer that preserves the number of channels in the input, we can directly write the operation in the form of a square matrix. In this case we can compute the relative gradient as explained in section 5.4, and we can obtain the gradients with respect to the filter entries by careful application of the chain rule. We however leave the precise theoretical derivation and experiments for future work.

D.8 Experiments

D.8.1 Computation of relative vs. ordinary gradient

Computational cost In section 5.5 and figure 5.1 we compared the computational cost of computing log-likelihood gradients with our newly proposed method and a naive backpropagation implementation when using hardware accelerators. Specifically, we used one Tesla P100 GPU card equipped with 16 GB of dedicated memory and circa 3500 computing cores. In figure D.1 we show the same comparison for a computation platform comprising 48 cpu threads (Intel Xeon Processor E5-2650 v4 @ 2.20 GHz base frequency, 2.90 GHz max frequency) operating in parallel with about 250 GB of available RAM memory. It is hard to spot the expected theoretical improvement from $O(n^3)$ to $O(n^2)$, but a practical gain of about 2 orders of magnitude in computation time emerges in favor of the relative gradient computation.

In order to directly compare the execution times disregarding bottlenecks due to memory operations, we performed all of the experiments with no garbage collection. Anyways, by using always the same batch we made our experiments not very memory intensive and repeating the experiments with garbage collection enabled didn't show any substantial difference; we therefore don't report the plot.

Memory consumption It is usual in deep learning to be constrained by the memory consumption of the models in use, as the available memory on hardware accelerators is typically scarce. To operate, a network needs to store the data, the intermediate activations (needed to compute gradients) and the parameters. For our simple architecture, the bottleneck is the storage of the parameters; this is because we don't employ very deep architectures, so the amount of intermediate activations to store is limited, and the size of the parameters grows quadratically with respect to the data size, meaning that parameters storage clearly dominate over data storage (this is assuming that data are loaded in small minibatches, which is the norm). This is certainly problematic for very high-dimensional datasets (i.e. high definition images) but even from this point of view we have a clear advantage over an explicit optimization of the Jacobian term with automatic differentiation. In this latter case, in fact, we need to compute the full Jacobian of the affine transformations for each individual data point; like for the weight matrices, the size of these terms grows quadratically with the input size, further increasing the memory footprint of the optimization procedure.

As a simple example, we can compare the approximate memory requirements of the two methods in the moderately high-dimensional case with $n = 20000$. For a modest 2-layers network and employing Float32 weights (each requiring 4 Bytes (B) for storage), the memory needed to store the parameters amounts to $n^2 \times 4B \times 2(\text{layers}) = 3.2GB$. Assuming a minibatch size of 100, data and activations require around 10-100 MB which is clearly negligible. The computed gradients will require the same space as the parameters, raising the memory footprint to over 6GB. For the gradient computations themselves, our method doesn't require additional memory (theoretically), while explicit automatic differentiation requires storing as many jacobian terms as the size of the minibatch, thus

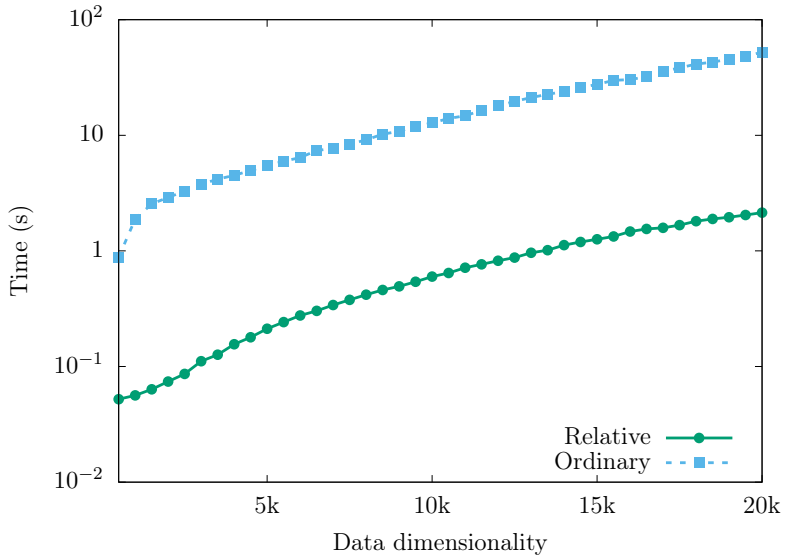


Figure D.1: Comparison of the average computation times of a single evaluation of the gradient of the log-likelihood over a batch of size 100. Values are the mean over 5 steps, and the experiments have been run 5 times on a CPU cluster.

requiring over 300GB in our simple case. As this is clearly unfeasible on common hardware accelerators, we can drop the parallelization of the jacobian terms computation to considerably reduce memory consumption (bringing it down to over 9GB in our case), but this comes at the cost of further slowing down an already inefficient procedure.

While the simple analysis above shows a clear advantage for our proposed method, from the practical point of view many additional technical details might play a role in incrementing the memory requirements of both methods (e.g. loading of libraries and computing environment, just-in-time compilation steps, intermediate computations that can't be fused together...). In figure D.2 we report a simple profiling of the memory consumption of the two methods, which shows how the difference is relevant in practice.

D.8.2 Relative gradient optimization behaviour with different optimizers

In this section we report some additional observations analyizing the relative gradient optimization behaviour with different optimizers.

In figures D.3 and D.4 we compare the optimization behaviour using vanilla Stochastic Gradient Descent (SGD) and Adam. Results on toy datasets like those in figure 2 in the main paper are shown in figure D.3. It can be seen that the data densities are modeled convincingly. We also report (figure D.4) the evolution of the loss with SGD and Adam on density estimation on MNIST. The two methods seem to reach convergence at comparable speed: SGD is faster initially, but in the longer run Adam appears to achieve a better performance faster. Ultimately, both methods achieve a comparably good result.

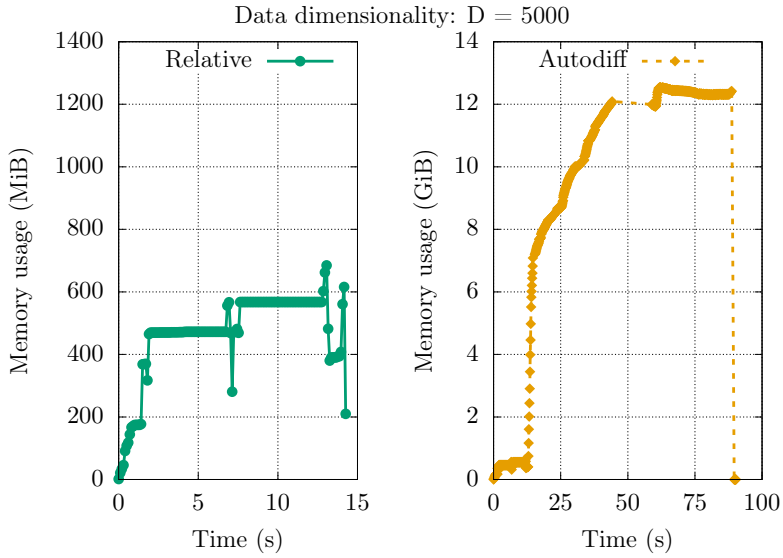


Figure D.2: Comparison of the memory consumption for a single gradient evaluation. With D = 5000 our simplified analysis predicts a lower bound in the memory consumption of 400 MB for storing the parameters and the computed gradients; given that at startup time we observe a base memory consumption of almost 200 MB (computing environment + loaded libraries) we can see that our relative gradient implementation comes very close to the theoretical limit. For the naive autodiff implementation, instead, we compute a lower bound of 10.4 GB, which is approximately reflected in the empirical measurements (maximum consumption is almost 13 GB). Note: memory consumption for the autodiff case is reported in GiB, effectively making the scale of the plot one order of magnitude higher then in the relative gradient plot.

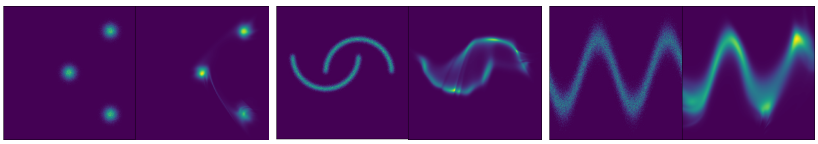


Figure D.3: 2D toy examples trained with SGD. True distribution on the left, predicted densities on the right.

D.8.3 Density estimation

Architecture Although mentioned all throughout the paper, let us recall the neural network used for these experiments. We here rely on the usual feedforward architecture, that is, a neural network for which the input is sequentially passed through an interleaving series of matrix multiplications and non-linear activation functions, being the last operation a matrix multiplication.

Nonlinearities Note that, since we make use of square weight matrices, the only two hyperparameters left in our architecture are the number of layers in the network, L, and the non-linearity used. We consider two types of non-linearities. First, a smooth version of the leaky-ReLU activation function with a hyperparameter α,

$$s_L(x) = \alpha x + (1 - \alpha)\log(1 + e^x). \tag{D.19}$$

Second, a weighted sum of the identity and hyperbolic tangent functions with two hyperparameters, α and β, controlling the steepness and "level

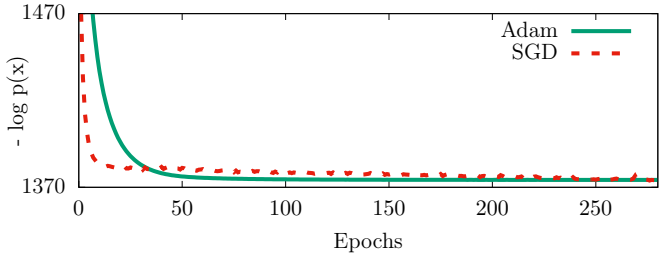


Figure D.4: Log-likelihood evolution on MNIST validation set.

of linearity" of the activation function,

$$s_T(x) = \tanh(\alpha x) + \beta x. \tag{D.20}$$

However, in our experiments, these two hyperparameters for the s_T nonlinearity are fixed to $\alpha = 1$ and $\beta = 0.1$ always. Both of these nonlinearities are relatively smooth, and while no closed form solution for their inverse is available they can be inverted easily with a Newton method; in practice, for our parameter choice, we use a fixed number of 100 iterations which seems to be (way) more than sufficient.

Toy examples For all the experiments shown in figure 5.2 of the main paper, we always use Adam as optimizer, fix the batch size and number of layers L to 100, use biases, and fix the activation function to s_L with $\alpha = 0.3$. We chose as base distribution (that is, the distribution of the latent variables) the standard normal distribution. We plot, as in the quantitative experiments, the best model found during the training. Regarding the data, we sampled five-thousand samples for the training set and five-hundred points for the test set. The only changing hyperparameters across the figures is the learning rate and the number of epochs, which are summarised in table D.1.

	MoG	half moons	sine
learning rate	0.001	0.001	0.005
no. of epochs	2000	1300	4000

Table D.1: Hyperparameters used for figure 2 of the main paper.

Quantitative results on MNIST To obtain the density results on the MNIST dataset, the same preprocessing as in [247] has been applied. Note that we do not include the contribution due to this preprocessing in the reported log-likelihood values. [1] For the model architecture, we fixed the number of layers to 2. Note that competing models reported in table D.3 of the main paper are taken from [247] and employ a higher number of parameters. We used the smooth Leaky-ReLU (D.19) with $\alpha = 0.01$ and a standard normal distribution as a distribution for the latent variables. The optimization has been performed with Adam with default parameters. The hyperparameters search has been performed over learning rate values of 0.001, 0.0005, 0.0001 and batch sizes of 10, 100. For each run, we selected the model whose performance did not improve in the successive 30 epochs of training (i.e. we chose the model at epoch 10 if all the values of the loss for epochs 11 to 40 were higher then the value after 10 epochs). The best hyperparameters selection is shown in table D.3.

1: We thank T. Anderson Keller and Emiel Hoogeboom for pointing this out.

Convergence time on MNIST To get an idea of the running time of our method in a real-world scenario, one epoch on MNIST ($n = 784$, 50k training samples) on a modern laptop CPU takes an order of tens of seconds, a $\sim 4.5\times$ speedup compared to "standard" optimization (which is roughly consistent with figure D.1, which was obtained with a slightly different experimental setup) and $\sim 50\times$ speedup with respect to "autodiff". Our convergence time is ~ 15 min. While the speed-up is

already visible at this data dimensionality, the difference is expected to be larger at higher dimensionality.

Quantitative results First, we want to remark that the data used for the experiments shown in table 5.1 was pre-processed in the exact same way as described in [247].

For the results shown in such table (MNIST excluded) a more exhaustive hyperparameter search has been performed. Particularly, for each dataset a grid-search was run with the options shown in table D.2, taking for each experiment the model with best validation log-likelihood obtained during training and, across experiments, getting the one with best test log-likelihood. Experiments were again trained using Adam and, instead of fixing the number of epochs, training was finished with an early-stopping criteria that evaluates the validation set every 25 epochs and has a patience of 5 trials. The best hyperparameters selection is shown in table D.3.

	Option #1	Option #2	Option #3
activation	$s_L, \alpha = 0.3$	$s_L, \alpha = 0.01$	s_T
no. layers	25	50	100
learning rate	0.001	0.0005	0.0001
batch size	10	50	100
base distribution	standard normal	hyperbolic secant	
bias	Yes	No	

Table D.2: Hyperparameters considered for the grid search.

	POWER	GAS	HEPMASS	MINIBOONE	BSDS300	MNIST
activation	$s_L, \alpha = 0.3$	$s_L, \alpha = 0.3$	$s_L, \alpha = 0.3$	s_T	s_T	$s_L, \alpha = 0.01$
no. layers	50	100	50	25	25	2
learning rate	0.001	0.001	0.001	0.0001	0.0001	0.0001
batch size	100	100	50	100	100	10
base dist.	std normal	std normal	hyper. secant	std normal	hyper. secant	std normal
bias	Yes	Yes	No	Yes	No	Yes

Table D.3: Hyperparameters for the results in table 1 in the main paper.

Regarding the rest of the models shown in that table, we reproduce the exact same experiments as those described in [247]. Therefore, the considered models have the same architecture and stopping criteria as the ones shown in table 1 of the aforementioned paper. The only difference with respect to the results shown in table 1 of [247] and table 5.1 in our paper is the number of trainable parameters. As mentioned in section 5.5, in order to perform a fair comparison, we tweaked the hyperparameters of each architecture so they have approximately the same number of parameters.

Specifically, we first trained our model as described above and, once we knew the number of parameters of the best-performing model (which is approximately Ln^2) we used the formulae shown in table 3 of [247] to find to which values we should fix the hyperparameters L and H of their models so that they have the same number of parameters.

As a final remark, note that there is one degree-of-freedom in those equations (for every L there is a value of H solving the given equation). Therefore, for each of the considered models and datasets, we run two different experiments, one with $L = 1$ and another with $L = 2$ (as similarly done in [247]), finding afterwards the proper value of H to

match the number of trainable parameters of our best model for that same dataset.

E

Additional Material on Chapter 6

E.1 Why does classification result in the log ratio?

Let us suppose that a variable X is drawn with equal probability from two distributions P_0 and P_1 with densities $p_0(x)$ and $p_1(x)$ respectively. We train a classifier $D : x \mapsto [0, 1]$ to estimate the posterior probability that a particular realization of X was drawn from P_0 with the cross entropy loss, i.e. the parameters of D are chosen to minimize

$$L(D) = \mathbb{E}_{X\sim P_0}\left[-\log D(X)\right] + \mathbb{E}_{X\sim P_1}\left[-\log(1 - D(X))\right].$$

As shown in, for instance, [82], the global optimum of this loss occurs when $D(x) = \frac{p_0(x)}{p_0(x)+p_1(x)}$, which can be rewritten as

[82]: Goodfellow et al. (2014), 'Generative adversarial nets'

$$D(x) = \frac{1}{1 + p_1(x)/p_0(x)} \tag{E.1}$$

$$= \frac{1}{1 + \exp(-\log(p_0(x)/p_1(x)))} \tag{E.2}$$

$$\tag{E.3}$$

Recall that in our setting, the function $r(x_1, x_2)$ is trained to classify between the two cases that (x_1, x_2) is drawn from the joint distribution $\mathbb{P}_{x_1,x_2}$ (*class* 0) or the product of marginals $\mathbb{P}_{x_1}\mathbb{P}_{x_2}$ (*class* 1). $r(x_1, x_2)$ is trained so that $\frac{1}{1+\exp(-r(x_1,x_2))}$ estimates the posterior probability of (x_1, x_2) belonging to class 0. By comparing to Equation E.2, it can be seen that

$$\begin{aligned} r(x_1, x_2) &= \log\left(p(x_1, x_2)/p(x_1)p(x_2)\right) \\ &= \log p(x_1|x_2) - \log p(x_1) \\ &= \log p(x_2|x_1) - \log p(x_2) \end{aligned}$$

Note that in order for the classification trick of contrastive learning to be useful, the variables x_1 and x_2 cannot be deterministically related. If this is the case, the log-ratio is everywhere either 0 or ∞ and hence the learnt features are not useful.

To see why this is the case, suppose that x_1, and x_2 are each N-dimensional vectors. If they are deterministically related, $p(x_1, x_2)$ puts mass on an N-dimensional submanifold of a $2N$-dimensional space. On the other

hand, $p(x_1)p(x_2)$ will put mass on a $2N$-dim manifold since it is the product of two distributions each of which are N-dimensional.

In this case, the distributions $p(x_1, x_2)$ and $p(x_1)p(x_2)$ are therefore not absolutely continuous with respect to one another and thus the log-ratio is ill-defined: $p(x_1, x_2)/p(x_1)p(x_2) = \infty$ at any point (x_1, x_2) at which $p(x_1, x_2)$ puts mass and zero at points where $p(x_1)p(x_2)$ puts mass and $p(x_1, x_2)$ does not.

E.2 The Sufficiently Distinct Views Assumption

We give the following two examples to provide intuition about the Sufficiently Distinct Views (SDV) assumption - one regarding a case in which it does not hold, and another one in which it does.

A simple case in which the assumption does not hold is when the conditional probability of $\boldsymbol{z}$ given $\boldsymbol{s}$ is Gaussian, as in

$$p(\boldsymbol{z}|\boldsymbol{s}) = \frac{1}{Z} \exp\left[-\sum_i (z_i - s_i)^2/(2\sigma_i^2)\right], \tag{E.4}$$

where Z is the normalization factor, $Z = (2\pi)^{n/2} \prod_i \sigma_i$. Since taking second derivatives of the log-probability with respect to s_i results in constants, it can be easily shown that there is no way to find $2D$ vectors $\boldsymbol{z}_j$, $j = 1, \ldots, 2D$, such that the corresponding $\boldsymbol{w}(\boldsymbol{s}, \boldsymbol{z}_j)$ (see Definition 1) are linearly independent.

The fact that the assumption breaks down in this case is reminiscent of the breakdown in the case of Gaussianity for linear ICA. Interestingly, in our work, the true latent sources **are** allowed to be Gaussian. In fact, the distribution of $\boldsymbol{s}$ does not enter the expression above.

An example in which the SDV assumption does hold is a conditional pdf given by

$$p(\boldsymbol{z}|\boldsymbol{s}) = \frac{1}{Z(\boldsymbol{s})} \exp\left[-\sum_i (z_i^2 s_i^2 + z_i^4 s_i^4)\right], \tag{E.5}$$

where $Z(\boldsymbol{s})$ is again a normalization function. Proving that this distribution satisfies the SDV assumption requires a few lines of computation. The idea is that $\boldsymbol{w}(\boldsymbol{s}, \boldsymbol{z})$ can be written as the product of a matrix and vector which are functions only of $\boldsymbol{s}$ and $\boldsymbol{z}$ respectively. Once written in this form, it is straightforward to show that the columns of the matrix are linearly independent for almost all values of $\boldsymbol{s}$ and that $2D$ linearly independent vectors can be realized by different choices of $\boldsymbol{z}$.

[66]: Hyvärinen et al. (2019), 'Nonlinear ICA using auxiliary variables and generalized contrastive learning'

E.3 Proof of Theorem 6.2.1 and corollary 6.2.2

E.3.1 Proof of Theorem 6.2.1

This proof is mainly inspired by the techniques employed by [66].

Proof. We have to show that, upon convergence, $h_i(\boldsymbol{x}_1)$ are s.t.

$$h_i(\boldsymbol{x}_1) \perp\!\!\!\perp h_j(\boldsymbol{x}_1), \forall i \neq j$$

We start by writing the difference in log-densities of the two classes:

$$\sum_i \psi_i(h_i(\boldsymbol{x}_1), \boldsymbol{x}_2) = \sum_i \alpha_i(\boldsymbol{f}_{1,i}^{-1}(\boldsymbol{x}_1), \boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2)) + \\ - \sum_i \delta_i(\boldsymbol{f}_{2,i}^{-1}(\boldsymbol{x}_2))$$

We now make the change of variables

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{h}(\boldsymbol{x}_1) \\ \boldsymbol{v}(\boldsymbol{y}) &= \boldsymbol{f}_1^{-1}(\boldsymbol{h}^{-1}(\boldsymbol{y})) \\ \boldsymbol{t} &= \boldsymbol{f}_2^{-1}(\boldsymbol{x}_2)) \end{aligned}$$

and rewrite the first equation in the following form:

$$\sum_i \psi_i(y_i, \boldsymbol{x}_2) = \sum_i \alpha_i(v_i(\boldsymbol{y}), t_i) \tag{E.6}$$

$$- \sum_i \delta_i(t_i) \tag{E.7}$$

We take derivatives with respect to y_j, $y_{j'}$, $j \neq j'$, of the LHS and RHS of equation E.15. Adopting the conventions in 6.8 and 6.9 and

$$v_i^j(\boldsymbol{y}) = \partial v_i(\boldsymbol{y}) / \partial y_j \tag{E.8}$$

$$v_i^{jj'}(\boldsymbol{y}) = \partial^2 v_i(\boldsymbol{y}) / \partial y_j \partial y_{j'}\,, \tag{E.9}$$

we have

$$\sum_i \alpha_i''(v_i(\boldsymbol{y}), t_i) v_i^j(\boldsymbol{y}) v_i^{j'}(\boldsymbol{y}) \\ + \alpha_i'(v_i(\boldsymbol{y}), t_i) v^{jj'}(\boldsymbol{y}) = 0\,,$$

where taking derivative w.r.t. y_j and y_j' for $j \neq j'$ makes LHS equal to zero, since the LHS has functions which depend only one y_i each. If we now rearrange our variables by defining vectors $\boldsymbol{a}_i(\boldsymbol{y})$ collecting all entries $v_i^j(\boldsymbol{y}) v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, and vectors $\boldsymbol{b}_i(\boldsymbol{y})$ with the variables $v_i^j(\boldsymbol{y}) v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, the above equality can be rewritten as

$$\sum_i \alpha_i''(v_i(\boldsymbol{y}), t_i) \boldsymbol{a}_i(\boldsymbol{y}) \\ + \alpha_i'(v_i(\boldsymbol{y}), t_i)) \boldsymbol{b}_i(\boldsymbol{y}) = 0\,.$$

The above expression can be recast in matrix form,

$$\boldsymbol{M}(\boldsymbol{y})\boldsymbol{w}(\boldsymbol{y},\boldsymbol{t}) = 0\,,$$

where $\boldsymbol{M}(\boldsymbol{y}) = (\boldsymbol{a}_1(\boldsymbol{y}), \ldots, \boldsymbol{a}_n(\boldsymbol{y}), \boldsymbol{b}_1(\boldsymbol{y}), \ldots, \boldsymbol{b}_n(\boldsymbol{y}))$ and $\boldsymbol{w}(\boldsymbol{y},\boldsymbol{t}) = (\alpha''_1, \ldots, \alpha''_n, \alpha'_1, \ldots, \alpha'_n)$. $\boldsymbol{M}(\boldsymbol{y})$ is therefore a $n(n-1)/2 \times 2n$ matrix, and $\boldsymbol{w}(\boldsymbol{y},\boldsymbol{t})$ is a $2n$ dimensional vector.

To show that $\boldsymbol{M}(\boldsymbol{y})$ is equal to zero, we invoke the SDV assumption. This implies the existence of $2n$ linearly independent $\boldsymbol{w}(\boldsymbol{y},\boldsymbol{t}_j)$. It follows that

$$\boldsymbol{M}(\boldsymbol{y})[\boldsymbol{w}(\boldsymbol{y},\boldsymbol{t}_1), \ldots, \boldsymbol{w}(\boldsymbol{y},\boldsymbol{t}_{2n})] = 0\,,$$

and hence $\boldsymbol{M}(\boldsymbol{y})$ is zero by elementary linear algebraic results. It follows that $v_i^j(\boldsymbol{y}) \neq 0$ for at most one value of j, since otherwise the product of two non-zero terms would appear in one of the entries of $\boldsymbol{M}(\boldsymbol{y})$, thus rendering it non-zero. Thus v_i is a function only of one y_j.

Observe that $\boldsymbol{v}(\boldsymbol{y}) = \boldsymbol{s}$. We have just proven that $v_i(y_{\pi(i)}) = s_i$. Since v_i is invertible, it follows that $h_{\pi(i)}(\boldsymbol{x}_1) = y_{\pi(i)} = v_i^{-1}(s_i)$ and hence the components of $\boldsymbol{h}(\boldsymbol{x}_1)$ recover the components of $\boldsymbol{s}$ up to the invertible component-wise ambiguity given by $\boldsymbol{v}$, and the permutation ambiguity.

□

E.3.2 Proof of Corollary 6.2.2

Proof. This follows exactly by repeating the proof of Theorem 6.2.1 where the roles of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are exchanged and the regression function in the statement of the corollary is used. □

E.4 Proof of Theorems 6.2.3 and 6.2.4

Theorem 6.2.3 is a special case of Theorem 6.2.4 by considering the case $\boldsymbol{g}_1(\boldsymbol{s},\boldsymbol{n}_1) = \boldsymbol{s}$. We therefore prove only the more general Theorem 6.2.4.

Proof. We have to show that, upon convergence, $h_i(\boldsymbol{x}_1)$ and $k_i(\boldsymbol{x}_2)$ are such that

$$h_{1,i}(\boldsymbol{x}_1) \perp\!\!\!\perp h_{1,j}(\boldsymbol{x}_1), \forall i \neq j \tag{E.10}$$
$$h_{2,i}(\boldsymbol{x}_2) \perp\!\!\!\perp h_{2,j}(\boldsymbol{x}_2), \forall i \neq j \tag{E.11}$$
$$h_{1,i}(\boldsymbol{x}_1) \perp\!\!\!\perp h_{2,j}(\boldsymbol{x}_2), \forall i \neq j. \tag{E.12}$$

We start by exploiting Equations 6.13 and 6.14 to write the difference in log-densities of the two classes

$$\begin{aligned}
&\sum_i \psi_i(h_{1,i}(\boldsymbol{x}_1), h_{2,i}(\boldsymbol{x}_2)) \\
&= \sum_i \eta_i(f_{1,i}^{-1}(\boldsymbol{x}_1), f_{2,i}^{-1}(\boldsymbol{x}_2)) - \sum_i \theta_i(f_{1,i}^{-1}(\boldsymbol{x}_1)) &\text{(E.13)}\\
&= \sum_i \lambda_i(f_{2,i}^{-1}(\boldsymbol{x}_2), f_{1,i}^{-1}(\boldsymbol{x}_1)) - \sum_i \mu_i(f_{2,i}^{-1}(\boldsymbol{x}_2)) &\text{(E.14)}
\end{aligned}$$

We now make the change of variables

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{h}_1(\boldsymbol{x}_1) \\ \boldsymbol{t} &= \boldsymbol{h}_2(\boldsymbol{x}_2) \\ \boldsymbol{v}(\boldsymbol{y}) &= \boldsymbol{f}_1^{-1}(\boldsymbol{h}_1^{-1}(\boldsymbol{y})) \\ \boldsymbol{u}(\boldsymbol{t}) &= \boldsymbol{f}_2^{-1}(\boldsymbol{h}_2^{-1}(\boldsymbol{t})) \end{aligned}$$

and rewrite equation E.13 in the following form:

$$\begin{aligned} &\sum_i \psi_i(y_i, t_i) \\ &= \sum_i \eta_i(v_i(\boldsymbol{y}), u_i(\boldsymbol{t})) - \sum_i \theta_i(v_i(\boldsymbol{y})) \end{aligned} \tag{E.15}$$

We first want to prove the condition in Equation E.10. We will show this is true by proving that

$$v_i(\boldsymbol{y}) \equiv v_i(y_{\pi(i)}) \tag{E.16}$$

for some permutation of the indices π with respect to the indexing of the sources $\boldsymbol{s} = (s_1, \ldots, s_D)$.

We take derivatives with respect to y_j, $y_{j'}$, $j \neq j'$, of the LHS and RHS of equation E.15, yielding

$$\begin{aligned} &\sum_i \eta_i''(v_i(\boldsymbol{y}), u_i(\boldsymbol{t}))v_i^j(\boldsymbol{y})v_i^{j'}(\boldsymbol{y}) \\ &+ \sum_i \eta_i'(v_i(\boldsymbol{y}), u_i(\boldsymbol{t}))v^{jj'}(\boldsymbol{y}) = 0 \end{aligned}$$

If we now rearrange our variables by defining vectors $\boldsymbol{a}_i(\boldsymbol{y})$ collecting all entries $v_i^j(\boldsymbol{y})v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, and vectors $\boldsymbol{b}_i(\boldsymbol{y})$ with the variables $v_i^j(\boldsymbol{y})v_i^{j'}(\boldsymbol{y})$, $j = 1, \ldots, n$, $j' = 1, \ldots, j-1$, the above equality can be rewritten as

$$\begin{aligned} &\sum_i \eta_i''(v_i(\boldsymbol{y}), u_i(\boldsymbol{t}))\boldsymbol{a}_i(\boldsymbol{y}) \\ &+ \eta_i'(v_i(\boldsymbol{y}), u_i(\boldsymbol{t}))\boldsymbol{b}_i(\boldsymbol{y}) = 0\,. \end{aligned}$$

Again following [66], we recast the above formula in matrix form,

$$\boldsymbol{M}(\boldsymbol{y})\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}) = 0\,, \tag{E.17}$$

where $\boldsymbol{M}(\boldsymbol{y}) = (\boldsymbol{a}_1(\boldsymbol{y}), \ldots, \boldsymbol{a}_n(\boldsymbol{y}), \boldsymbol{b}_1(\boldsymbol{y}), \ldots, \boldsymbol{b}_n(\boldsymbol{y}))$ and $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}) = (\eta_1'', \ldots, \eta_n'', \eta_1', \ldots, \eta_n')$. $\boldsymbol{M}(\boldsymbol{y})$ is therefore a $n(n-1)/2 \times 2n$ matrix, and $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t})$ is a $2n$ dimensional vector.

To show that $\boldsymbol{M}(\boldsymbol{y})$ is equal to zero, we invoke the SDV assumption on $\boldsymbol{\eta}$. This implies the existence of $2n$ linearly independent $\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_j)$. It follows that

$$\boldsymbol{M}(\boldsymbol{y})[\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_1), \ldots, \boldsymbol{w}(\boldsymbol{y}, \boldsymbol{t}_{2n})] = 0\,,$$

and hence $\boldsymbol{M}(\boldsymbol{y})$ is zero by elementary linear algebraic results. It follows that $v_i^j(\boldsymbol{y}) \neq 0$ for at most one value of j, since otherwise the product of

two non-zero terms would appear in one of the entries of $\boldsymbol{M}(\boldsymbol{y})$, thus rendering it non-zero. Thus v_i is a function only of one $y_j = y_{\pi(i)}$.

Observe that $\boldsymbol{v}(\boldsymbol{y}) = \boldsymbol{s}$. We have just proven that $v_i(y_{\pi(i)}) = s_i$. Since v_i is invertible, it follows that $h_{\pi(i)}(\boldsymbol{x}_1) = y_{\pi(i)} = v_i^{-1}(s_i)$ and hence the components of $\boldsymbol{h}(\boldsymbol{x}_1)$ recover the components of $\boldsymbol{s}$ up to the invertible component-wise ambiguity given by $\boldsymbol{v}$, and the permutation ambiguity.

For the condition in Equation E.11, we need

$$u_i(\boldsymbol{t}) \equiv u_i(t_{\tilde{\pi}(i)}) \,, \tag{E.18}$$

where the permutation $\tilde{\pi}$ doesn't need to be equal to π. By symmetry, exactly the same argument as used to prove the condition in Equation E.16 holds, by replacing $(\boldsymbol{v}, \boldsymbol{y}, \boldsymbol{\eta}, \boldsymbol{\theta})$ with $(\boldsymbol{u}, \boldsymbol{t}, \boldsymbol{\lambda}, \boldsymbol{\mu})$, noting that the SDV assumption is also assumed for $\boldsymbol{\lambda}$.

We have shown that $\boldsymbol{y} = \boldsymbol{h}_1(\boldsymbol{x}_1)$ and $\boldsymbol{t} = \boldsymbol{h}_2(\boldsymbol{x}_2)$ estimate $\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)$ and $\boldsymbol{g}_2(\boldsymbol{s}, \boldsymbol{n}_2)$ up to two different gauges of all possible scalar invertible functions.

A remaining ambiguity could be that the two representations might be misaligned; that is, defining $\boldsymbol{z}_1 = \boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1)$ and $\boldsymbol{z}_2 = \boldsymbol{g}_2(\boldsymbol{s}, \boldsymbol{n}_2)$, while

$$z_{1,i} \perp\!\!\!\perp z_{2,j} \forall i \neq j \tag{E.19}$$

we might have

$$y_{\pi(i)} \perp\!\!\!\perp t_{\tilde{\pi}(j)} \forall i \neq j \,,$$

where $\pi(i), \tilde{\pi}(i)$ are two different permutations of the indices $i = 1, \ldots, n$. We want to show that this ambiguity is also resolved; that means, our goal is to show that

$$y_i \perp\!\!\!\perp t_j, \;\; \forall i \neq j \tag{E.20}$$

We recall that, by definition, we have $v_i(y_{\pi(i)}) = z_{1,i}$ and $u_j(t_{\tilde{\pi}(j)}) = z_{2,j}$. Then, due to equation E.19,

$$v_i(y_{\pi(i)}) \perp\!\!\!\perp u_j(t_{\tilde{\pi}(j)}) \;\; \forall i \neq j \tag{E.21}$$

$$\implies y_{\pi(i)} \perp\!\!\!\perp t_{\tilde{\pi}(j)} \;\; \forall i \neq j \tag{E.22}$$

$$\implies y_i \perp\!\!\!\perp t_{\tilde{\pi} \circ \pi^{-1}(j)} \;\; \forall i \neq j \,, \tag{E.23}$$

where the implication E.21-E.22 follows from invertibility of v_i and u_j, and the implication E.22-E.23 follows from considering that, given that we know E.22, we can define $l = \pi(j)$ and $k = \pi(i)$ and have

$$y_k \perp\!\!\!\perp t_{\tilde{\pi} \circ \pi^{-1}(l)} \;\; \forall k \neq l.$$

Define

$$\tau = \tilde{\pi} \circ \pi^{-1}$$

and note that it is a permutation. Then

$$y_i \perp\!\!\!\perp t_{\tau(j)} \forall i \neq j \tag{E.24}$$

Fix any particular i. Our goal is to show that for any $j \neq i$ the independence relation in Equation E.20 holds. There are two possibilities:

i $\tau(i) = i$
ii $\tau(i) \neq i$

In the first case, τ restricted to the set $\{1, \dots, D\} \setminus \{i\}$ is still a permutation, and thus considering the independences of Equation E.24 for all $j \neq i$ implies each of the independences of Equation E.20 and we are done.

Let us consider the second case. Then,

$$\exists l \in \{1, \dots, D\} \setminus \{i\} \text{ s.t. } l = \tau(i)\,.$$

We then need to prove

$$y_i \perp\!\!\!\perp t_l\,, \tag{E.25}$$

which is the only independence implied by Equation E.20 which is not implied by Equation E.24.

In order to do so, we rewrite equation E.15, yielding

$$\begin{aligned}&\sum_m \psi_m(y_m, t_m)\\ =&\sum_m \eta_m(v_m(y_{\pi(m)}), u_m(t_{\tilde{\pi}(m)})) - \sum_m \theta_i(v_m(y_{\pi(m)}))\end{aligned} \tag{E.26}$$

We now take derivative with respect to y_i and t_l in E.25; noting that $\tilde{\pi}^{-1}(l) = \pi^{-1}(i)$, we get

$$\begin{aligned}0 =&\frac{\partial^2}{\partial v_{\pi^{-1}(i)} \partial u_{\pi^{-1}(i)}} \eta_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i), u_{\pi^{-1}(i)}(t_l))\\ &\times \frac{\partial}{\partial y_i} v_{\pi^{-1}(i)}(y_i) \frac{\partial}{\partial t_l} u_{\pi^{-1}(i)}(t_l)\end{aligned} \tag{E.27}$$

Since $v_{\pi^{-1}(i)}(y_i)$ is a smooth and invertible function of its argument, the set of y_i such that $\frac{\partial}{\partial y_i} v_{\pi^{-1}(i)}(y_i) = 0$ has measure zero. Similarly, $\frac{\partial}{\partial t_l} u_{\pi^{-1}(i)}(t_l) = 0$ on a set of measure zero.

It therefore follows that

$$\frac{\partial}{\partial y_i} v_{\pi^{-1}(i)}(y_i) \frac{\partial}{\partial t_l} u_{\pi^{-1}(i)}(t_l) \neq 0$$

almost everywhere and hence that

$$\frac{\partial^2}{\partial v_{\pi^{-1}(i)} \partial u_{\pi^{-1}(i)}} \eta_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i), u_{\pi^{-1}(i)}(t_l)) = 0\,. \tag{E.28}$$

almost everywhere. We can thus conclude that

$$\begin{aligned}&\eta_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i), u_{\pi^{-1}(i)}(t_l)) =\\ &\eta^y_{\pi^{-1}(i)}(v_{\pi^{-1}(i)}(y_i)) + \eta^t_{\pi^{-1}(i)}(u_{\pi^{-1}(i)}(t_l))\end{aligned}$$

This in turn implies that, for some functions A and B, we can write

$$\begin{aligned}&\log p(z_{1,\pi^{-1}(i)} | z_{2,\pi^{-1}(i)}) - \log p(z_{1,\pi^{-1}(i)})\\ &= A(v_{\pi^{-1}(i)}(y_i)) + B(u_{\pi^{-1}(i)}(t_l))\end{aligned}$$

and therefore

$$\log p(z_{1,\pi^{-1}(i)}, z_{2,\pi^{-1}(i)}) = C(v_{\pi^{-1}(i)}(y_i)) + D(u_{\pi^{-1}(i)}(t_l))$$

for some functions C and D. This decomposition of the log-pdf implies

$$\begin{aligned} & z_{1,\pi^{-1}(i)} \perp\!\!\!\perp z_{2,\pi^{-1}(i)} \\ \implies & z_{1,\pi^{-1}(i)} \perp\!\!\!\perp z_{2,\tilde{\pi}^{-1}(l)} \\ \implies & v_{\pi^{-1}(i)}(y_i) \perp\!\!\!\perp u_{\tilde{\pi}^{-1}(l)}(t_l) \\ \implies & y_i \perp\!\!\!\perp t_l \,, \end{aligned}$$

where the last implication holds due to invertibility of $v_{\pi^{-1}(i)}$ and $u_{\tilde{\pi}^{-1}(l)}$. We have thus concluded the proof.

□

E.5 Proof of Corollary 6.2.5

Proof. Denoting by $\boldsymbol{d}_1^{(k)}$ the component-wise invertible ambiguity up to which $\boldsymbol{g}(\boldsymbol{s}, \boldsymbol{n}_1^{(k)})$ is recovered, we have that

$$\inf_{\boldsymbol{e}\in\boldsymbol{E}} \mathbb{E}_{\boldsymbol{x}_1}\left[\left\|\boldsymbol{s} - \boldsymbol{e}(\boldsymbol{h}_1^{(k)}(\boldsymbol{x}_1))\right\|_2^2\right] \tag{E.29}$$

$$= \inf_{\boldsymbol{e}\in\boldsymbol{E}} \mathbb{E}_{(\boldsymbol{n}_1^{(k)},\boldsymbol{s})}\left[\left\|\boldsymbol{s} - \boldsymbol{e}\circ\boldsymbol{d}_1^{(k)}\circ\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1^{(k)})\right\|_2^2\right] \tag{E.30}$$

$$= \inf_{\tilde{\boldsymbol{e}}\in\boldsymbol{E}} \mathbb{E}_{(\boldsymbol{n}_1^{(k)},\boldsymbol{s})}\left[\left\|\boldsymbol{s} - \tilde{\boldsymbol{e}}\circ\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1^{(k)})\right\|_2^2\right] \tag{E.31}$$

$$\leq \mathbb{E}_{(\boldsymbol{n}_1^{(k)},\boldsymbol{s})}\left[\left\|\boldsymbol{s} - \boldsymbol{e}^*\circ\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1^{(k)})\right\|_2^2\right] \tag{E.32}$$

The lower bound holds for any $\boldsymbol{e}^* \in \boldsymbol{E}$ by definition of infimum and in particular for $\boldsymbol{e}^* = \boldsymbol{g}_1|_{\boldsymbol{n}=0}^{-1}$, the existence of which is guaranteed by the assumptions on $\boldsymbol{g}_1$. Taking a Taylor expansion of $\boldsymbol{e}^*\circ\boldsymbol{g}_1(\boldsymbol{s}, \boldsymbol{n}_1^{(k)})$ around $\boldsymbol{n}_1^{(k)} = 0$ yields

$$\begin{aligned} &\mathbb{E}_{(\boldsymbol{n}_1^{(k)},\boldsymbol{s})}\Bigg[\Bigg\|\boldsymbol{s} - \boldsymbol{e}^*\circ\boldsymbol{g}_1(\boldsymbol{s}, 0) \\ &\quad + \frac{\partial \boldsymbol{e}^*}{\partial \boldsymbol{g}_1}\frac{\partial \boldsymbol{g}_1(\boldsymbol{s}, 0)}{\partial \boldsymbol{n}_1^{(k)}}\cdot\boldsymbol{n}_1^{(k)} + \mathcal{O}(\|\boldsymbol{n}_1^{(k)}\|^2)\Bigg\|_2^2\Bigg] \\ &= \mathbb{E}_{(\boldsymbol{n}_1^{(k)},\boldsymbol{s})}\left[\left\|\frac{\partial \boldsymbol{e}^*}{\partial \boldsymbol{g}_1}\frac{\partial \boldsymbol{g}_1(\boldsymbol{s}, 0)}{\partial \boldsymbol{n}_1^{(k)}}\cdot\boldsymbol{n}_1^{(k)} + \mathcal{O}(\|\boldsymbol{n}_1^{(k)}\|^2)\right\|_2^2\right] \\ &\longrightarrow 0 \text{ as } k \longrightarrow \infty \end{aligned}$$

where the last equality follows from fact that $\boldsymbol{e}^* = \boldsymbol{g}|_{\boldsymbol{n}=0}^{-1}$ and the convergence follows from the fact that $\boldsymbol{n}_1^{(k)} \longrightarrow 0$ as $k \to \infty$. □

E.6 Proof of Lemma 6.2.6

We will make crucial use of *Kolmogorov's strong law*:

Theorem E.6.1 *Suppose that X_n is a sequence of independent (but not necessarily identically distributed) random variables with*

$$\sum_{n=1}^{\infty} \frac{1}{n^2} \mathrm{Var}[X_n] < \infty$$

Then,

$$\frac{1}{N} \sum_{n=1}^{N} X_n - \mathbb{E}[X_n] \xrightarrow{a.s.} 0$$

Fix $\boldsymbol{s}$ and consider $\Omega_{\mathbf{e}}^N(\boldsymbol{s}, \boldsymbol{n})$ as a random variable with randomness induced by $\boldsymbol{n}$. We will show that for almost all $\boldsymbol{s}$ this converges $\boldsymbol{n}$-almost surely to a constant, and hence $\Omega_{\mathbf{e}}^N(\boldsymbol{s}, \boldsymbol{n})$ converges almost surely to a function of $\boldsymbol{s}$.

The law of total expectation says that

$$\begin{aligned}
&\mathrm{Var}_{\boldsymbol{s},\boldsymbol{n}_i}[\boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s} + \boldsymbol{n}_i)] \\
&= \mathbb{E}_{\boldsymbol{s}}\left[V_i(\boldsymbol{s})\right] + \mathrm{Var}_{\boldsymbol{s}}\left[\mathbb{E}_{\boldsymbol{n}_i}[\boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s} + \boldsymbol{n}_i)]\right] \\
&\geq \mathbb{E}_{\boldsymbol{s}}\left[V_i(\boldsymbol{s})\right].
\end{aligned}$$

Since by assumption $\mathrm{Var}_{\boldsymbol{s},\boldsymbol{n}_i}[\boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s} + \boldsymbol{n}_i)] \leq K$, we have that

$$\mathbb{E}_{\boldsymbol{s}}\left[\sum_{i=1}^{\infty} \frac{V_i(\boldsymbol{s})}{i^2}\right] \leq \frac{K\pi^2}{6}$$

and therefore $\sum_{i=1}^{\infty} \frac{V_i(\boldsymbol{s})}{i^2} < \infty$ with probability 1 over $\boldsymbol{s}$, else the expectation above would be unbounded since $V_i(\boldsymbol{s}) \geq 0$.

We have further that for almost all $\boldsymbol{s}$,

$$\Omega_e(\boldsymbol{s}) = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} E_{e_i}(\boldsymbol{s})$$

exists. Therefore, for almost all s the conditions of Kolmogorov's strong law are met by $\Omega_e^N(\boldsymbol{s}, \boldsymbol{n})$ and so

$$\Omega_e^N(\boldsymbol{s}, \boldsymbol{n}) - \mathbb{E}_{\boldsymbol{n}}[\Omega_e^N(\boldsymbol{s}, \boldsymbol{n})] \xrightarrow{\boldsymbol{n}-a.s.} 0$$

Since $\mathbb{E}_{\boldsymbol{n}}[\Omega_e^N(\boldsymbol{s}, \boldsymbol{n})] \xrightarrow{\boldsymbol{n}-a.s.} \Omega_e(\boldsymbol{s})$, it follows that

$$\Omega_e^N(\boldsymbol{s}, \boldsymbol{n}) \xrightarrow{\boldsymbol{n}-a.s.} \Omega_e(\boldsymbol{s}).$$

Since this holds with probability 1 over $\boldsymbol{s}$, we have that

$$\Omega_e^N(\boldsymbol{s}, \boldsymbol{n}) \xrightarrow{\boldsymbol{n}-a.s.} \Omega_e(\boldsymbol{s}).$$

It follows that we can write

$$R^N_{\boldsymbol{e},i}(\boldsymbol{s},\boldsymbol{n}) = \boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s}+\boldsymbol{n}_i) - \Omega^N_{\boldsymbol{e}}(\boldsymbol{s},\boldsymbol{n})$$
$$\xrightarrow{a.s.} R_{\boldsymbol{e},i}(\boldsymbol{s},\boldsymbol{n}_i) := \boldsymbol{e}_i \circ \boldsymbol{k}_i(\boldsymbol{s}+\boldsymbol{n}_i) - \Omega_{\boldsymbol{e}}(\boldsymbol{s})$$

E.7 Proof of Theorem 6.2.7

We will begin by showing that if $K \geq \mathrm{Var}(\boldsymbol{s}) + C$ then $\{\boldsymbol{k}_i^{-1}\} \in \mathcal{G}_K$.

For $\boldsymbol{e}_i = \boldsymbol{k}_i^{-1}$, we have that

$$\Omega^N_{\boldsymbol{e}}(\boldsymbol{s},\boldsymbol{n}) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{s}+\boldsymbol{n}_i \xrightarrow{a.s.} \boldsymbol{s} = \Omega^N_{\boldsymbol{e}}(\boldsymbol{s})$$
$$R^N_i = \boldsymbol{s}+\boldsymbol{n}_i - \Omega_{\boldsymbol{e}}(\boldsymbol{s},\boldsymbol{n}) \xrightarrow{a.s.} \boldsymbol{n}_i = R_{\boldsymbol{e},i}(\boldsymbol{n}_i)$$

where the convergences follow from application of Kolmogorov's strong law, using the fact that $\mathrm{Var}(\boldsymbol{n}_i) \leq C$ for all i. Satisfaction of condition 6.17 follows from the fact that $\mathrm{Var}_{\boldsymbol{s},\boldsymbol{n}_i}(\boldsymbol{s}+\boldsymbol{n}_i) \leq C + \mathrm{Var}(\boldsymbol{s}) \leq K$. Since $\boldsymbol{s}$ is a well-defined random variable, $\Omega_{\boldsymbol{e}}(\boldsymbol{s}) < \infty$ with probability 1, satisfying condition 6.18. It follows from the mutual independence of $\boldsymbol{n}_i$ and $\boldsymbol{n}_j$ that $R_{\boldsymbol{e},i}$ and $R_{\boldsymbol{e},j}$ satisfy condition 6.19. Condition 6.20 follows from the fact that $\mathbb{E}[\boldsymbol{n}_i] = 0$ Condition 6.21 follows from $R_{\boldsymbol{e},i}$ being constant as a function of $\boldsymbol{s}$.

It therefore follows that $\{\boldsymbol{k}_i^{-1}\} \in \mathcal{G}_K$ for K sufficiently large.

We will next show that if $\{\boldsymbol{e}_i\} \in \mathcal{G}_K$ then there exist a matrix $\boldsymbol{\alpha}$ and vector $\boldsymbol{\beta}$ such that $\boldsymbol{e}_i = \boldsymbol{\alpha}\boldsymbol{k}_i^{-1} + \boldsymbol{\beta}$ for all i. Since $\boldsymbol{e}_i$ acts coordinate-wise, it moreover follows that $\boldsymbol{\alpha}$ is diagonal.

First, we will show that each $\boldsymbol{e}_i \circ \boldsymbol{k}_i$ is affine, i.e. there exist potentially different $\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i$ such that $\boldsymbol{e}_i = \boldsymbol{\alpha}_i\boldsymbol{k}_i^{-1} + \boldsymbol{\beta}_i$ for each i.

Then we will show that we must have $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ for all i, j.

To see that $\boldsymbol{e}_i$ is affine, we make use of that fact that $R_{\boldsymbol{e},i}$ is constant as a function of $\boldsymbol{s}$. It follows that for any x and y

$$\begin{aligned}\boldsymbol{e}_i \circ \boldsymbol{k}_i(x+y) &= R_{\boldsymbol{e},i}(x) + \Omega_{\boldsymbol{e}}(y)\\ &= R_{\boldsymbol{e},i}(x) + \Omega_{\boldsymbol{e}}(0) + R_{\boldsymbol{e},i}(0) + \Omega_{\boldsymbol{e}}(y)\\ &\quad - (R_{\boldsymbol{e},i}(0) + \Omega_{\boldsymbol{e}}(0))\\ &= \boldsymbol{e}_i \circ \boldsymbol{k}_i(x) + \boldsymbol{e}_i \circ \boldsymbol{k}_i(y) - \boldsymbol{e}_i \circ \boldsymbol{k}_i(0)\end{aligned}$$

It therefore follows that $\boldsymbol{e}_i \circ \boldsymbol{k}_i$ is affine, since if we define

$$\begin{aligned}L(x+y) &= \boldsymbol{e}_i \circ \boldsymbol{k}_i(x+y) - \boldsymbol{e}_i \circ \boldsymbol{k}_i(0)\\ &= (\boldsymbol{e}_i \circ \boldsymbol{k}_i(x) - \boldsymbol{e}_i \circ \boldsymbol{k}_i(0))\\ &\quad + (\boldsymbol{e}_i \circ \boldsymbol{k}_i(y) - \boldsymbol{e}_i \circ \boldsymbol{k}_i(0))\\ &= L(x) + L(y)\end{aligned}$$

then L is linear and we can write $\boldsymbol{e}_i \circ \boldsymbol{k}_i(x)$ as the sum of a linear function and a constant:

$$\boldsymbol{e}_i \circ \boldsymbol{k}_i(x) = L(x) + \boldsymbol{e}_i \circ \boldsymbol{k}_i(0)$$

Thus $\boldsymbol{e}_i \circ \boldsymbol{k}_i$ is affine, and we have some (diagonal) matrix $\boldsymbol{\alpha}_i$ and vector $\boldsymbol{\beta}_i$ such that for any x

$$\begin{aligned} \boldsymbol{e}_i \circ \boldsymbol{k}_i(x) &= \boldsymbol{\alpha}_i x + \boldsymbol{\beta}_i \\ \implies \boldsymbol{e}_i\,(x) &= \boldsymbol{\alpha}_i \boldsymbol{k}_i^{-1} x + \boldsymbol{\beta}_i. \end{aligned}$$

Next we show that for the set of $\{\boldsymbol{e}_i = \boldsymbol{\alpha}_i \boldsymbol{k}_i^{-1} + \boldsymbol{\beta}_i\}$, it must be the case that each $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$.

Observe that

$$\begin{aligned} \Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n}) &= \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\alpha}_i \boldsymbol{s} + \boldsymbol{\alpha}_i \boldsymbol{n}_i + \boldsymbol{\beta}_i \\ &= \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\alpha}_i \right) \boldsymbol{s} + \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\beta}_i + \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\alpha}_i \boldsymbol{n}_i \\ \mathbb{E}_{\boldsymbol{n}}[\Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n})] &= \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\alpha}_i \right) \boldsymbol{s} + \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\beta}_i \end{aligned}$$

Define

$$\begin{aligned} \boldsymbol{\alpha} &= \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\alpha}_i \\ \boldsymbol{\beta} &= \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\beta}_i \end{aligned}$$

which exist by the assumption that $\Omega_{\boldsymbol{e}}^N(\boldsymbol{s}, \boldsymbol{n})$ converges as $N \to \infty$. Thus

$$\begin{aligned} \Omega_{\boldsymbol{e}}(\boldsymbol{s}) &= \boldsymbol{\alpha} \boldsymbol{s} + \boldsymbol{\beta} \\ R_{\boldsymbol{e},i}(\boldsymbol{s}, \boldsymbol{n}_i) &= (\boldsymbol{\alpha}_i - \boldsymbol{\alpha}) \boldsymbol{s} + \boldsymbol{\alpha}_i \boldsymbol{n}_i + \boldsymbol{\beta}_i - \boldsymbol{\beta} \end{aligned}$$

Now, suppose that there exist i and j such that such that $\boldsymbol{\alpha}_i \neq \boldsymbol{\alpha}_j$. It follows that

$$\begin{aligned} R_{\boldsymbol{e},i}(\boldsymbol{s}, \boldsymbol{n}_i) &= (\boldsymbol{\alpha}_i - \boldsymbol{\alpha}) \boldsymbol{s} + \boldsymbol{\alpha}_i \boldsymbol{n}_i + \boldsymbol{\beta}_i - \boldsymbol{\beta} \\ R_{\boldsymbol{e},j}(\boldsymbol{s}, \boldsymbol{n}_j) &= (\boldsymbol{\alpha}_j - \boldsymbol{\alpha}) \boldsymbol{s} + \boldsymbol{\alpha}_j \boldsymbol{n}_j + \boldsymbol{\beta}_j - \boldsymbol{\beta} \end{aligned}$$

There are two cases. If $\boldsymbol{\alpha}_i \neq \boldsymbol{\alpha}$, then $R_{\boldsymbol{e},i}(\boldsymbol{s}, \boldsymbol{n}_i)$ is not a constant function of $\boldsymbol{s}$. But if $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}$, then $\boldsymbol{\alpha}_j \neq \boldsymbol{\alpha}$ and so $R_{\boldsymbol{e},j}(\boldsymbol{s}, \boldsymbol{n}_j)$ is not a constant function of $\boldsymbol{s}$. This is a contradiction, and so $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j$ for all i, j.

Suppose similarly that there exist $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}_j$. If $\boldsymbol{\beta}_i \neq \boldsymbol{\beta}$, then $\mathbb{E}[R_{\boldsymbol{e},i}(\boldsymbol{n}_i)] = \boldsymbol{\beta}_i - \boldsymbol{\beta}$ which is non-zero. If $\boldsymbol{\beta}_i = \boldsymbol{\beta}$, then $\boldsymbol{\beta}_j \neq \boldsymbol{\beta}$ and so $\mathbb{E}[R_{\boldsymbol{e},j}(\boldsymbol{n}_j)] = \boldsymbol{\beta}_j - \boldsymbol{\beta}$ is non-zero. This is a contradiction, and so $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ for all i, j.

We have thus proven that set $\{\boldsymbol{e}_i\} \in \mathcal{G}_K$ is of the form $\boldsymbol{e}_i = \boldsymbol{\alpha} \boldsymbol{k}_i^{-1} + \boldsymbol{\beta}$ for all i.

E.8 Other Related Work on Multi-view Latent Variable Models

Bearing a strong resemblance to our considered setting, [518] proposes a sequence of diffusion maps to find the common source of variability captured by multiple sensors, discarding irrelevant sensor-specific effects. It computes the distance among the samples measured by different sensors to form a similarity matrix for the measurements of each sensor; each similarity matrix is then associated to a diffusion operator, which is a Markov matrix by construction. A Markov chain is then run by alternately applying these Markov matrices on the initial state. During these Markovian dynamics, sensor specific information will eventually vanish, and the final state will only contain information on the common source. While the method focuses on recovering the common information in the form of a parametrization of the common variable, our method both inverts the mixing mechanisms of each view and recovers the common latent variables.

[519] proves identifiability for multi-view, latent variable models, unifying previously proposed spectral techniques [520]. However, while the setting is similar to the one considered in this work, both the objectives and the employed methods are different. The paper considers the setting in which L variables X_l, $l = 1, \ldots, L$ are observed; additionally, there exists an unobserved latent variable H, such that conditional distributions $P(X_l|H)$ are independent. While the setting bears obvious similarities with our multi-view ICA, the method proposed in [519] is aimed at learning the mixture parameters, rather than the exact realization of latent variables. Their method is based on the mean embedding of distributions in a Reproducing Kernel Hilbert Space and a result of identifiability for the parameters of the mean embeddings of $P(H)$ and $P(X|H)$ is proved.

Another related field of study is multi-view clustering, which considers a multiview setting and aims at performing clustering on a given dataset, see e.g. [521] and [522]. While related to our setting, this line of work is different from it in two key ways. Firstly, clustering can be thought of as assigning a discrete latent label per datapoint. In contrast, our setting seeks to recover a continuous latent vector per datapoint. Second, since no underlying generative model with discrete latent variable is assumed, identifiability results are not given.

F

Additional Material on Chapter 7

Overview

F.1 Likelihood

F.1.1 Initial form of likelihood

To derive the likelihood, we start by conditioning on $\mathbf{s}$. Then, we make a variable transformation from $\mathbf{x}^i$ to $\mathbf{n}^i = \mathbf{W}^i\mathbf{x}^i - \mathbf{s}$, as opposed to the transformation to $\mathbf{s}$ as is usual in ICA. Using the probability transformation formula, we obtain

$$p(\mathbf{x}^i|\mathbf{s}) = |\mathbf{W}^i| p_n^i(\mathbf{W}^i\mathbf{x}^i - \mathbf{s}) \tag{F.1}$$

where p_n^i is the distribution of $\mathbf{n}^i$. Note that the $\mathbf{x}^i$ are conditionally independent given $\mathbf{s}$, so we have their joint probability as

$$p(\mathbf{x}|\mathbf{s}) = \prod_{i=1}^{m} |\mathbf{W}^i| p_n^i(\mathbf{W}^i\mathbf{x}^i - \mathbf{s}) \tag{F.2}$$

and we next get the joint probability as

$$p(\mathbf{x}, \mathbf{s}) = p(\mathbf{s}) \prod_{i=1}^{m} |\mathbf{W}^i| p_n^i(\mathbf{W}^i\mathbf{x}^i - \mathbf{s}) \tag{F.3}$$

Integrating out $\mathbf{s}$ gives Eq. (7.2).

F.1.2 Integrating out the sources

The integral in question, after factorization, is given by

$$\int_{\mathbf{s}} \prod_{j=1}^{k} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{m} ((\mathbf{w}_j^i)^\top \mathbf{x}^i - s_j)^2\right) d(s_j) d\mathbf{s} \tag{F.4}$$

which factorizes for each j. Denote $y_j^i = (\mathbf{w}_j^i)^\top \mathbf{x}^i$ and $\tilde{s_j} = \frac{1}{m}\sum_{i=1}^m y_j^i$. Fix j, and drop it to simplify notation. Then we need to solve the integral

$$\begin{aligned}
&\int_s \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^m (y^i - s)^2\right) d(s)ds \\
&= \int_s \exp\left(-\frac{1}{2\sigma^2}[m(\tilde{s} - s)^2 + \sum_{i=1}^m (y^i - \tilde{s})^2]\right) d(s)ds \\
&= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^m (y^i - \tilde{s})^2\right) \int_z \exp\left(-\frac{m}{2\sigma^2}z^2\right) d(\tilde{s} - z)dz
\end{aligned}$$

where we have made the change of variable $z = \tilde{s} - s$. The remaining integral simply means that d is smoothed by a Gaussian kernel, which can be computed exactly if d is a Gaussian mixture. We therefore define $f(s) = \log\left(\int_z \exp\left(-\frac{m}{2\sigma^2}z^2\right) d(s - z)dz\right)$.

F.2 Initialization of MultiViewICA

Since the cost function $\mathcal{L}$ is non-convex, having a good initialization can make a difference in the final result. We propose a two stage approach. We begin by applying PermICA on the datasets, which gives us a first set of unimixing matrices $\mathbf{W}_1^1, \ldots, \mathbf{W}_1^m$. Note that we could also use GroupICA for this task. Next, we perform a diagonal scaling of the mixing matrices, i.e. we find the diagonal matrices $\mathbf{D}^1, \ldots, \mathbf{D}^m$ such that $\mathcal{L}(\mathbf{D}^1\mathbf{W}_1^1, \ldots, \mathbf{D}^m\mathbf{W}_1^m)$ is minimized. To do so, we employ Algorithm 1 but only take into account the diagonal of the descent direction at each step: the update rule becomes $\mathbf{W}^i \leftarrow (\mathbf{I}_k + \rho\,\mathrm{diag}(\mathbf{S}))\mathbf{W}^i$. The initial unmixing matrices for Algorithm 1 are then taken as $\mathbf{D}^1\mathbf{W}_1^1, \ldots, \mathbf{D}^m\mathbf{W}_1^m$.

Empirically, we find that this two stage procedure allows for the algorithm to start close from a satisfactory solution.

F.3 Proofs of Section 7.2

F.3.1 Proof of Prop. 7.2.1

We fix a subject i. Since $\mathbf{s}$ has independent components, so does $\mathbf{s} + \mathbf{n}^i$. Following [37], Theorem 11, there exists a scale-permutation matrix $\mathbf{P}^i$ such that $\mathbf{A}'^i = \mathbf{A}^i\mathbf{P}^i$. As a consequence, we have $\mathbf{s} + \mathbf{n}^i = \mathbf{P}^i(\mathbf{s}' + \mathbf{n}'^i)$ for all i.

Then, we focus on subject 1 and subject $i \neq 1$:

$$\mathbf{s} + \mathbf{n}^1 - (\mathbf{s} + \mathbf{n}^i) = \mathbf{P}^1(\mathbf{s}' + \mathbf{n}'^1) - \mathbf{P}^i(\mathbf{s}' + \mathbf{n}'^i) \tag{F.5}$$

$$\mathbf{n}^1 - \mathbf{n}^i = \mathbf{P}^1(\mathbf{s}' + \mathbf{n}'^1) - \mathbf{P}^i(\mathbf{s}' + \mathbf{n}'^i) \tag{F.6}$$

$$\iff \mathbf{P}^1\mathbf{s}' - \mathbf{P}^i\mathbf{s}' = \mathbf{P}^i\mathbf{n}'^i - \mathbf{n}^i + \mathbf{n}^1 - \mathbf{P}^1\mathbf{n}'^1 \tag{F.7}$$

Since the right hand side of equation (F.7) is a linear combination of Gaussian random variables, this would imply that $\mathbf{P}^1\mathbf{s}' - \mathbf{P}^i\mathbf{s}'$ is also

Gaussian. However, given that $\mathbf{s}'$ is assumed to be non-Gaussian, the equality can only hold if $\mathbf{P}^1 = \mathbf{P}^i$ and both the right and the left hand side vanish. Therefore, the matrices $\mathbf{P}^i$ are all equal, and there exists a scale and permutation matrix $\mathbf{P}$ such that $\mathbf{A}'^i = \mathbf{A}^i\mathbf{P}$.

F.3.2 Proof of Prop. 7.2.2

We consider $\mathbf{W}^i = \mathbf{D}(\mathbf{A}^i)^{-1}$, where $\mathbf{D}$ is a diagonal matrix. We recall $\mathbf{x}^i = \mathbf{A}^i(\mathbf{s} + \mathbf{n}^i)$, so that $\mathbf{y}^i = \mathbf{W}^i\mathbf{x}^i = \mathbf{D}(\mathbf{s} + \mathbf{n}^i)$. The gradient of $\mathscr{L}$ is given by eq. 7.7:

$$\mathbf{G}^i = \frac{1}{m} f'(\tilde{\mathbf{s}})(\mathbf{s} + \mathbf{n}^i)^\top \mathbf{D} + \frac{1 - 1/m}{\sigma^2} \mathbf{D}\left(\mathbf{n}_i - \frac{1}{m-1}\sum_{j\neq i} \mathbf{n}^j\right)(\mathbf{s} + \mathbf{n}^i)^\top \mathbf{D} - \mathbf{I}_k \tag{F.8}$$

$$= \frac{1}{m} f'(\mathbf{D}(\mathbf{s} + \frac{1}{m}\sum_j \mathbf{n}^j))(\mathbf{s} + \mathbf{n}^i)^\top \mathbf{D} + \frac{\sigma'^2(1 - 1/m)}{\sigma^2}\mathbf{D}^2 - \mathbf{I}_k \tag{F.9}$$

where we write $f'(\mathbf{s}) = \begin{bmatrix} f'(s_1) \\ \vdots \\ f'(s_k) \end{bmatrix}$. Therefore, $\mathbf{G}^i$ is diagonal and constant across subjects (because $f'(\mathbf{D}(\mathbf{s}+\frac{1}{m}\sum_j \mathbf{n}^j))(\mathbf{n}^i)^\top = f'(\mathbf{D}(\mathbf{s}+\frac{1}{m}\sum_j \mathbf{n}^j))(\mathbf{n}^{i'})^\top$). Let us therefore consider only its coefficient (a, a), and let $\mathbf{D} = \mathbf{D}_{aa}$:

$$G^i_{aa} = G(\lambda) = \phi(\lambda)\lambda + \frac{\sigma'^2(1 - 1/m)}{\sigma^2}\lambda^2 - 1,$$

where $\phi(\lambda) = \frac{1}{m} f'(\lambda(s_a + \frac{1}{m}\sum_j n^j_a))(s_a + n^i_a)$. One the one hand, we have $G(0) = -1$. On the other hand, if we assume for instance that f' has sub linear growth (i.e. $|f'(x)| \leq c|x|^\alpha + d$ for some $\alpha < 1$) or that ϕ is positive, we find that $G(+\infty) = +\infty$. Therefore, G cancels, which concludes the proof.

F.3.3 Stability conditions

We consider $\mathbf{W}^i = \mathbf{D}(\mathbf{A}^i)^{-1}$ where $\mathbf{D}$ is such that the gradients $\mathbf{G}^i$ all cancel. We consider a small relative perturbation of $\mathbf{W}^i$ of the form $\mathbf{W}^i \leftarrow (\mathbf{I}_k + \boldsymbol{\epsilon}^i)\mathbf{W}^i$, and consider the effect on the gradient. We define $\Delta^i = \mathbf{G}^i\left((\mathbf{I}_k + \boldsymbol{\epsilon}^1)\mathbf{W}^1, \ldots, (\mathbf{I}_k + \boldsymbol{\epsilon}^m)\mathbf{W}^m\right)$. Denoting $C = \frac{1-1/m}{\sigma^2}$ and $\tilde{\mathbf{n}} = \frac{1}{m}\sum_{i=1}^m \mathbf{n}^i$, we find:

$$\Delta^i = \underbrace{\frac{1}{m} f'\left(\mathbf{D}(\mathbf{s} + \tilde{\mathbf{n}}) + \frac{1}{m}\sum_{j=1}^m \boldsymbol{\epsilon}^j \mathbf{D}(\mathbf{s} + \mathbf{n}^j)\right)(\mathbf{s} + \mathbf{n}^i)^\top \mathbf{D}(\mathbf{I}_k + \boldsymbol{\epsilon}^i)^\top}_{\Delta^i_1} +$$

$$\underbrace{C\left(\mathbf{D}\mathbf{n}^i - \frac{1}{m-1}\sum_{j\neq i}\mathbf{D}\mathbf{n}^j + \boldsymbol{\epsilon}^i\mathbf{D}(\mathbf{s} + \mathbf{n}^i) - \frac{1}{m-1}\sum_{j\neq i}\boldsymbol{\epsilon}^j\mathbf{D}(\mathbf{s} + \mathbf{n}^j)\right)(\mathbf{s} + \mathbf{n}^i)^\top \mathbf{D}(\mathbf{I}_k + \boldsymbol{\epsilon}^i)^\top}_{\Delta^i_2}$$

$$- \mathbf{I}_k \tag{F.10}$$

The first term is expanded at the first order, denoting $S = \sum_{j=1}^m \boldsymbol{\epsilon}^j$:

$$\begin{aligned}\Delta_1^i &= \frac{1}{m}\left(f'(\mathbf{D}(\mathbf{s}+\tilde{\mathbf{n}})) + f''(\mathbf{D}(\mathbf{s}+\tilde{\mathbf{n}})) \odot \left(\frac{1}{m}\sum_{j=1}^m \boldsymbol{\epsilon}^j \mathbf{D}(\mathbf{s}+\mathbf{n}^j)\right)\right)(\mathbf{s}+\mathbf{n}^i)^\top \mathbf{D}(\mathbf{I}_k + \boldsymbol{\epsilon}^i)^\top \\ &= \frac{1}{m} f'(\mathbf{D}(\mathbf{s}+\tilde{\mathbf{n}}))(\mathbf{s}+\mathbf{n}^i)^\top \mathbf{D}(\mathbf{I}_k + \boldsymbol{\epsilon}^i)^\top + \frac{1}{m^2} S \odot \left(f''(\mathbf{D}(\mathbf{s}+\tilde{\mathbf{n}}))(\mathbf{s}^2)^\top \mathbf{D}^2\right) \\ &+ \frac{1}{m^2}\boldsymbol{\epsilon}^i \odot \left(f''(\mathbf{D}(\mathbf{s}+\tilde{\mathbf{n}}))((\mathbf{n}^i)^2)^\top \mathbf{D}^2\right) \qquad \text{(F.11)}\end{aligned}$$

The symbol $\odot$ denotes the element-wise multiplication, $f'(\mathbf{s}) = \begin{bmatrix} f'(s_1) \\ \vdots \\ f'(s_k) \end{bmatrix}$ and $f''(\mathbf{s}) = \begin{bmatrix} f''(s_1) \\ \vdots \\ f''(s_k) \end{bmatrix}$. Similarly, the second term gives at the first order:

$$\Delta_2^i = \sigma'^2 \mathbf{D}^2(\mathbf{I}_k + \boldsymbol{\epsilon}^i)^\top + (1+\sigma'^2)\boldsymbol{\epsilon}^i \mathbf{D}^2 - \frac{1}{m-1}(S - \boldsymbol{\epsilon}^i)\mathbf{D}^2 \qquad \text{(F.12)}$$

Combining this, we find:

$$\Delta^i = (\boldsymbol{\epsilon}^i)^\top + \boldsymbol{\epsilon}^i \odot \Gamma^E + S \odot \Gamma^S \qquad \text{(F.13)}$$

where

$$\Gamma^E = \left(\frac{1}{m^2} f''(\mathbf{D}(\mathbf{s}+\tilde{\mathbf{n}}))((\mathbf{n}^i)^2)^\top + (1-\frac{1}{m})\frac{\sigma'^2}{\sigma^2} + \frac{1}{\sigma^2}\right)\mathbf{D}^2$$

$$\Gamma^S = \left(\frac{1}{m^2} f''(\mathbf{D}(\mathbf{s}+\tilde{\mathbf{n}}))(\mathbf{s}^2)^\top - \frac{1}{m\sigma^2}\right)\mathbf{D}^2$$

are $k \times k$ matrices, independent of the subject. This linear operator is the Hessian block corresponding to the i-th subject: Denoting $\mathscr{H}$ the Hessian, it is the mapping $\mathscr{H}(\boldsymbol{\epsilon}^1, \dots, \boldsymbol{\epsilon}^m) = (\Delta^1, \dots, \Delta^m)$.

The coefficient Δ^i_{ab} only depends on $(\boldsymbol{\epsilon}^i_{ab}, \boldsymbol{\epsilon}^i_{ba}, \boldsymbol{\epsilon}^1_{ab}, \dots, \boldsymbol{\epsilon}^m_{ab})$. Therefore, the Hessian is block diagonal with respect to the blocks of coordinates $(\boldsymbol{\epsilon}^1_{ab}, \boldsymbol{\epsilon}^1_{ba}, \dots, \boldsymbol{\epsilon}^m_{ab}, \boldsymbol{\epsilon}^m_{ba})$. Denote $\varepsilon = \Gamma^E_{ab}$, $\varepsilon' = \Gamma^E_{ba}$, $\beta = \Gamma^S_{ab}$ and $\beta' = \Gamma^S_{ba}$. The linear operator for the block is:

$$K(\varepsilon, \varepsilon', \beta, \beta') = \left(\begin{array}{cc|cc|c|cc} \varepsilon+\beta & 1 & \beta & 0 & \dots & \beta & 0 \\ 1 & \varepsilon'+\beta' & 0 & \beta' & \dots & 0 & \beta' \\ \hline \beta & 0 & \varepsilon+\beta & 1 & & \beta & 0 \\ 0 & \beta' & 1 & \varepsilon'+\beta' & \ddots & 0 & \beta' \\ \hline \vdots & \vdots & & \ddots & \ddots & \vdots & \vdots \\ \hline \beta & 0 & \beta & 0 & \dots & \varepsilon+\beta & 1 \\ 0 & \beta' & 0 & \beta' & \dots & 1 & \varepsilon'+\beta' \end{array}\right)$$

The positivity of $\mathscr{H}$ is equivalent to the positivity of this operator for all pairs a, b. We now assume $\beta\beta' > 0$.

First, we should note that $K(\varepsilon, \varepsilon', \beta, \beta')$ is congruent to $K(\varepsilon\sqrt{\frac{\beta'}{\beta}}, \varepsilon'\sqrt{\frac{\beta}{\beta'}}, \sqrt{\beta\beta'}, \sqrt{\beta\beta'})$ via the basis $\mathrm{diag}((\frac{\beta'}{\beta})^{1/4}, (\frac{\beta}{\beta'})^{1/4}, \cdots, (\frac{\beta'}{\beta})^{1/4}, (\frac{\beta}{\beta'})^{1/4})$. We denote to simplify notation $\alpha = \varepsilon\sqrt{\frac{\beta'}{\beta}}$, $\alpha' = \varepsilon'\sqrt{\frac{\beta}{\beta'}}$ and $\gamma = \sqrt{\beta\beta'}$. We only have to study the positivity of $K(\alpha, \alpha', \gamma, \gamma)$. We have:

$$K(\alpha, \alpha', \gamma, \gamma) = \mathbf{I}_m \otimes M_\alpha + \gamma \mathbb{1} \otimes \mathbf{I}_2, \quad M_\alpha = \begin{pmatrix} \alpha & 1 \\ 1 & \alpha' \end{pmatrix}$$

Since $\mathbf{I}_m \otimes M_\alpha$ and $\gamma\mathbb{1} \otimes \mathbf{I}_2$ commute, the minimum value of $\mathrm{Sp}(K)$ is $\min(\mathbf{I}_m \otimes M_\alpha) + \min(\gamma \mathrm{Sp}(\mathbb{1})) = \frac{1}{2}(\alpha + \alpha' - \sqrt{(\alpha - \alpha')^2 + 4}) + m \min(0, \gamma)$. Since we assumed $\beta\beta' > 0$ we have $\gamma > 0$. This is similar to the usual ICA case, we find that the condition is $\alpha\alpha' > 1$.

If the following conditions hold for all pair of sources a, b, the sources are a local minimum of the cost function:

- $\Gamma^S_{ab}\Gamma^S_{ba} \geq 0$
- $\Gamma^E_{ab}\Gamma^E_{ba} > 1$

F.4 Identifiability for Shared Response Model

The shared response model [264] (SRM) models the data $\mathbf{x}^i \in \mathbb{R}^v$ of subject i for $i = 1, \ldots, m$ as

$$\mathbf{x}^i = \mathbf{A}^i\mathbf{s} + \mathbf{n}^i \text{ with } \mathbf{s} \sim \mathcal{N}(0, \mathbf{\Sigma}), \quad \mathbf{n}^i \sim \mathcal{N}(0, \rho_i^2\mathbf{I}_v), \quad {\mathbf{A}^i}^\top \mathbf{A}^i = \mathbf{I}_k$$

where $\mathbf{A}^i \in \mathbb{R}^{v\times k}$, $\mathbf{s} \in \mathbb{R}^k$ and $\mathbf{\Sigma} \in \mathbb{R}^{k\times k}$ is a symmetric positive definite matrix.

Proposition F.4.1 *SRM is not identifiable*

Proof. Let us assume the data $\mathbf{x}^i$ $i = 1, \ldots, m$ follow the SRM model with parameters $\mathbf{\Sigma}, \mathbf{A}^i, \rho_i^2$ $i = 1, \ldots, m$.

Let us consider an orthogonal matrix $\mathbf{O} \in \mathbb{O}_k$. We call $\mathbf{A'}^i = \mathbf{A}^i\mathbf{O}$ and $\mathbf{\Sigma}' = \mathbf{O}^\top\mathbf{\Sigma}\mathbf{O}$. $\mathbf{\Sigma}'$ is trivially symmetric positive definite.

Then the data also follows the SRM model with different parameters $\mathbf{\Sigma}', \mathbf{A'}^i, \rho_i^2$ $i = 1, \ldots, m$. □

Proposition F.4.2 *We consider the decorrelated SRM model with an additional decorrelation assumption on the shared responses.*

$$\mathbf{x}^i = \mathbf{A}^i\mathbf{s} + \mathbf{n}^i \text{ with } \mathbf{s} \sim \mathcal{N}(0, \mathbf{\Sigma}), \quad \mathbf{n}^i \sim \mathcal{N}(0, \rho_i^2\mathbf{I}_v), \quad {\mathbf{A}^i}^\top \mathbf{A}^i = \mathbf{I}_k$$

where $\mathbf{\Sigma}$ is a positive diagonal *matrix. We further assume that the values in $\mathbf{\Sigma}$ are all distinct and ranked in ascending order. The decorrelated SRM is identifiable up to sign indeterminacies on the columns of* $\begin{bmatrix} \mathbf{A}^1 \\ \vdots \\ \mathbf{A}^m \end{bmatrix}$.

Proof. The decorrelated SRM model can be written

$$\mathbf{x}^i \sim \mathcal{N}(0, \mathbf{A}^i \mathbf{\Sigma} {\mathbf{A}^i}^\top + \rho_i^2 \mathbf{I}_v) \text{ with } {\mathbf{A}^i}^\top \mathbf{A}^i = \mathbf{I}_k$$

where $\mathbf{\Sigma}$ is a positive diagonal matrix with distincts values ranked in ascending order.

Let us assume the data $\mathbf{x}^i \;\; i = 1, \ldots, m$ follow the decorrelated SRM model with parameters $\mathbf{\Sigma}, \mathbf{A}^i, {\rho_i}^2 \;\; i = 1, \ldots, m$. Let us further assume that the data $\mathbf{x}^i \;\; i = 1, \ldots, m$ follow the decorrelated SRM model with an other set of parameters $\mathbf{\Sigma}', \mathbf{A}'^i, \rho_i'^2 \;\; i = 1, \ldots, m$.

Since the model is Gaussian, we look at the covariances. We have for $i \neq j$

$$\mathbb{E}[\mathbf{x}^i \left(\mathbf{x}^j\right)^\top] = \mathbf{A}^i \mathbf{\Sigma} {\mathbf{A}^j}^\top = \mathbf{A}'^i \mathbf{\Sigma}' {\mathbf{A}'^j}^\top \;\; ,$$

The singular value decomposition is unique up to sign flips and permutation. Since eigenvalues are positive and ranked the only indeterminacies left are on the eigenvectors. For each eigenvalue a sign flip can occur simultaneously on the corresponding left and right eigenvector.

Therefore we have $\mathbf{\Sigma}' = \mathbf{\Sigma}, \mathbf{A}^i = \mathbf{A}'^i \mathbf{D}^{ij}$ and $\mathbf{A}^j = \mathbf{A}'^j \mathbf{D}^{ij}$ where $\mathbf{D}^{ij} \in \mathbb{R}^{k \times k}$ is a diagonal matrix with values in $\{-1, 1\}$. This analysis holds for every $j \neq i$ and therefore $\mathbf{D}^{ij} = \mathbf{D}$ is the same for all subjects.

We also have for all i

$$\mathbb{E}[\mathbf{x}^i \left(\mathbf{x}^i\right)^\top] = \mathbf{A}^i \mathbf{\Sigma} {\mathbf{A}^i}^\top + \rho_i^2 \mathbf{I}_v = \mathbf{A}'^i \mathbf{\Sigma}' {\mathbf{A}'^i}^\top + \rho_i'^2 \mathbf{I}_v$$

We therefore conclude $\rho_i'^2 = \rho_i^2, i = 1 \ldots m$.

Note that if the diagonal subject specific noise covariance $\rho_i^2 \mathbf{I}_v$ is replaced by any positive definite matrix, the model still enjoys identifiability. □

F.5 fMRI experiments

F.5.1 Dataset description and preprocessing

The full brain mask used to select brain regions is available in the Python package associated with the paper.

Sherlock In *sherlock* dataset, 17 participants are watching "Sherlock" BBC TV show (beginning of episode 1). These data are downloaded from `http://arks.princeton.edu/ark:/88435/dsp01nz8062179`. Data were acquired using a 3T scanner with an isotropic spatial resolution of 3 mm. More information including the preprocessing pipeline is available in [317]. Subject 5 is removed because of missing data leaving us with 16 participants. Although *sherlock* data are downloaded as a temporal concatenation of two runs, we split it manually into 4 runs of 395 timeframes and one run of 396 timeframes so that we can perform 5 fold cross-validation in our experiments.

FORREST In FORREST dataset 20 participants are listening to an audio version of the Forrest Gump movie. FORREST data are downloaded from OpenfMRI [523]. Data were acquired using a 7T scanner with an isotropic spatial resolution of 1 mm (see more details in [318]) and resampled to an isotropic spatial resolution of 3 mm. More information about the forrest project can be found at http://studyforrest.org. Subject 10 is discarded because not all runs available for other subjects were available for subject 10 at the time of writing. Run 8 is discarded because it is not present in most subjects.

RAIDERS In RAIDERS dataset, 11 participants are watching the movie "Raiders of the lost ark". The RAIDERS dataset belongs to the Individual Brain Charting dataset [319]. Data were acquired using a 3T scanner and resampled to an isotropic spatial resolution of 3 mm. The RAIDERS dataset reproduces the protocol described in [265]. Preprocessing details are described in [319].

CLIPS In CLIPS dataset, 12 participants are exposed to short video clips. The CLIPS dataset also belongs to the Individual Brain Charting dataset ([319]). Data were acquired using a 3T scanner and resampled to an isotropic spatial resolution of 3 mm. It reproduces the protocol of original studies described in [524] and [525]. Preprocessing details are described in [319].

At the time of writing of the original paper [116], the CLIPS and RAIDERS dataset from the individual brain charting dataset https://project.inria.fr/IBC/ are available at https://openneuro.org/datasets/ds002685. Protocols on the visual stimuli presented are available in a dedicated repository on Github: https://github.com/hbp-brain-charting/public_protocols.

F.5.2 Reconstructing the BOLD signal of missing subjects: Discussion on ROIs choice

The quality of the reconstructed BOLD signal varies depending on the choice of the region of interest. In Figure F.1, we plot for GroupICA, SRM and MultiViewICA, the R2 score per voxel using 50 components for datasets *sherlock*, *forrest*, *raiders* and *clips*. As could be anticipated from the task definition, *forrest* obtains high reconstruction accuracy in the auditory cortices, while *clips* shows good reconstruction in the visual cortex (occipital lobe mostly); the richer *sherlock* and *raiders* datasets yield good reconstructions in both domains, but also in other systems (language, motor). We also see visually see that data reconstructed by MultiViewICA are a better approximation of the original data than other methods. This is particularly obvious for the *clips* datasets where it is clear that voxels in the posterior part of the superior temporal sulcus are better recovered by MultiViewICA than by SRM or GroupICA.

In order to determine the ROIs, we focus on the R2 score per voxel between the BOLD signal reconstructed by GroupICA and the actual bold signal. We run GroupICA with 10, 20 and 50 components and select the voxels that obtained a positive R2 score for all sets of components. We

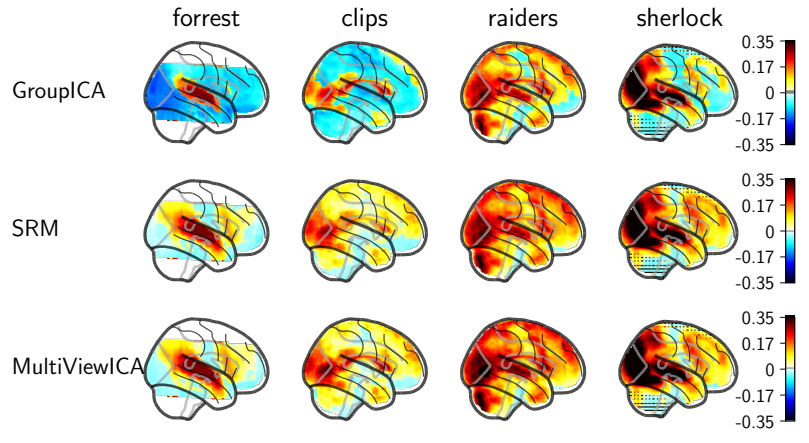


Figure F.1: Reconstructing the BOLD signal of missing subjects: Reconstruction R2 score per voxel We plot for GroupICA, SRM and MultiViewICA, the R2 score per voxel using 50 components for datasets *sherlock*, *forrest*, *raiders* and *clips*. We visually see that data reconstructed by MultiViewICA are more faithful reproduction of the original data than other methods.

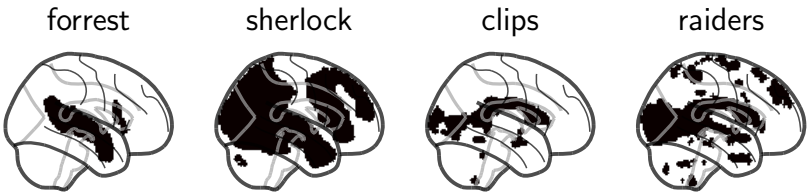


Figure F.2: Data-driven choice of ROI Chosen ROIs for the experiment: Reconstructing the BOLD signal of missing subjects.

discard voxels with an R2 score above 80% as they visually correspond to artefacts and apply a binary opening using a unit cube as the structuring element. The chosen regions are plotted in figure F.2.

F.5.3 Between-runs time-segment matching

We measure the ability of each algorithm to extract meaningful shared sources that correlate more when they correspond to the same stimulus than when they correspond to distinct stimuli. We use the *raiders-full* dataset, which allows this kind of analysis because subjects watch some selected scenes from the movie twice, during the first two runs (1 and 2) and the last two (11 and 12). First, the forward operators are learnt by fitting each algorithm with 20 components on the data of all 11 subjects using all 12 runs. We then select a subset of 8 subjects and the shared sources are computed by applying the forward operators and averaging. We select a large target time-segment (50 timeframes) taken at random from run 1 and 2, and we try to localize the corresponding sample time-segment from the 10 last runs using a single component of the shared sources. The time-segment is said to be correctly classified if the correlation between the target and corresponding sample time-segment is higher than with any other time-segment (partially overlapping windows are excluded). In contrast to the *between subject time-segment matching* experiment, we obtain one accuracy score per component. We repeat the experiment 10 times with different subsets of subjects randomly chosen and report the mean accuracy of the 3 best performing components in Figure F.3. Error bars correspond to a 95 % confidence interval. MultiView ICA achieves the highest accuracy.

We then focus on the 3 best performing components of MultiView ICA. For each component, we plot in Figure F.4 (left) the shared sources during two sets of runs where subjects were exposed to the same scenes of the movie. We then study the localisation of these sources. We average the

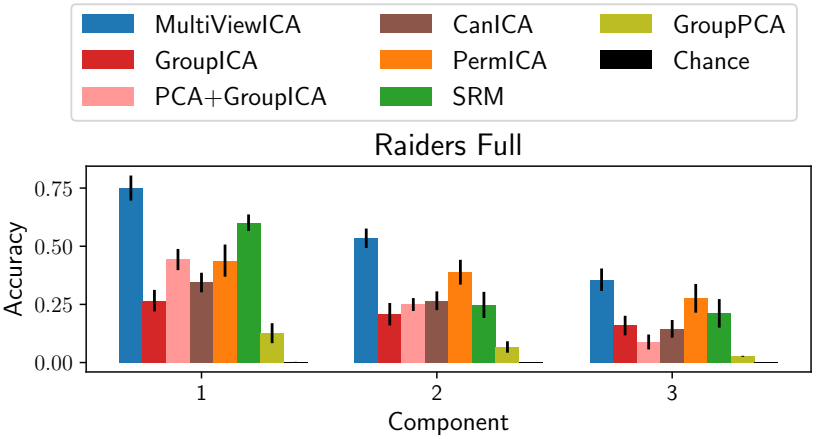


Figure F.3: Between runs time-segment matching. Interesting sources correlates more when they correspond to the same stimulus (same scenes of the movie) than when they correspond to distinct stimuli (different scenes). We extract 20 sources and report the mean accuracy of the 3 best performing sources

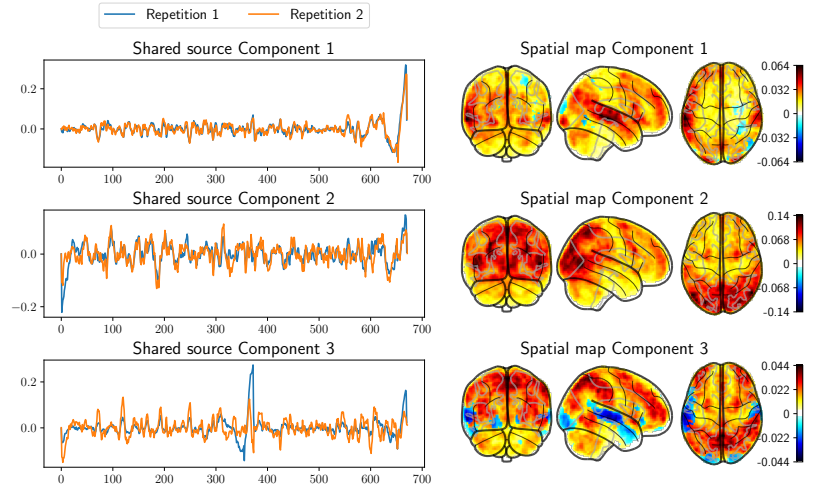


Figure F.4: Between-runs time segment matching: spatial maps and timecourses *Left:* Timecourses of the 3 shared sources yielding the highest accuracy. The two displayed set of runs correspond to the same scenes in the movie. *Right:* Localisation of the same shared sources in the brain

forward operators across subjects and plot the columns corresponding to the components of interest in Figure F.4 (right). As each column is seen as a set of weights over all voxels, it represents a spatial map.

The component 1 of the shared responses follows almost the same pattern in the two set of runs corresponding to the same scenes of the movie. The spatial map corresponding to component 1 highlights the language network. In component 2, the temporal patterns during the viewing of identical scenes are also very similar. The corresponding spatial map highlights the visual network especially the visual dorsal pathway. In component 3, there exists a similarity however less striking than with the two previous components. The corresponding spatial map highlights a contrast between the spatial attention network and the auditory network.

F.5.4 Reproducing time-segment matching experiment

We reproduce the time-segment matching experiments described in [304] and [292] and use two fold classification over runs instead of 5-fold as we have done in the main paper. We used the sherlock data available at `http://arks.princeton.edu/ark:/88435/dsp01nz8062179` and the full brain mask provided in the Python package associated with the

paper. We applied high-pass filtering (140 s cutoff) and the time series of each voxel were normalized to zero mean and unit variance.

The results are available in Figure F.5.

F.5.5 Impact of the hyperparameter σ

On top of the theoretical guarantees about the robustness of our method to the choice of the σ parameter, we investigate its practical impact on the time-matching segment experiment, on the Sherlock dataset with 10 components. We compute the accuracy of the multi-view ICA pipeline with different choice of σ. This is reported in Fig. F.6. The accuracy is constant for a wide range of σ, only decreasing when σ attains very high values.

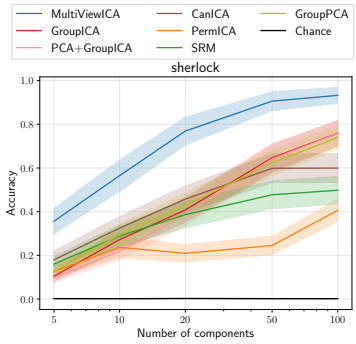


Figure F.5: Reproducing the time-segment matching experiment of [292, 304] Mean classification accuracy - error bars represent 95% confidence interval

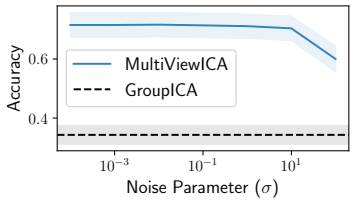


Figure F.6: Effect of the parameter σ: We compute the accuracy of the multiview-ICA pipeline on the time-segment matching experiment for various values of the σ hyperparameter over a grid. The accuracy varies only marginally with σ.

F.6 Related Work

The following table describes some usual method for extracting shared sources from multiple subjects datasets. The column "Modality/Source" describes the type of data for which each algorithm was *initially* proposed, even though each algorithm could be applied on any type of data. The source type can be either temporal if extracted sources are time courses or spatial if they are spatial patterns.

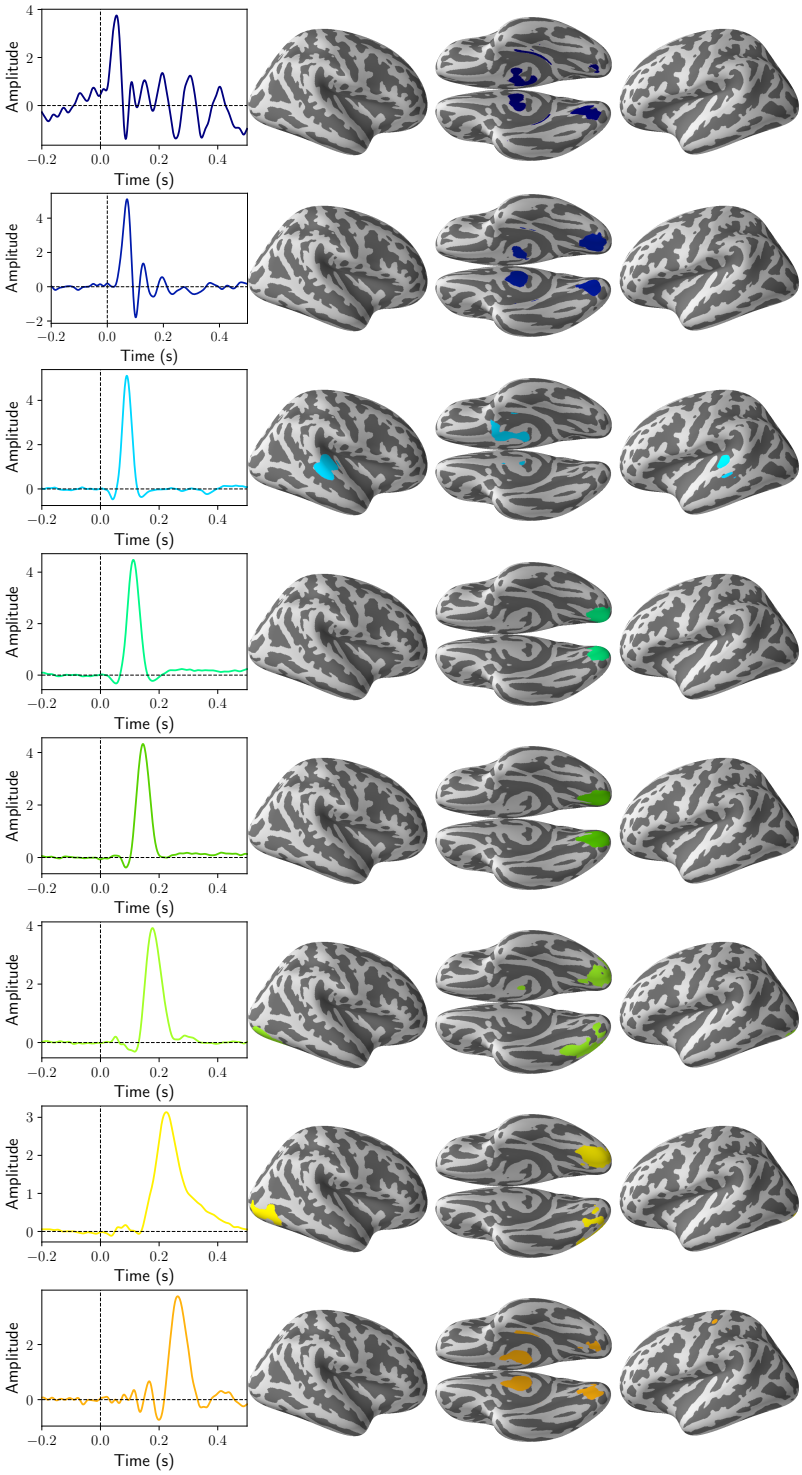
Method	**Modality/Source**	**Dimensionality reduction**	**Description**
SRM [264]	fMRI/Temporal	SRM	The model is $\mathbf{x}^i = \mathbf{A}^i\mathbf{s} + \mathbf{n}^i$, with *Gaussian* sources and *orthogonal* mixing matrices $\mathbf{A}^i$
GroupPCA [316]	fMRI/Spatial	GroupPCA	A memory efficient implementation of PCA applied on temporally concatenated data.
GIFT [288]	fMRI/Spatial	Individual PCA + Group PCA (on component-wise concatenated data)	Single-subject ICA is applied on the aggregated data
EEGIFT [290]	EEG/Temporal	Individual PCA + Group PCA (on component-wise concatenated data)	Single-subject ICA is applied on the aggregated data
PermICA	Any	Any	Single-subject ICA is applied on each subject's data, and the components are matched using the Hungarian algorithm
Clustering approach [302]	fMRI/Spatial	Individual PCA	Single-subject ICA is applied on each subject's data, and the components are matched using a hierarchical clustering algorithm.
Measure projection analysis [303]	EEG/Temporal	Individual PCA	Single-subject ICA is applied on each subject's data, and the components are matched using a hierarchical clustering algorithm.
TensorICA [296]	fMRI/Spatial	Group PCA (on spatially concatenated data)	TensorICA incorporates ICA assumptions into the PARAFAC model. The mixing matrices $A_1 \cdots A_n$ are such that $A_i = AD_i$ where A is common to all subjects and D_i are subject specific diagonal matrices.
Unifying Approach of [293]	fMRI/Spatial	Group PCA (on spatially concatenated data) + GroupPCA (on component-wise concatenated data).	The model is $\mathbf{x}^i = \mathbf{A}^i\mathbf{s} + \mathbf{n}^i$ with a Gaussian mixture model on independent sources and a matrix normal prior on the noise.
SR-ICA [292]	fMRI/Temporal	SR-ICA	SR-ICA incorporates ICA assumptions into the shared response model.
CAE-SRM [304]	fMRI/Temporal	CAE-SRM	A convolutional auto-encoder is used to perform the unmixing.

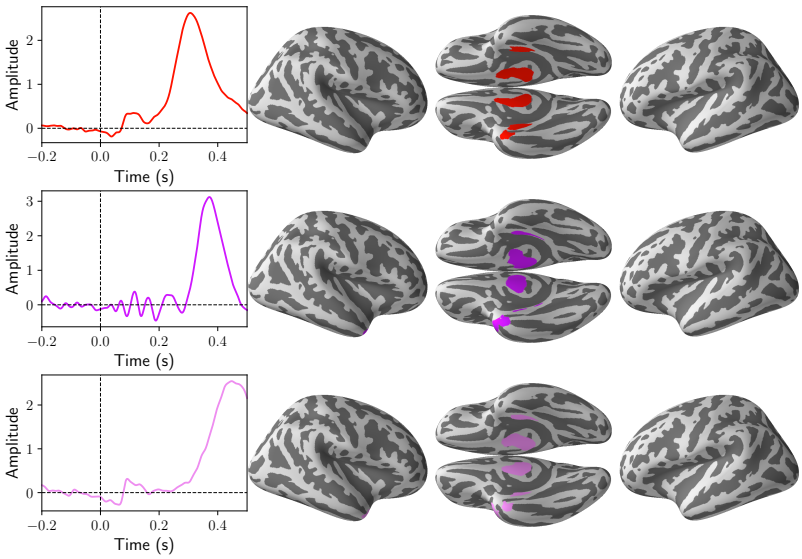
CanICA [289]	fMRI/Spatial	Individual PCA + multi set CCA (on component-wise concatenated data)	CanICA applies single-subject ICA on data reduced with PCA and CCA.
Spatial Concat-ICA [282]	fMRI/Spatial	Group PCA (on spatially concatenated data)	ICA is applied on spatially concatenated data. The mixing is constrained to be the same across all subjects.
Temporal ConcatICA [297]	EEG/Temporal	Group PCA (on temporally concatenated data)	ICA is applied on temporaly concatenated data. The mixing is constrained to be the same across all subjects.
coroICA [281]	Any	Any	The model is $\mathbf{x}^i = \mathbf{A}\mathbf{s}_i + \mathbf{n}^i$. The mixing is constrained to be the same across all subjects.

An additional related model is described in [83]. Similarly to our work, the ICA model has noise on the source side. However, the model involves nonlinear mixings, which are computationally unfeasible to optimize via maximum likelihood; a contrastive learning scheme is therefore adopted, and the likelihood is not derived in closed form. No evaluation on neuroimaging datasets is presented.

F.7 Detailed Cam-CAN sources

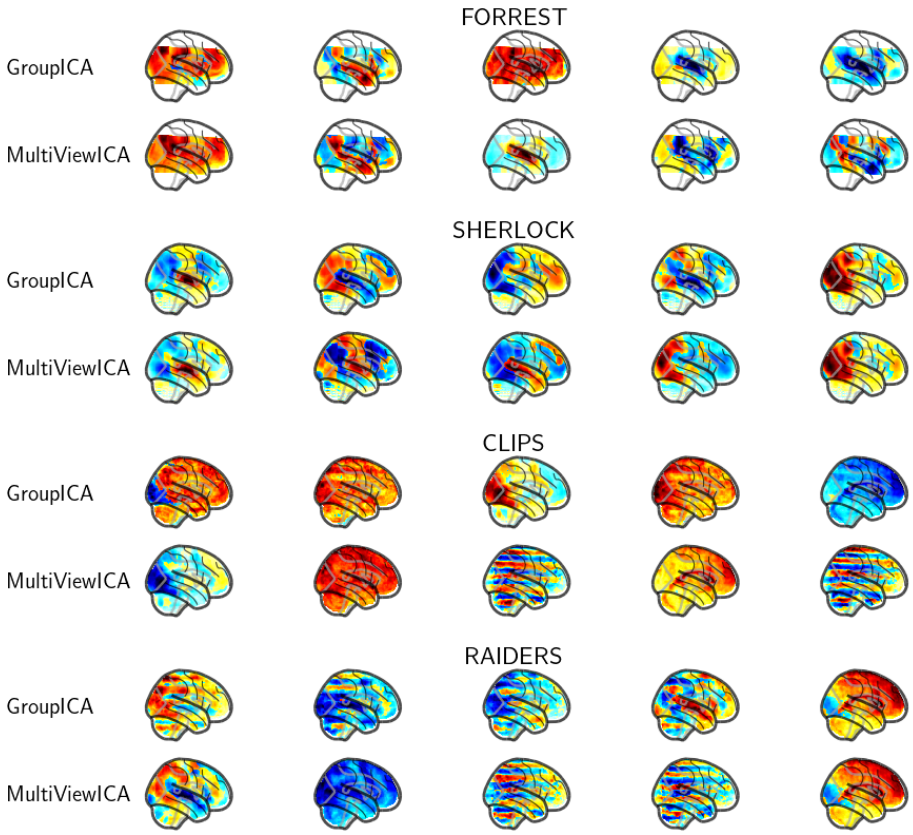
We display each of the 11 shared sources found by Multiview ICA on the Cam-CAN. The time-courses are on the left, the corresponding brain maps are on the right.





F.8 Average forward operators on fMRI datasets

We display the average forward operator across subjects on the Raiders, Forrest, Clips and Sherlock datasets obtained with MultiViewICA and GroupICA with 5 components. A 5 mm spatial smoothing was applied on all datasets, and the confound signals corresponding to the 5 components with the highest variance were removed before applying MultiViewICA or GroupICA.



F.9 Synthetic benchmark using the model $\mathbf{x}^i = \mathbf{A}^i\mathbf{s} + \mathbf{n}^i$

We generate data according to the model $\mathbf{x}^i = \mathbf{A}^i\mathbf{s} + \mathbf{n}^i$, where $\mathbf{x}^i \in \mathbb{R}^{50}$, $\mathbf{s} \in \mathbb{R}^{20}$, and $\mathbf{n}^i \sim \mathcal{N}(0, \sigma^2\mathbf{I}_{50})$. After applying individual PCA to obtain signals of dimension 20, we apply the different ICA algorithms and report the reconstruction error in fig. F.7.

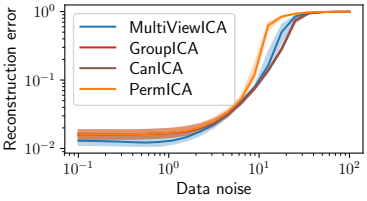


Figure F.7: Synthetic experiment with model $\mathbf{x}^i = \mathbf{A}^i\mathbf{s}^i + \mathbf{n}^i$

F.10 Summary of our quantitative results

Our quantitative results for the fMRI experiments of time-segment matching and BOLD signal reconstruction and on for the MEG phantom data experiment are summarized, respectively, in Table F.2, Table F.3 and Table F.4. All methods are compared upon extraction of sources with the same dimensionality (20 components).

Dataset	Method	Accuracy	Confidence interval
clips	Chance	0.002	[0.001, 0.003]
	CanICA	0.130	[0.112, 0.147]
	PCA + GroupICA	0.124	[0.109, 0.139]
	GroupICA	0.152	[0.133, 0.171]
	PermICA	0.147	[0.126, 0.169]
	SRM	0.115	[0.104, 0.126]
	MultiViewICA	**0.167**	[0.142, 0.192]
forrest	Chance	0.002	[0.001, 0.002]
	CanICA	0.192	[0.170, 0.214]
	PCA + GroupICA	0.088	[0.077, 0.098]
	GroupICA	0.154	[0.137, 0.170]
	PermICA	0.135	[0.118, 0.152]
	SRM	0.188	[0.173, 0.203]
	MultiViewICA	**0.448**	[0.411, 0.484]
raiders	Chance	0.002	[0.001, 0.003]
	CanICA	0.256	[0.220, 0.291]
	PCA + GroupICA	0.331	[0.289, 0.372]
	GroupICA	0.321	[0.281, 0.361]
	PermICA	0.381	[0.341, 0.421]
	SRM	0.265	[0.240, 0.289]
	MultiViewICA	**0.408**	[0.358, 0.458]
sherlock	Chance	0.005	[0.003, 0.006]
	CanICA	0.607	[0.567, 0.648]
	PCA + GroupICA	0.454	[0.416, 0.492]
	GroupICA	0.519	[0.481, 0.556]
	PermICA	0.399	[0.365, 0.434]
	SRM	0.493	[0.465, 0.520]
	MultiViewICA	**0.873**	[0.844, 0.903]

Table F.2: Timesegment matching: Summary of our quantitative results. We report the mean accuracy across cross-validation splits.

Dataset	Method	R2 score	Confidence interval
clips	Chance	0.000	[0.000 ,0.000]
	CanICA	0.110	[0.097 , 0.123]
	PCA + GroupICA	0.075	[0.058 , 0.092]
	GroupICA	0.077	[0.059 , 0.094]
	PermICA	0.099	[0.087 , 0.111]
	SRM	0.081	[0.069 , 0.094]
	MultiViewICA	**0.114**	[0.099 , 0.128]
forrest	Chance	0.000	[0.000 ,0.000]
	CanICA	0.181	[0.169 , 0.193]
	PCA + GroupICA	0.072	[0.054 , 0.090]
	GroupICA	0.081	[0.062 , 0.099]
	PermICA	0.098	[0.090 , 0.106]
	SRM	0.180	[0.168 , 0.193]
	MultiViewICA	**0.191**	[0.177 , 0.204]
raiders	Chance	0.000	[0.000 ,0.000]
	CanICA	0.136	[0.122 , 0.149]
	PCA + GroupICA	0.063	[0.045 , 0.080]
	GroupICA	0.062	[0.043 , 0.081]
	PermICA	0.107	[0.091 , 0.124]
	SRM	0.138	[0.121 , 0.154]
	MultiViewICA	**0.144**	[0.124 , 0.164]
sherlock	Chance	0.000	[0.000 ,0.000]
	CanICA	0.156	[0.141 , 0.172]
	PCA + GroupICA	0.087	[0.065 , 0.108]
	GroupICA	0.091	[0.070 , 0.112]
	PermICA	0.067	[0.055 , 0.078]
	SRM	**0.164**	[0.147 , 0.181]
	MultiViewICA	0.161	[0.142 , 0.180]

Table F.3: Reconstructing the BOLD signal of missing subjects: Summary of our quantitative results. We report the mean R2 score across cross-validation splits.

Method	Reconstruction error	1st and 3d quartiles
MultiViewICA	**0.0045**	[0.0039, 0.0052]
GroupICA	0.1098	[0.0549, 0.1734]
PCA+GroupICA	0.1111	[0.0760, 0.1502]
PermICA	0.0730	[0.0423, 0.1037]

Table F.4: Phantom MEG data: Summary of our quantitative results with 2 epochs. We report the median reconstruction error across cross-validation splits.

G

Additional Material on Chapter 8

Overview:

- Appendix G.1 contains the full proofs for all theoretical results from the main paper.
- Appendix G.2 contains additional details and plots for the *Causal3DIdent* dataset.
- Appendix G.3 contains additional experimental results and analysis.
- Appendix G.4 contains additional implementation details for our experiments.

G.1 Proofs

We now present the full detailed proofs of our three theorems which were briefly sketched in the main paper. We remark that these proofs build on each other, in the sense that the (main) step 2 of the proof of Thm. 8.4.1 is also used in the proofs of Thms. 8.4.2 and 8.4.3.

G.1.1 Proof of Thm. 8.4.1

Theorem 8.4.1 (Identifying content with a generative model) *Consider the data generating process described in § 8.3, i.e., the pairs* $(\mathbf{x}, \tilde{\mathbf{x}})$ *of original and augmented views are generated according to* (8.2) *and* (8.3) *with* $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ *as defined in Assumptions 8.3.1 and 8.3.2. Assume further that*

(i) $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ *is smooth and invertible with smooth inverse (i.e., a diffeomorphism);*

(ii) $p_{\mathbf{z}}$ *is a smooth, continuous density on* $\mathcal{Z}$ *with* $p_{\mathbf{z}}(\mathbf{z}) > 0$ *almost everywhere;*

(iii) for any $l \in \{1, \dots, n_s\}$, $\exists A \subseteq \{1, \dots, n_s\}$ *s.t.* $l \in A$; $p_A(A) > 0$; $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ *is smooth w.r.t. both* $\mathbf{s}_A$ *and* $\tilde{\mathbf{s}}_A$; *and for any* $\mathbf{s}_A$, $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\cdot|\mathbf{s}_A) > 0$ *in some open, non-empty subset containing* $\mathbf{s}_A$.

If, for a given n_s $(1 \leq n_s < n)$, *a generative model* $(\hat{p}_{\mathbf{z}}, \hat{p}_A, \hat{p}_{\tilde{\mathbf{s}}|\mathbf{s},A}, \hat{\mathbf{f}})$ *assumes the same generative process (§ 8.3), satisfies the above assumptions (i)-(iii), and matches the data likelihood,*

$$p_{\mathbf{x},\tilde{\mathbf{x}}}(\mathbf{x}, \tilde{\mathbf{x}}) = \hat{p}_{\mathbf{x},\tilde{\mathbf{x}}}(\mathbf{x}, \tilde{\mathbf{x}}) \qquad \forall (\mathbf{x}, \tilde{\mathbf{x}}) \in \mathcal{X} \times \mathcal{X},$$

then it block-identifies the true content variables via $\mathbf{g} = \hat{\mathbf{f}}^{-1}$ *in the sense of Defn. 8.4.1.*

Proof. The proof consists of two main steps.

In the first step, we use assumption *(i)* and the matching likelihoods to show that the representation $\hat{\mathbf{z}} = \mathbf{g}(\mathbf{x})$ extracted by $\mathbf{g} = \hat{\mathbf{f}}^{-1}$ is related to the true latent $\mathbf{z}$ by a smooth invertible mapping $\mathbf{h}$, and that $\hat{\mathbf{z}}$ must satisfy invariance across $(\mathbf{x}, \tilde{\mathbf{x}})$ in the first n_c (content) components almost surely (a.s.) with respect to (w.r.t.) the true generative process.

In the second step, we then use assumptions *(ii)* and *(iii)* to prove (by contradiction) that $\hat{\mathbf{c}} := \hat{\mathbf{z}}_{1:n_c} = \mathbf{h}(\mathbf{z})_{1:n_c}$ can, in fact, only depend on the true content $\mathbf{c}$ and not on the true style $\mathbf{s}$, for otherwise the invariance established in the first step would have be violated with probability greater than zero.

To provide some further intuition for the second step, the assumed generative process implies that $(\mathbf{c}, \tilde{\mathbf{s}})|A$ is constrained to take values (a.s.) in a subspace $\mathcal{R}$ of $\mathcal{C} \times \mathcal{S} \times \mathcal{S}$ of dimension $n_c + n_s + |A|$ (as opposed to dimension $n_c + 2n_s$ for $\mathcal{C} \times \mathcal{S} \times \mathcal{S}$). In this context, assumption *(iii)* implies that $(\mathbf{c}, \tilde{\mathbf{s}})|A$ has a density with respect to a measure on this subspace equivalent to the Lebesgue measure on $\mathbb{R}^{n_c+n_s+|A|}$. This equivalence implies, in particular, that this "subspace measure" is strictly positive: it takes strictly positive values on open sets of $\mathcal{R}$ seen as a topological subspace of $\mathcal{C} \times \mathcal{S} \times \mathcal{S}$. These open sets are defined by the induced topology: they are the intersection of the open sets of $\mathcal{C} \times \mathcal{S} \times \mathcal{S}$ with $\mathcal{R}$. An open set B of V on which $p(\mathbf{c}, \tilde{\mathbf{s}}|A) > 0$ then satisfies $P(B|A) > 0$. We look for such an open set to prove our result.

Step 1. From the assumed data generating process described in § 8.3 it follows that

$$\mathbf{g}(\mathbf{x})_{1:n_c} = \mathbf{g}(\tilde{\mathbf{x}})_{1:n_c} \tag{G.1}$$

a.s., i.e., with probability one, w.r.t. the model distribution $\hat{p}_{\mathbf{x},\tilde{\mathbf{x}}}$.

Due to the assumption of matching likelihoods, the invariance in (G.1) must also hold (a.s.) w.r.t. the true data distribution $p_{\mathbf{x},\tilde{\mathbf{x}}}$.

Next, since $\mathbf{f}, \hat{\mathbf{f}} : \mathcal{Z} \to \mathcal{X}$ are smooth and invertible functions by assumption *(i)*, there exists a smooth and invertible function $\mathbf{h} = \mathbf{g} \circ \mathbf{f} : \mathcal{Z} \to \mathcal{Z}$ such that

$$\mathbf{g} = \mathbf{h} \circ \mathbf{f}^{-1}. \tag{G.2}$$

Substituting (G.2) into (G.1), we obtain (a.s. w.r.t. p):

$$\hat{\mathbf{c}} := \hat{\mathbf{z}}_{1:n_c} = \mathbf{g}(\mathbf{x})_{1:n_c} = \mathbf{h}(\mathbf{f}^{-1}(\mathbf{x}))_{1:n_c} = \mathbf{h}(\mathbf{f}^{-1}(\tilde{\mathbf{x}}))_{1:n_c} \tag{G.3}$$

Substituting $\mathbf{z} = \mathbf{f}^{-1}(\mathbf{x})$ and $\tilde{\mathbf{z}} = \mathbf{f}^{-1}(\tilde{\mathbf{x}})$ into (G.3), we obtain (a.s. w.r.t. p)

$$\hat{\mathbf{c}} = \mathbf{h}(\mathbf{z})_{1:n_c} = \mathbf{h}(\tilde{\mathbf{z}})_{1:n_c}. \tag{G.4}$$

It remains to show that $\mathbf{h}(\cdot)_{1:n_c}$ can only be a function of $\mathbf{c}$, i.e., does not depend on any other (style) dimension of $\mathbf{z} = (\mathbf{c}, \mathbf{s})$.

Step 2. Suppose *for a contradiction* that $\mathbf{h}_c(\mathbf{c}, \mathbf{s}) := \mathbf{h}(\mathbf{c}, \mathbf{s})_{1:n_c} = \mathbf{h}(\mathbf{z})_{1:n_c}$ depends on some component of the style variable $\mathbf{s}$:

$$\exists l \in \{1, ..., n_s\}, (\mathbf{c}^*, \mathbf{s}^*) \in \mathcal{C} \times \mathcal{S}, \qquad \text{s.t.} \qquad \frac{\partial \mathbf{h}_c}{\partial s_l}(\mathbf{c}^*, \mathbf{s}^*) \neq 0, \tag{G.5}$$

that is, we assume that the partial derivative of $\mathbf{h}_c$ w.r.t. some style variable s_l is non-zero at some point $\mathbf{z}^* = (\mathbf{c}^*, \mathbf{s}^*) \in \mathcal{Z} = \mathcal{C} \times \mathcal{S}$.

Since $\mathbf{h}$ is smooth, so is $\mathbf{h}_c$. Therefore, $\mathbf{h}_c$ has continuous (first) partial derivatives.

By continuity of the partial derivative, $\frac{\partial \mathbf{h}_c}{\partial s_l}$ must be non-zero in a neighbourhood of $(\mathbf{c}^*, \mathbf{s}^*)$, i.e.,

$$\exists \eta > 0 \qquad \text{s.t.} \qquad s_l \mapsto \mathbf{h}_c\big(\mathbf{c}^*, (\mathbf{s}^*_{-l}, s_l)\big) \quad \text{is strictly monotonic on} \quad (s_l^* - \eta, s_l^* + \eta), \tag{G.6}$$

where $\mathbf{s}_{-l} \in \mathcal{S}_{-l}$ denotes the vector of remaining style variables except s_l.

Next, define the auxiliary function $\psi : \mathcal{C} \times \mathcal{S} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ as follows:

$$\psi(\mathbf{c}, \tilde{\mathbf{s}}) := |\mathbf{h}_c(\mathbf{c}, \mathbf{s}) - \mathbf{h}_c(\mathbf{c}, \tilde{\mathbf{s}})| \geq 0\,. \tag{G.7}$$

To obtain a contradiction to the invariance condition (G.4) from Step 1 under assumption (G.5), it remains to show that ψ from (G.7) is *strictly positive* with probability greater than zero (w.r.t. p).

First, the strict monotonicity from (G.6) implies that

$$\psi\big(\mathbf{c}^*, (\mathbf{s}^*_{-l}, s_l), (\mathbf{s}^*_{-l}, \tilde{s}_l)\big) > 0\,, \quad \forall (s_l, \tilde{s}_l) \in (s_l^*, s_l^* + \eta) \times (s_l^* - \eta, s_l^*)\,. \tag{G.8}$$

Note that in order to obtain the strict inequality in (G.8), it is important that s_l and $\tilde{s}_l$ take values in *disjoint* open subsets of the interval $(s_l^* - \eta, s_l^* + \eta)$ from (G.6).

Since ψ is a composition of continuous functions (absolute value of the difference of two continuous functions), ψ is continuous.

Consider the open set $\mathbb{R}_{>0}$, and recall that, under a continuous function, pre-images (or inverse images) of open sets are always *open*.

Applied to the continuous function ψ, this pre-image corresponds to an *open* set

$$\mathcal{U} \subseteq \mathcal{C} \times \mathcal{S} \times \mathcal{S} \tag{G.9}$$

in the domain of ψ on which ψ is strictly positive.

Moreover, due to (G.8):

$$\{\mathbf{c}^*\} \times \big(\{\mathbf{s}^*_{-l}\} \times (s_l^*, s_l^* + \eta)\big) \times \big(\{\mathbf{s}^*_{-l}\} \times (s_l^* - \eta, s_l^*)\big) \subset \mathcal{U}, \tag{G.10}$$

so $\mathcal{U}$ is *non-empty*.

Next, by assumption *(iii)*, there exists at least one subset $A \subseteq \{1, ..., n_s\}$ of changing style variables such that $l \in A$ and $p_A(A) > 0$; pick one such subset and call it A.

Then, also by assumption *(iii)*, for any $\mathbf{s}_A \in \mathcal{S}_A$, there is an open subset $\mathcal{O}(\mathbf{s}_A) \subseteq \mathcal{S}_A$ containing $\mathbf{s}_A$, such that $p_{\tilde{\mathbf{s}}_A | \mathbf{s}_A}(\cdot | \mathbf{s}_A) > 0$ within $\mathcal{O}(\mathbf{s}_A)$.

Define the following space

$$\mathcal{R}_A := \{(\mathbf{s}_A, \tilde{\mathbf{s}}_A) : \mathbf{s}_A \in \mathcal{S}_A, \tilde{\mathbf{s}}_A \in \mathcal{O}(\mathbf{s}_A)\} \tag{G.11}$$

and, recalling that $A^c = \{1, ..., n_s\} \setminus A$ denotes the complement of A, define

$$\mathcal{R} := \mathcal{C} \times \mathcal{S}_{A^c} \times \mathcal{R}_A \tag{G.12}$$

which is a topological subspace of $\mathcal{C} \times \mathcal{S} \times \mathcal{S}$.

By assumptions *(ii)* and *(iii)*, $p_{\mathbf{z}}$ is smooth and fully supported, and $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\cdot|\mathbf{s}_A)$ is smooth and fully supported on $\mathcal{O}(\mathbf{s}_A)$ for any $\mathbf{s}_A \in \mathcal{S}_A$. Therefore, the measure $\mu_{(\mathbf{c},\mathbf{s}_{A^c},\mathbf{s}_A,\tilde{\mathbf{s}}_A)|A}$ has fully supported, strictly-positive density on $\mathcal{R}$ w.r.t. a strictly positive measure on $\mathcal{R}$. In other words, $p_{\mathbf{z}} \times p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ is fully supported (i.e., strictly positive) on $\mathcal{R}$.

Now consider the intersection $\mathcal{U} \cap \mathcal{R}$ of the open set $\mathcal{U}$ with the topological subspace $\mathcal{R}$.

Since $\mathcal{U}$ is open, by the definition of topological subspaces, the intersection $\mathcal{U} \cap \mathcal{R} \subseteq \mathcal{R}$ is *open* in $\mathcal{R}$, (and thus has the same dimension as $\mathcal{R}$ if non-empty).

Moreover, since $\mathcal{O}(\mathbf{s}_A^*)$ is open containing $\mathbf{s}_A^*$, there exists $\eta' > 0$ such that $\{\mathbf{s}_{-l}^*\} \times (s_l^* - \eta', s_l^*) \subset \mathcal{O}(\mathbf{s}_A^*)$. Thus, for $\eta'' = \min(\eta, \eta') > 0$,

$$\{\mathbf{c}^*\} \times \{\mathbf{s}_{A^c}^*\} \times \left(\{\mathbf{s}_{A\setminus\{l\}}^*\} \times (s_l^*, s_l^* + \eta)\right) \times \left(\{\mathbf{s}_{A\setminus\{l\}}^*\} \times (s_l^* - \eta'', s_l^*)\right) \subset \mathcal{R}. \tag{G.13}$$

In particular, this implies that

$$\{\mathbf{c}^*\} \times \left(\{\mathbf{s}_{-l}^*\} \times (s_l^*, s_l^* + \eta)\right) \times \left(\{\mathbf{s}_{-l}^*\} \times (s_l^* - \eta'', s_l^*)\right) \subset \mathcal{R}, \tag{G.14}$$

Now, since $\eta'' \leq \eta$, the LHS of (G.14) is also in $\mathcal{U}$ according to (G.10), so the intersection $\mathcal{U} \cap \mathcal{R}$ is *non-empty*.

In summary, the intersection $\mathcal{U} \cap \mathcal{R} \subseteq \mathcal{R}$:

- is non-empty (since both $\mathcal{U}$ and $\mathcal{R}$ contain the LHS of (G.10));
- is an open subset of the topological subspace $\mathcal{R}$ of $\mathcal{C} \times \mathcal{S} \times \mathcal{S}$ (since it is the intersection of an open set, $\mathcal{U}$, with $\mathcal{R}$);
- satisfies $\psi > 0$ (since this holds for all of $\mathcal{U}$);
- is fully supported w.r.t. the generative process (since this holds for all of $\mathcal{R}$).

As a consequence,

$$\mathbb{P}\left(\psi(\mathbf{c}, \tilde{\mathbf{s}}) > 0|A\right) \geq \mathbb{P}(\mathcal{U} \cap \mathcal{R}) > 0, \tag{G.15}$$

where $\mathbb{P}$ denotes probability w.r.t. the true generative process p.

Since $p_A(A) > 0$, this is a **contradiction** to the invariance (G.4) from Step 1.

Hence, assumption (G.5) cannot hold, i.e., $\mathbf{h}_c(\mathbf{c}, \mathbf{s})$ does not depend on any style variable s_l. It is thus only a function of $\mathbf{c}$, i.e., $\hat{\mathbf{c}} = \mathbf{h}_c(\mathbf{c})$.

Finally, smoothness and invertibility of $\mathbf{h}_c : \mathcal{C} \to \mathcal{C}$ follow from smoothness and invertibility of $\mathbf{h}$, as established in Step 1.

This concludes the proof that $\hat{\mathbf{c}}$ is related to the true content $\mathbf{c}$ via a smooth invertible mapping. □

G.1.2 Proof of Thm. 8.4.2

Theorem 8.4.2 (Identifying content with an invertible encoder) *Assume the same data generating process (§ 8.3) and conditions (i)-(iv) as in Thm. 8.4.1. Let* $\mathbf{g} : \mathcal{X} \to \mathcal{Z}$ *be any smooth and* invertible *function which minimises the following functional:*

$$\mathcal{L}_{\text{Align}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathbf{x},\tilde{\mathbf{x}}}} \left[\left\| \mathbf{g}(\mathbf{x})_{1:n_c} - \mathbf{g}(\tilde{\mathbf{x}})_{1:n_c} \right\|_2^2 \right] \tag{8.4}$$

Then $\mathbf{g}$ *block-identifies the true content variables in the sense of Definition 8.4.1.*

Proof. As in the proof of Thm. 8.4.1, the proof again consists of two main steps.

In the first step, we show that the representation $\hat{\mathbf{z}} = \mathbf{g}(\mathbf{x})$ extracted by any $\mathbf{g}$ that minimises $\mathcal{L}_{\text{Align}}$ is related to the true latent $\mathbf{z}$ through a smooth invertible mapping $\mathbf{h}$, and that $\hat{\mathbf{z}}$ must satisfy invariance across $(\mathbf{x}, \tilde{\mathbf{x}})$ in the first n_c (content) components almost surely (a.s.) with respect to (w.r.t.) the true generative process.

In the second step, we use the same argument by contradiction as in Step 2 of the proof of Thm. 8.4.1, to show that $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{z})_{1:n_c}$ can only depend on the true content $\mathbf{c}$ and not on style $\mathbf{s}$.

Step 1. From the form of the objective (8.4), it is clear that $\mathcal{L}_{\text{Align}} \geq 0$ with equality if and only if $\mathbf{g}(\tilde{\mathbf{x}})_{1:n_c} = \mathbf{g}(\mathbf{x})_{1:n_c}$ for all $(\mathbf{x}, \tilde{\mathbf{x}})$ s.t. $p_{\mathbf{x},\tilde{\mathbf{x}}}(\mathbf{x}, \tilde{\mathbf{x}}) > 0$.

Moreover, it follows from the assumed generative process that the global minimum of zero is attained by the true unmixing $\mathbf{f}^{-1}$ since

$$\mathbf{f}^{-1}(\mathbf{x})_{1:n_c} = \mathbf{c} = \tilde{\mathbf{c}} = \mathbf{f}^{-1}(\tilde{\mathbf{x}})_{1:n_c} \tag{G.16}$$

holds a.s. (i.e., with probability one) w.r.t. the true generative process p.

Hence, there exists at least one smooth invertible function ($\mathbf{f}^{-1}$) which attains the global minimum.

Let $\mathbf{g}$ be *any* function attaining the global minimum of $\mathcal{L}_{\text{Align}}$ of zero.

As argued above, this implies that (a.s. w.r.t. p):

$$\mathbf{g}(\tilde{\mathbf{x}})_{1:n_c} = \mathbf{g}(\mathbf{x})_{1:n_c}. \tag{G.17}$$

Writing $\mathbf{g} = \mathbf{h} \circ \mathbf{f}^{-1}$, where $\mathbf{h}$ is the smooth, invertible function $\mathbf{h} = \mathbf{g} \circ \mathbf{f}$ we obtain (a.s. w.r.t. p):

$$\hat{\mathbf{c}} = \mathbf{h}(\tilde{\mathbf{z}})_{1:n_c} = \mathbf{h}(\mathbf{z})_{1:n_c}. \tag{G.18}$$

Note that this is the same invariance condition as (G.4) derived in Step 1 of the proof of Thm. 8.4.1.

Step 2. It remains to show that $\mathbf{h}(\mathbf{z})_{1:n_c}$ can only depend on the true content $\mathbf{c}$ and not on any of the style variables $\mathbf{s}$. To show this, we use the same Step 2 as in the proof of Thm. 8.4.1. □

G.1.3 Proof of Thm. 8.4.3

Theorem 8.4.3 (Identifying content with discriminative learning and a non-invertible encoder) *Assume the same data generating process (§ 8.3) and conditions (i)-(iv) as in Thm. 8.4.1. Let $\mathbf{g} : \mathcal{X} \to (0,1)^{n_c}$ be any smooth function which minimises the following functional:*

$$\mathscr{L}_{\text{AlignMaxEnt}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathbf{x},\tilde{\mathbf{x}}}} \left[\left\| \mathbf{g}(\mathbf{x}) - \mathbf{g}(\tilde{\mathbf{x}}) \right\|_2^2 \right] - H\left(\mathbf{g}(\mathbf{x})\right) \tag{8.5}$$

where $H(\cdot)$ denotes the differential entropy of the random variable $\mathbf{g}(\mathbf{x})$ taking values in $(0,1)^{n_c}$. Then $\mathbf{g}$ block-identifies the true content variables in the sense of Defn. 8.4.1.

Proof. The proof consists of three main steps.

In the first step, we show that the representation $\hat{\mathbf{c}} = \mathbf{g}(\mathbf{x})$ extracted by any smooth function $\mathbf{g}$ that minimises (8.5) is related to the true latent $\mathbf{z}$ through a smooth mapping $\mathbf{h}$; that $\hat{\mathbf{c}}$ must satisfy invariance across $(\mathbf{x}, \tilde{\mathbf{x}})$ almost surely (a.s.) with respect to (w.r.t.) the true generative process p; and that $\hat{\mathbf{c}}$ must follow a uniform distribution on $(0,1)^{n_c}$.

In the second step, we use the same argument by contradiction as in Step 2 of the proof of Thm. 8.4.1, to show that $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{z})$ can only depend on the true content $\mathbf{c}$ and not on style $\mathbf{s}$.

Finally, in the third step, we show that $\mathbf{h}$ must be a bijection, i.e., invertible, using a result from [88].

Step 1. The global minimum of $\mathscr{L}_{\text{AlignMaxEnt}}$ is reached when the first term (alignment) is minimised (i.e., equal to zero) and the second term (entropy) is maximised.

Without additional moment constraints, the *unique* maximum entropy distribution on $(0,1)^{n_c}$ is the uniform distribution [211, 526].

First, we show that there exists a smooth function $\mathbf{g}^* : \mathcal{X} \to (0,1)^{n_c}$ which attains the global minimum of $\mathscr{L}_{\text{AlignMaxEnt}}$.

To see this, consider the function $\mathbf{f}^{-1}_{1:n_c} : \mathcal{X} \to \mathcal{C}$, i.e., the inverse of the true mixing $\mathbf{f}$, restricted to its first n_c dimensions. This exists and is smooth since $\mathbf{f}$ is smooth and invertible by assumption *(i)*. Further, we have $\mathbf{f}^{-1}(\mathbf{x})_{1:n_c} = \mathbf{c}$ by definition.

We now build a function $\mathbf{d} : \mathcal{C} \to (0,1)^{n_c}$ which maps $\mathbf{c}$ to a uniform random variable on $(0,1)^{n_c}$ using a recursive construction known as the *Darmois construction* [69, 70].

Specifically, we define

$$d_i(\mathbf{c}) := F_i(c_i|\mathbf{c}_{1:i-1}) = \mathbb{P}(C_i \leq c_i|\mathbf{c}_{1:i-1}), \qquad i = 1, \ldots, n_c, \tag{G.19}$$

where F_i denotes the conditional cumulative distribution function (CDF) of c_i given $\mathbf{c}_{1:i-1}$.

By construction, $\mathbf{d}(\mathbf{c})$ is uniformly distributed on $(0,1)^{n_c}$ [69, 70].

Further, $\mathbf{d}$ is smooth by the assumption that $p_{\mathbf{z}}$ (and thus $p_{\mathbf{c}}$) is a smooth density.

Finally, we define

$$\mathbf{g}^* := \mathbf{d} \circ \mathbf{f}^{-1}_{1:n_c} : \mathcal{X} \to (0,1)^{n_c}, \tag{G.20}$$

which is a smooth function since it is a composition of two smooth functions.

Claim G.1.1 $\mathbf{g}^*$ as defined in (G.20) attains the global minimum of $\mathcal{L}_{\text{AlignMaxEnt}}$.

Proof of Claim G.1.1. Using $\mathbf{f}^{-1}(\mathbf{x})_{1:n_c} = \mathbf{c}$ and $\mathbf{f}^{-1}(\tilde{\mathbf{x}})_{1:n_c} = \tilde{\mathbf{c}}$, we have

$$\mathcal{L}_{\text{AlignMaxEnt}}(\mathbf{g}^*) = \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{(\mathbf{x},\tilde{\mathbf{x}})}} \left[\left\| \mathbf{g}^*(\mathbf{x}) - \mathbf{g}^*(\tilde{\mathbf{x}}) \right\|_2^2 \right] - H\left(\mathbf{g}^*(\mathbf{x})\right) \tag{G.21}$$

$$= \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{(\mathbf{x},\tilde{\mathbf{x}})}} \left[\left\| \mathbf{d}(\mathbf{c}) - \mathbf{d}(\tilde{\mathbf{c}}) \right\|_2^2 \right] - H\left(\mathbf{d}(\mathbf{c})\right) \tag{G.22}$$

$$= 0 \tag{G.23}$$

where in the last step we have used the fact that $\mathbf{c} = \tilde{\mathbf{c}}$ almost surely w.r.t. to the ground truth generative process p described in § 8.3, so the first term is zero; and the fact that $\mathbf{d}(\mathbf{c})$ is uniformly distributed on $(0,1)^{n_c}$ and the uniform distribution on the unit hypercube has zero entropy, so the second term is also zero.

Next, let $\mathbf{g} : \mathcal{X} \to (0,1)^{n_c}$ be *any* smooth function which attains the global minimum of (8.5), i.e.,

$$\mathcal{L}_{\text{AlignMaxEnt}}(\mathbf{g}) = \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{(\mathbf{x},\tilde{\mathbf{x}})}} \left[\left\| \mathbf{g}(\mathbf{x}) - \mathbf{g}(\tilde{\mathbf{x}}) \right\|_2^2 \right] - H\left(\mathbf{g}(\mathbf{x})\right) = 0. \tag{G.24}$$

Define $\mathbf{h} := \mathbf{g} \circ \mathbf{f} : \mathcal{Z} \to (0,1)^{n_c}$ which is smooth because both $\mathbf{g}$ and $\mathbf{f}$ are smooth.

Writing $\mathbf{x} = \mathbf{f}(\mathbf{z})$, (G.24) then implies in terms of $\mathbf{h}$:

$$\mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{(\mathbf{x},\tilde{\mathbf{x}})}} \left[\left\| \mathbf{h}(\mathbf{z}) - \mathbf{h}(\tilde{\mathbf{z}}) \right\|_2^2 \right] = 0\,, \tag{G.25}$$

$$H\left(\mathbf{h}(\mathbf{z})\right) = 0\,. \tag{G.26}$$

Equation (G.25) implies that the same invariance condition (G.4) used in the proofs of Thms. 8.4.1 and 8.4.2 must hold (a.s. w.r.t. p), and (G.26) implies that $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{z})$ must be uniformly distributed on $(0,1)^{n_c}$.

Step 2. Next, we show that $\mathbf{h}(\mathbf{z}) = \mathbf{h}(\mathbf{c}, \mathbf{s})$ can only depend on the true content $\mathbf{c}$ and not on any of the style variables $\mathbf{s}$. For this we use the same Step 2 as in the proofs of Thms. 8.4.1 and 8.4.2.

Step 3. Finally, we show that the mapping $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{c})$ is invertible.

To this end, we make use of the following result from [88].

Proposition G.1.2 (Proposition 5 of [88]) *Let $\mathcal{M}, \mathcal{N}$ be simply connected and oriented $\mathcal{C}^1$ manifolds without boundaries and $h : \mathcal{M} \to \mathcal{N}$ be a differentiable map. Further, let the random variable $\mathbf{z} \in \mathcal{M}$ be distributed according to $\mathbf{z} \sim p(\mathbf{z})$ for a regular density function p, i.e., $0 < p < \infty$. If the pushforward $p_{\#h}(\mathbf{z})$ of p through h is also a regular density, i.e., $0 < p_{\#h} < \infty$, then h is a bijection.*

We apply this result to the simply connected and oriented $\mathcal{C}^1$ manifolds without boundaries $\mathcal{M} = \mathcal{C}$ and $\mathcal{N} = (0,1)^{n_c}$, and the smooth (hence, differentiable) map $\mathbf{h} : \mathcal{C} \to (0,1)^{n_c}$ which maps the random variable $\mathbf{c}$ to a uniform random variable $\hat{\mathbf{c}}$ (as established in Step 1).

Since both $p_{\mathbf{c}}$ (by assumption) and the uniform distribution (the pushforward of $p_{\mathbf{c}}$ through $\mathbf{h}$) are regular densities in the sense of Proposition G.1.2, we conclude that $\mathbf{h}$ is a bijection, i.e., invertible.

We have shown that for any smooth $\mathbf{g} : \mathcal{X} \to (0,1)^{n_c}$ which minimises $\mathcal{L}_{\text{AlignMaxEnt}}$, we have that $\hat{\mathbf{c}} = \mathbf{g}(\mathbf{x}) = \mathbf{h}(\mathbf{c})$ for a smooth and invertible $\mathbf{h} : \mathcal{C} \to (0,1)^{n_c}$, i.e., $\mathbf{c}$ is block-identified by $\mathbf{g}$. □

G.2 Additional details on the Causal3DIdent data set

Using the Blender rendering engine [527], 3DIdent [88] is a recently proposed benchmark which contains hallmarks of natural environments (e.g. shadows, different lighting conditions, a 3D object), but allows for identifiability evaluation by exposing the underlying generative factors.

Each $224 \times 224 \times 3$ image in the dataset shows a coloured 3D object which is located and rotated above a coloured ground in a 3D space. Furthermore, each scene contains a coloured spotlight which is focused on the object and located on a half-circle around the scene. The images are rendered based on a 10-dimensional latent, where: (i) three dimensions describe the XYZ position of the object, (ii) three dimensions describe the rotation of the object in Euler angles, (iii) two dimensions describe the colour (hue) of the object and the ground of the scene, respectively, and (iv) two dimensions describe the position and colour (hue) of the spotlight. For influence of the latent factors on the renderings, see Fig. 2 of [88].

G.2.1 Details on introduced object classes

3DIdent contained a single object class, Teapot [383]. We add **six** additional object classes: Hare [384], Dragon [385], Cow [386], Armadillo [387], Horse [388], Head [389].

G.2.2 Details on latent causal graph

In 3DIdent, the latents are uniformly sampled independently. We instead impose a causal graph over the variables (see Fig. 8.2). While object class and all environment variables (spotlight position, spotlight hue, background hue) are sampled independently, all object variables are dependent. Specifically, for spotlight position, spotlight hue, and background hue, we sample from $U(-1, 1)$. We impose the dependence by varying the mean (μ) of a truncated normal distribution with standard deviation $\sigma = 0.5$, truncated to the range $[-1, 1]$.

Object rotation is dependent solely on object class, see Tab. G.1 for details. Object position is dependent on both object class & spotlight position, see Tab. G.2. Object hue is dependent on object class, background hue, & object hue, see Tab. G.3. Hares blending into their environment as a form of active camouflage has been observed in Alaskan [390], Arctic [391], & Snowshoe hares.

object class	$\mu(\phi)$	$\mu(\theta)$	$\mu(\psi)$
Teapot	-0.35	0.35	0.35
Hare	0.35	-0.35	0.35
Dragon	0.35	0.35	-0.35
Cow	0.35	-0.35	-0.35
Armadillo	-0.35	0.35	-0.35
Horse	-0.35	-0.35	0.35
Head	-0.35	-0.35	-0.35

Table G.1: Given a certain object class, the center of the truncated normal distribution from which we sample *rotation* latents varies.

object class	$\mu(x)$	$\mu(y)$	$\mu(z)$
Teapot	0	0	0
Hare	$-\sin(\text{pos}_{\text{spl}})$	$-\cos(\text{pos}_{\text{spl}})$	0
Dragon	$-\sin(\text{pos}_{\text{spl}})$	$-\cos(\text{pos}_{\text{spl}})$	0
Cow	$\sin(\text{pos}_{\text{spl}})$	$\cos(\text{pos}_{\text{spl}})$	0
Armadillo	$\sin(\text{pos}_{\text{spl}})$	$\cos(\text{pos}_{\text{spl}})$	0
Horse	$-\sin(\text{pos}_{\text{spl}})$	$-\cos(\text{pos}_{\text{spl}})$	0
Head	$\sin(\text{pos}_{\text{spl}})$	$\cos(\text{pos}_{\text{spl}})$	0

Table G.2: Given a certain object class & spotlight position, the center of the truncated normal distribution from which we sample *xy-position* latents varies. Note the spotlight position pos_{spl} is rescaled from $[-1, 1]$ to $[-\pi/2, \pi/2]$.

G.2.3 Dataset Visuals

We show 40 random samples from the marginal of each object class in Causal3DIdent in Figs. G.1 to G.7.

object class	μ(hue)
Teapot	0
Hare	$\frac{\text{hue}_{\text{bg}}+\text{hue}_{\text{spl}}}{2}$
Dragon	$-\frac{\text{hue}_{\text{bg}}+\text{hue}_{\text{spl}}}{2}$
Cow	-0.35
Armadillo	0.7
Horse	-0.7
Head	0.35

Table G.3: Given a certain object class, background hue, and spotlight hue, the center of the truncated normal distribution from which we sample the *object hue* latent varies. Note that for the Hare and Dragon classes, in particular, the object either blends in or stands out from the environment.

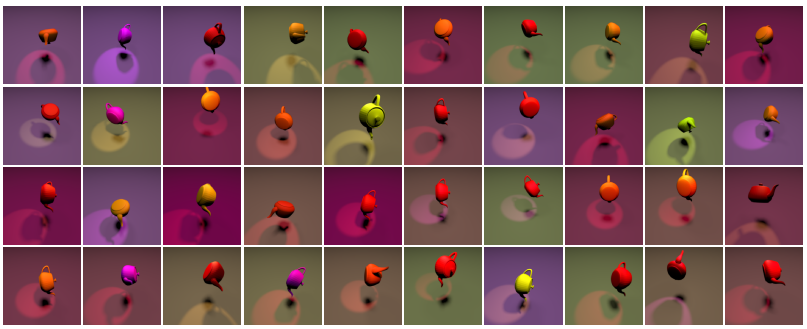


Figure G.1: 40 random samples from the marginal distribution of the *Teapot* object class.

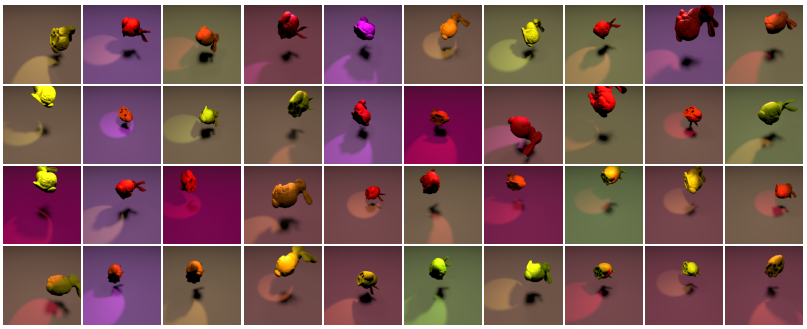


Figure G.2: 40 random samples from the marginal distribution of the *Hare* object class.

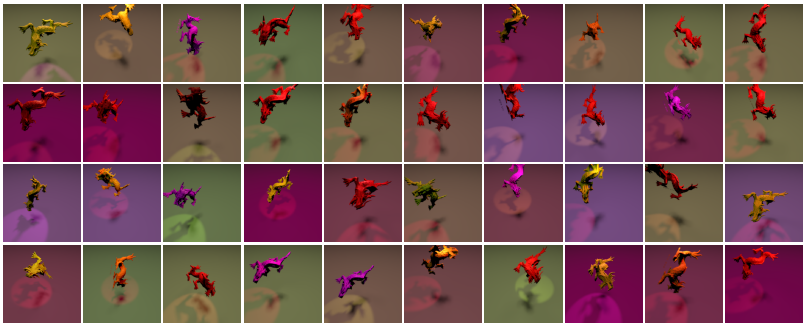


Figure G.3: 40 random samples from the marginal distribution of the *Dragon* object class.

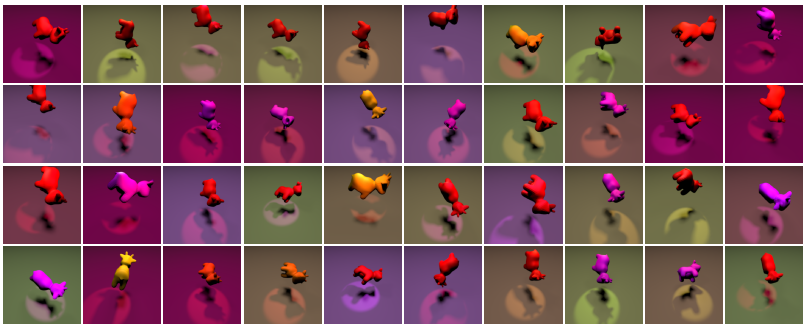


Figure G.4: 40 random samples from the marginal distribution of the *Cow* object class.

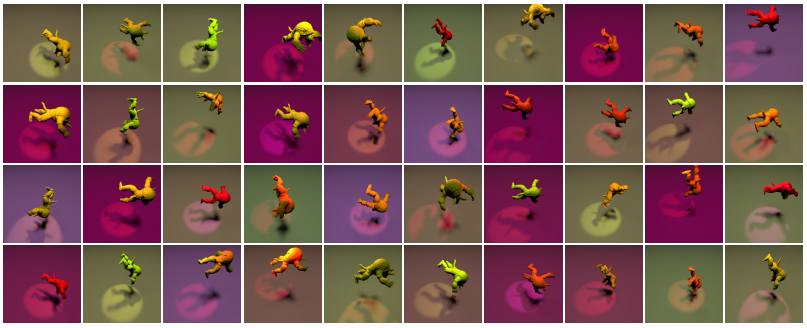


Figure G.5: 40 random samples from the marginal distribution of the *Armadillo* object class.

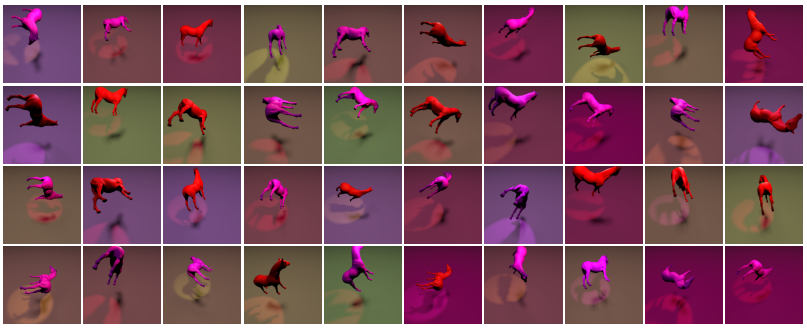


Figure G.6: 40 random samples from the marginal distribution of the *Horse* object class.

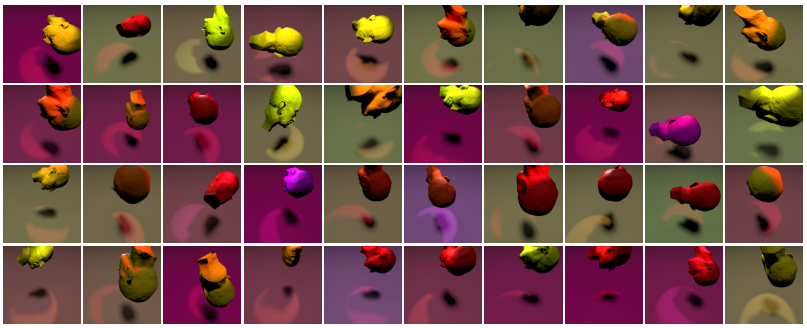


Figure G.7: 40 random samples from the marginal distribution of the *Head* object class.

G.3 Additional results

- Appendix G.3.1 contains numerical experiments, namely linear evaluation & an ablation on $\dim(\hat{\mathbf{c}})$.
- Appendix G.3.2 contains experiments on *Causal3DIdent*, namely (i) nonlinear & linear evaluation results of the output & intermediate feature representation of `SimCLR` with results for the individual axes of object position & rotation, and (ii) evaluation of `BarlowTwins`.
- Appendix G.3.3 contains experiments on the *MPI3D-real* dataset [393], namely SimCLR & a supervised sanity check.

G.3.1 Numerical Data

In Tab. G.4, we report mean ± std. dev. R^2 over 3 random seeds across four generative processes of increasing complexity using *linear* (instead of nonlinear) regression to predict $\mathbf{c}$ from $\hat{\mathbf{c}}$. The block-identification of content can clearly still be seen even if we consider a linear fit.

In Fig. G.8, we perform an ablation on $\dim(\hat{\mathbf{c}})$, visualising how varying the dimensionality of the learnt representation affects identifiability of the ground-truth content & style partition. Generally, if $\dim(\hat{\mathbf{c}}) < n_c$, there is insufficient capacity to encode all content, so a lower-dimensional mixture of content is learnt. Conversely, if $\dim(\hat{\mathbf{c}}) > n_c$, the excess capacity is used to encode some style information, as that increases entropy.

Generative process			R^2 (linear)	
p(chg.)	Stat.	Cau.	Content c	Style s
1.0	✗	✗	**1.00** ± 0.00	0.00 ± 0.00
0.75	✗	✗	**0.99** ± 0.00	0.00 ± 0.00
0.75	✓	✗	**0.97** ± 0.03	0.37 ± 0.05
0.75	✓	✓	**0.98** ± 0.01	**0.78** ± 0.07

Table G.4: Results using linear regression for the experiment on numerical data presented in Subsection 8.5.1

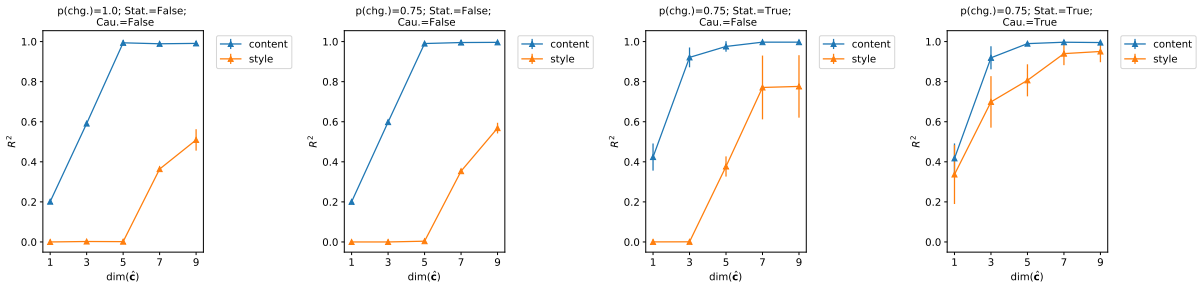


Figure G.8: Identifiability of the content & style partition in the numerical experiment as a function of the model latent dimensionality

On Dependence. As can be seen from Tab. G.4, the corresponding inset table in § 8.5.1, and Fig. G.8, scores for identifying style increase substantially when statistical dependence within blocks and causal dependence between blocks are included. This finding can be explained as follows.

If we compare the performance for small latent dimensionalities ($\dim(\hat{\mathbf{c}}) < n_c$) between the first two (without) and the third plot (with statistical

dependence) of Fig. G.8, we observe a significantly higher score in identifying content for the latter (e.g., R^2 of ca. 0.4 vs 0.2 at $\dim(\hat{\mathbf{c}}) = 1$). This suggest that the introduction of statistical dependence between content variables (as well as between style variables, and in how style variables change) in the third plot/row, reduces the effective dimensionality of the ground-truth latents and thus leads to higher content identifiability for the same $\dim(\hat{\mathbf{c}}) < n_c$. Since the R^2 for content is already close to 1 for $\dim(\hat{\mathbf{c}}) = 3$ in the third plot of Fig. G.8 (due to the smaller effective dimensionality induced by statistical dependence between $\mathbf{c}$), when $\dim(\hat{\mathbf{c}}) = n_c = 5$ is used (as reported in Tab. G.4), excess capacity is used to encode style, leading to a positive R^2.

Regarding causal dependence (i.e., the fourth plot in Fig. G.8 and fourth row in Tab. G.4), we note that the ground truth dependence between $\mathbf{c}$ and $\mathbf{s}$ is linear, i.e., $p(\cdot|\mathbf{c})$ is centred at a linear transformation $\mathbf{a} + B\mathbf{c}$ of $\mathbf{c}$, see the data generating process in Appendix G.4 for details. Given that our evaluation consists of predicting the ground truth $\mathbf{c}$ and $\mathbf{s}$ from the learnt representation $\hat{\mathbf{c}} = \mathbf{g}(\mathbf{x})$, if we were to block-identify $\mathbf{c}$ according to Defn. 8.4.1, we should be able to also predict some aspects of $\mathbf{s}$ from $\hat{\mathbf{c}}$, due to the linear dependence between $\mathbf{c}$ and $\mathbf{s}$. This manifests in a relatively large R^2 for $\mathbf{s}$ in the last row of Tab. G.4 and the corresponding table in § 8.5.1.

To summarise, we highlight two main takeaways: (i) when latent dependence is present, this may reduce the effective dimensionality, so that some style is encoded in addition to content unless a smaller representation size is chosen; (ii) even though the learnt representation isolates content in the sense of Defn. 8.4.1, it may still be predictive of style when content and style are (causally) dependent.

G.3.2 *Causal3DIdent*

Full version of Tab. 8.2: In Tab. G.5, we a) provide the results for the individual axes of object position & rotation and b) present additional rows omitted from Tab. 8.2 for space considerations.

Interestingly, we find that the variance across the individual axes is significantly higher for object position than object rotation. If we compare the causal dependence imposed for object position (see Tab. G.2) to the causal dependence imposed for object rotation (see Tab. G.1), we can observe that the dependence imposed over individual axes is also significantly more variable for position than rotation, i.e., for x the sine nonlinearity is used, for y the cosine nonlinearity is used, while for z, no dependence is imposed.

Regarding the additional rows, we can observe that the composition of image-level rotation & crops yields results quite similar to solely using crops, a relationship which mirrors how transforming the rotation & position latents yields results quite similar to solely transforming the position latents. This suggests that the rotation variables are difficult to disentangle from the position variables in Causal3DIdent, regardless of whether data augmentation or latent transforms are used.

Finally, we can observe that applying image-level rotation in conjunction with small crops & colour distortion does lead to a difference in the encoding, background hue is preserved, while the scores for object position & rotation appear to slightly decrease. When using three augmentations as opposed to two, the effects of the individual augmentations are lessened. While colour distortion discourages the encoding of background hue, both small crops & image-level rotation encourages it, and thus it is preserved when all three augmentations are used. While colour distortion encourages the encoding of object position & rotation, both small crops & image-level rotation discourage it, but as a causal relationship exists between the class variable and said latents, the scores merely decrease, the latents are still for the most part preserved. In reality, where complex interactions between latent variables abound, the effect of data augmentations may be uninterpretable, however with Causal3DIdent, we are able to interpret their effects in the presence of rich visual complexity and causal dependencies, even when applying three distinct augmentations in tandem.

Table G.5: Full version of Tab. 8.2.

Views generated by	Class	Positions				Hues			Rotations		
		object(x)	object(y)	object(z)	spotlight	object	spotlight	background	object(ϕ)	object(θ)	object(ψ)
DA: colour distortion	0.42 ± 0.01	**0.58** ± 0.01	**0.75** ± 0.00	**0.52** ± 0.01	0.17 ± 0.00	0.10 ± 0.01	0.01 ± 0.00	0.01 ± 0.00	0.36 ± 0.01	0.33 ± 0.01	0.32 ± 0.00
LT: change hues	**1.00** ± 0.00	**0.81** ± 0.02	**0.81** ± 0.02	0.15 ± 0.02	**0.91** ± 0.00	0.30 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.30 ± 0.02	0.30 ± 0.01	0.30 ± 0.01
DA: crop (large)	0.28 ± 0.04	0.04 ± 0.02	0.03 ± 0.01	0.19 ± 0.02	0.21 ± 0.13	**0.87** ± 0.00	0.09 ± 0.02	**1.00** ± 0.00	0.00 ± 0.00	0.05 ± 0.00	0.02 ± 0.00
DA: crop (small)	0.14 ± 0.00	0.00 ± 0.00	0.01 ± 0.02	0.00 ± 0.00	0.00 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	**1.00** ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
LT: change positions	**1.00** ± 0.00	0.01 ± 0.00	0.47 ± 0.01	0.01 ± 0.00	0.00 ± 0.01	0.46 ± 0.02	0.00 ± 0.00	**0.97** ± 0.00	0.30 ± 0.00	0.29 ± 0.00	0.28 ± 0.00
DA: crop (large) + colour distortion	**0.97** ± 0.00	**0.59** ± 0.03	**0.52** ± 0.01	**0.68** ± 0.01	**0.59** ± 0.05	0.28 ± 0.00	0.01 ± 0.01	0.01 ± 0.00	**0.74** ± 0.01	**0.78** ± 0.00	**0.72** ± 0.00
DA: crop (small) + colour distortion	**1.00** ± 0.00	**0.72** ± 0.02	**0.65** ± 0.02	**0.70** ± 0.00	**0.93** ± 0.00	0.30 ± 0.01	0.00 ± 0.00	0.02 ± 0.03	**0.53** ± 0.00	**0.57** ± 0.01	**0.58** ± 0.01
LT: change positions + hues	**1.00** ± 0.00	0.10 ± 0.10	0.49 ± 0.02	0.06 ± 0.05	0.07 ± 0.08	0.32 ± 0.02	0.00 ± 0.01	0.02 ± 0.03	0.34 ± 0.09	0.34 ± 0.04	0.34 ± 0.08
DA: rotation	0.33 ± 0.06	0.29 ± 0.03	0.11 ± 0.01	0.12 ± 0.04	0.23 ± 0.12	**0.83** ± 0.01	0.30 ± 0.12	**0.99** ± 0.00	0.02 ± 0.01	0.06 ± 0.03	0.07 ± 0.01
LT: change rotations	**1.00** ± 0.00	**0.78** ± 0.01	**0.72** ± 0.03	0.09 ± 0.03	**0.90** ± 0.00	0.41 ± 0.00	0.00 ± 0.00	**0.97** ± 0.00	0.28 ± 0.00	0.28 ± 0.00	0.28 ± 0.00
DA: rotation + colour distortion	**0.59** ± 0.01	**0.63** ± 0.01	**0.57** ± 0.08	**0.54** ± 0.02	0.21 ± 0.01	0.12 ± 0.02	0.01 ± 0.00	0.01 ± 0.00	0.36 ± 0.03	0.34 ± 0.04	0.30 ± 0.03
LT: change rotations + hues	**1.00** ± 0.00	**0.80** ± 0.02	**0.77** ± 0.01	0.13 ± 0.02	**0.91** ± 0.00	0.30 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.28 ± 0.00	0.28 ± 0.01	0.28 ± 0.00
DA: rot. + crop (lg)	0.26 ± 0.01	0.03 ± 0.02	0.03 ± 0.01	0.15 ± 0.04	0.04 ± 0.03	**0.84** ± 0.06	0.10 ± 0.01	**1.00** ± 0.00	0.00 ± 0.00	0.04 ± 0.02	0.02 ± 0.00
DA: rot. + crop (sm)	0.15 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	**1.00** ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
LT: change rot. + pos.	**1.00** ± 0.00	0.02 ± 0.03	0.48 ± 0.02	0.01 ± 0.01	0.02 ± 0.03	0.49 ± 0.03	0.03 ± 0.02	**0.98** ± 0.00	0.29 ± 0.01	0.28 ± 0.01	0.28 ± 0.01
DA: rot. + crop (lg) + col. dist.	**0.99** ± 0.00	**0.69** ± 0.03	**0.60** ± 0.01	**0.70** ± 0.02	**0.86** ± 0.03	0.28 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	**0.60** ± 0.01	**0.64** ± 0.02	**0.61** ± 0.01
DA: rot. + crop (sm) + col. dist.	**1.00** ± 0.00	**0.61** ± 0.02	**0.59** ± 0.01	**0.64** ± 0.01	**0.82** ± 0.01	0.38 ± 0.00	0.01 ± 0.01	**0.78** ± 0.03	0.44 ± 0.00	0.48 ± 0.02	0.45 ± 0.01
LT: change rot. + pos. + hues	**1.00** ± 0.00	0.20 ± 0.12	**0.50** ± 0.04	0.14 ± 0.11	0.15 ± 0.12	0.32 ± 0.01	0.00 ± 0.00	0.02 ± 0.01	0.33 ± 0.04	0.33 ± 0.02	0.32 ± 0.03

Linear identifiability: In Tab. G.6, we present results evaluating all continuous variables with linear regression. While, as expected, R^2 scores are reduced across the board, we can observe that even with a linear fit, the patterns observed in Tab. G.5 persist.

Table G.6: Evaluation results using a linear fit for not only class, but all continuous variables.

Views generated by	Class	Positions				Hues			Rotations		
		object(x)	object(y)	object(z)	spotlight	object	spotlight	background	object(ϕ)	object(θ)	object(ψ)
DA: colour distortion	0.42 ± 0.01	0.37 ± 0.03	0.20 ± 0.16	0.23 ± 0.02	0.01 ± 0.01	0.03 ± 0.01	−0.00 ± 0.00	−0.00 ± 0.00	0.13 ± 0.01	0.04 ± 0.01	0.09 ± 0.02
LT: change hues	1.00 ± 0.00	0.72 ± 0.07	0.56 ± 0.04	−0.00 ± 0.00	0.65 ± 0.07	0.29 ± 0.01	−0.00 ± 0.00	−0.00 ± 0.00	0.27 ± 0.01	0.26 ± 0.03	0.26 ± 0.01
DA: crop (large)	0.28 ± 0.04	0.00 ± 0.00	0.02 ± 0.00	0.04 ± 0.07	0.08 ± 0.13	0.51 ± 0.05	0.03 ± 0.02	0.20 ± 0.04	0.00 ± 0.00	0.02 ± 0.00	0.01 ± 0.00
DA: crop (small)	0.14 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00	0.17 ± 0.05	−0.00 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00
LT: change positions	1.00 ± 0.00	−0.00 ± 0.00	0.44 ± 0.02	−0.00 ± 0.00	−0.00 ± 0.00	0.29 ± 0.04	0.00 ± 0.00	0.73 ± 0.16	0.26 ± 0.01	0.25 ± 0.03	0.25 ± 0.04
DA: crop (large) + colour distortion	0.97 ± 0.00	0.12 ± 0.02	0.24 ± 0.03	0.21 ± 0.00	0.08 ± 0.03	0.13 ± 0.01	−0.00 ± 0.00	−0.00 ± 0.00	0.14 ± 0.04	0.18 ± 0.05	0.22 ± 0.02
DA: crop (small) + colour distortion	1.00 ± 0.00	0.35 ± 0.02	0.50 ± 0.01	0.19 ± 0.03	0.80 ± 0.01	0.28 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00	0.29 ± 0.00	0.30 ± 0.00	0.29 ± 0.01
LT: change positions + hues	1.00 ± 0.00	0.00 ± 0.00	0.42 ± 0.06	0.00 ± 0.00	0.00 ± 0.00	0.27 ± 0.02	−0.00 ± 0.00	−0.00 ± 0.00	0.23 ± 0.07	0.26 ± 0.03	0.25 ± 0.04
DA: rotation	0.33 ± 0.06	0.04 ± 0.04	0.04 ± 0.00	0.02 ± 0.03	0.12 ± 0.08	0.46 ± 0.06	0.06 ± 0.04	0.30 ± 0.13	0.00 ± 0.00	0.04 ± 0.02	0.02 ± 0.00
LT: change rotations	1.00 ± 0.00	0.34 ± 0.21	0.48 ± 0.03	−0.00 ± 0.00	0.60 ± 0.15	0.28 ± 0.00	0.00 ± 0.00	0.59 ± 0.26	0.27 ± 0.01	0.27 ± 0.00	0.27 ± 0.01
DA: rotation + colour distortion	0.59 ± 0.01	0.31 ± 0.02	0.26 ± 0.06	0.25 ± 0.07	0.02 ± 0.00	0.03 ± 0.02	−0.00 ± 0.00	−0.00 ± 0.00	0.07 ± 0.01	0.06 ± 0.01	0.10 ± 0.01
LT: change rotations + hues	1.00 ± 0.00	0.68 ± 0.02	0.57 ± 0.01	−0.00 ± 0.00	0.72 ± 0.10	0.29 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00	0.28 ± 0.00	0.28 ± 0.00	0.28 ± 0.00
DA: rot. + crop (lg)	0.26 ± 0.01	−0.00 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.59 ± 0.05	0.02 ± 0.01	0.20 ± 0.04	0.00 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
DA: rot. + crop (sm)	0.15 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00	0.29 ± 0.21	−0.00 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00
LT: change rot. + pos.	1.00 ± 0.00	−0.00 ± 0.00	0.45 ± 0.01	−0.00 ± 0.00	−0.00 ± 0.00	0.32 ± 0.02	0.00 ± 0.00	0.80 ± 0.09	0.27 ± 0.00	0.27 ± 0.01	0.27 ± 0.01
DA: rot. + crop (lg) + col. dist.	0.99 ± 0.00	0.23 ± 0.04	0.26 ± 0.07	0.26 ± 0.01	0.51 ± 0.14	0.21 ± 0.01	−0.00 ± 0.00	−0.00 ± 0.00	0.21 ± 0.04	0.28 ± 0.02	0.22 ± 0.02
DA: rot. + crop (sm) + col. dist.	1.00 ± 0.00	0.26 ± 0.02	0.48 ± 0.01	0.21 ± 0.02	0.61 ± 0.01	0.31 ± 0.00	−0.00 ± 0.00	0.34 ± 0.02	0.30 ± 0.00	0.30 ± 0.01	0.29 ± 0.01
LT: change rot. + pos. + hues	1.00 ± 0.00	0.03 ± 0.05	0.46 ± 0.01	0.01 ± 0.01	0.01 ± 0.02	0.29 ± 0.01	−0.00 ± 0.00	−0.00 ± 0.00	0.27 ± 0.00	0.28 ± 0.01	0.28 ± 0.01

Intermediate feature evaluation: In Tab. G.7 and Tab. G.8, we present evaluation based on the representation from an intermediate layer (i.e., prior to applying a projection layer [329]) with nonlinear and linear regression for the continuous variables, respectively. Note the intermediate layer has an output dimensionality of 100. While it is clear that all R^2 scores are increased across the board, we can notice that certain latents which were discarded in the final layer, were not in an intermediate layer. For example, with "LT: change hues", in the final layer the z-position was discarded ($R^2 = 0.15$ in Tab. G.5), inexplicably we may add, as position is content regardless of axis with this latent transformation. But in the intermediate layer, z-position was not discarded ($R^2 = 0.88$ in Tab. G.7).

Table G.7: Evaluation of an intermediate layer. Logistic regression used for class, kernel ridge regression used for all continuous variables.

Views generated by	Class	Positions				Hues			Rotations		
		object(x)	object(y)	object(z)	spotlight	object	spotlight	background	object(ϕ)	object(θ)	object(ψ)
DA: colour distortion	0.71 ± 0.02	0.68 ± 0.02	0.80 ± 0.01	0.63 ± 0.01	0.25 ± 0.01	0.13 ± 0.00	0.02 ± 0.01	0.01 ± 0.01	0.44 ± 0.01	0.48 ± 0.01	0.39 ± 0.00
LT: change hues	1.00 ± 0.00	0.98 ± 0.00	0.97 ± 0.00	0.88 ± 0.01	0.98 ± 0.00	0.34 ± 0.01	−0.00 ± 0.00	0.20 ± 0.10	0.71 ± 0.02	0.68 ± 0.03	0.68 ± 0.02
DA: crop (large)	0.43 ± 0.03	0.41 ± 0.05	0.35 ± 0.05	0.32 ± 0.04	0.41 ± 0.13	0.88 ± 0.00	0.14 ± 0.03	1.00 ± 0.00	0.03 ± 0.02	0.06 ± 0.01	0.08 ± 0.00
DA: crop (small)	0.20 ± 0.01	0.04 ± 0.05	0.20 ± 0.02	0.01 ± 0.02	0.20 ± 0.03	−0.00 ± 0.00	−0.00 ± 0.00	1.00 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00
LT: change positions	1.00 ± 0.00	0.78 ± 0.02	0.90 ± 0.01	0.75 ± 0.01	0.59 ± 0.02	0.82 ± 0.01	0.18 ± 0.02	0.99 ± 0.00	0.64 ± 0.02	0.55 ± 0.02	0.56 ± 0.02
DA: crop (large) + colour distortion	1.00 ± 0.00	0.92 ± 0.00	0.83 ± 0.00	0.92 ± 0.00	0.90 ± 0.01	0.29 ± 0.00	0.01 ± 0.01	0.01 ± 0.01	0.87 ± 0.00	0.90 ± 0.00	0.85 ± 0.00
DA: crop (small) + colour distortion	1.00 ± 0.00	0.92 ± 0.00	0.87 ± 0.01	0.90 ± 0.00	0.97 ± 0.00	0.46 ± 0.04	0.02 ± 0.02	0.58 ± 0.12	0.79 ± 0.01	0.83 ± 0.00	0.79 ± 0.00
LT: change positions + hues	1.00 ± 0.00	0.83 ± 0.04	0.90 ± 0.01	0.81 ± 0.04	0.75 ± 0.08	0.42 ± 0.09	0.04 ± 0.02	0.52 ± 0.20	0.72 ± 0.05	0.69 ± 0.07	0.67 ± 0.06
DA: rotation	0.46 ± 0.04	0.35 ± 0.04	0.19 ± 0.02	0.28 ± 0.04	0.34 ± 0.08	0.85 ± 0.01	0.35 ± 0.12	1.00 ± 0.00	0.03 ± 0.01	0.08 ± 0.02	0.10 ± 0.01
LT: change rotations	1.00 ± 0.00	0.97 ± 0.00	0.96 ± 0.01	0.84 ± 0.01	0.98 ± 0.00	0.82 ± 0.01	0.17 ± 0.02	0.99 ± 0.00	0.64 ± 0.02	0.59 ± 0.01	0.60 ± 0.03
DA: rotation + colour distortion	0.87 ± 0.02	0.76 ± 0.01	0.81 ± 0.01	0.71 ± 0.01	0.39 ± 0.08	0.19 ± 0.02	−0.00 ± 0.00	0.02 ± 0.02	0.55 ± 0.03	0.55 ± 0.03	0.48 ± 0.02
LT: change rotations + hues	1.00 ± 0.00	0.98 ± 0.00	0.97 ± 0.00	0.87 ± 0.00	0.99 ± 0.00	0.39 ± 0.05	0.04 ± 0.02	0.37 ± 0.21	0.69 ± 0.01	0.68 ± 0.01	0.68 ± 0.00
DA: rot. + crop (lg)	0.43 ± 0.03	0.38 ± 0.04	0.34 ± 0.02	0.28 ± 0.03	0.30 ± 0.05	0.86 ± 0.04	0.17 ± 0.02	1.00 ± 0.00	0.02 ± 0.00	0.05 ± 0.01	0.10 ± 0.01
DA: rot. + crop (sm)	0.20 ± 0.01	0.07 ± 0.03	0.09 ± 0.10	0.01 ± 0.01	0.20 ± 0.01	−0.00 ± 0.00	−0.00 ± 0.00	1.00 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00	−0.00 ± 0.00
LT: change rot. + pos.	1.00 ± 0.00	0.81 ± 0.01	0.90 ± 0.01	0.76 ± 0.01	0.67 ± 0.04	0.84 ± 0.01	0.28 ± 0.04	0.99 ± 0.00	0.62 ± 0.02	0.57 ± 0.01	0.55 ± 0.01
DA: rot. + crop (lg) + col. dist.	1.00 ± 0.00	0.92 ± 0.01	0.89 ± 0.00	0.92 ± 0.00	0.95 ± 0.01	0.30 ± 0.00	0.02 ± 0.02	0.18 ± 0.16	0.81 ± 0.00	0.84 ± 0.00	0.79 ± 0.00
DA: rot. + crop (sm) + col. dist.	1.00 ± 0.00	0.87 ± 0.00	0.85 ± 0.00	0.87 ± 0.00	0.93 ± 0.00	0.71 ± 0.02	0.33 ± 0.05	0.96 ± 0.00	0.72 ± 0.00	0.75 ± 0.00	0.71 ± 0.00
LT: change rot. + pos. + hues	1.00 ± 0.00	0.84 ± 0.02	0.91 ± 0.01	0.82 ± 0.02	0.78 ± 0.06	0.40 ± 0.01	0.06 ± 0.01	0.50 ± 0.05	0.72 ± 0.04	0.70 ± 0.05	0.67 ± 0.04

Table G.8: Evaluation of an intermediate layer. Logistic regression used for class, linear regression used for all continuous variables.

Views generated by	Class	Positions				Hues			Rotations		
		object(x)	object(y)	object(z)	spotlight	object	spotlight	background	object(ϕ)	object(θ)	object(ψ)
DA: colour distortion	0.71 ± 0.02	0.53 ± 0.01	0.70 ± 0.01	0.46 ± 0.01	0.13 ± 0.01	0.11 ± 0.01	−0.01 ± 0.00	0.00 ± 0.00	0.28 ± 0.01	0.19 ± 0.01	0.25 ± 0.01
LT: change hues	1.00 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.60 ± 0.04	0.95 ± 0.00	0.31 ± 0.00	0.01 ± 0.01	0.06 ± 0.04	0.44 ± 0.02	0.41 ± 0.02	0.42 ± 0.00
DA: crop (large)	0.43 ± 0.03	0.18 ± 0.06	0.06 ± 0.01	0.17 ± 0.02	0.19 ± 0.14	0.82 ± 0.02	0.08 ± 0.04	0.98 ± 0.00	0.01 ± 0.00	0.05 ± 0.01	0.05 ± 0.01
DA: crop (small)	0.20 ± 0.01	0.01 ± 0.01	0.03 ± 0.02	0.00 ± 0.01	0.02 ± 0.01	−0.00 ± 0.00	−0.01 ± 0.00	0.99 ± 0.00	−0.01 ± 0.00	−0.01 ± 0.00	−0.00 ± 0.01
LT: change positions	1.00 ± 0.00	0.49 ± 0.04	0.72 ± 0.03	0.43 ± 0.03	0.19 ± 0.03	0.71 ± 0.02	0.09 ± 0.02	0.98 ± 0.00	0.39 ± 0.01	0.36 ± 0.01	0.35 ± 0.00
DA: crop (large) + colour distortion	1.00 ± 0.00	0.67 ± 0.03	0.56 ± 0.01	0.66 ± 0.02	0.67 ± 0.03	0.28 ± 0.00	−0.01 ± 0.00	0.01 ± 0.01	0.58 ± 0.02	0.61 ± 0.02	0.56 ± 0.01
DA: crop (small) + colour distortion	1.00 ± 0.00	0.76 ± 0.01	0.70 ± 0.02	0.68 ± 0.01	0.90 ± 0.00	0.38 ± 0.03	0.00 ± 0.01	0.39 ± 0.13	0.50 ± 0.02	0.50 ± 0.01	0.49 ± 0.01
LT: change positions + hues	1.00 ± 0.00	0.61 ± 0.09	0.74 ± 0.02	0.51 ± 0.08	0.40 ± 0.15	0.34 ± 0.04	0.02 ± 0.01	0.25 ± 0.22	0.47 ± 0.04	0.40 ± 0.02	0.41 ± 0.03
DA: rotation	0.46 ± 0.04	0.21 ± 0.02	0.10 ± 0.01	0.10 ± 0.02	0.21 ± 0.09	0.77 ± 0.01	0.25 ± 0.11	0.97 ± 0.01	0.02 ± 0.01	0.06 ± 0.02	0.08 ± 0.01
LT: change rotations	1.00 ± 0.00	0.92 ± 0.00	0.88 ± 0.01	0.51 ± 0.02	0.95 ± 0.00	0.70 ± 0.06	0.07 ± 0.02	0.98 ± 0.00	0.36 ± 0.01	0.34 ± 0.00	0.34 ± 0.01
DA: rotation + colour distortion	0.87 ± 0.02	0.60 ± 0.01	0.62 ± 0.03	0.52 ± 0.02	0.23 ± 0.02	0.18 ± 0.02	−0.01 ± 0.00	0.02 ± 0.01	0.33 ± 0.04	0.29 ± 0.01	0.28 ± 0.01
LT: change rotations + hues	1.00 ± 0.00	0.94 ± 0.00	0.92 ± 0.01	0.58 ± 0.01	0.96 ± 0.00	0.33 ± 0.02	0.02 ± 0.01	0.15 ± 0.10	0.40 ± 0.02	0.38 ± 0.01	0.41 ± 0.03
DA: rot. + crop (lg)	0.43 ± 0.03	0.24 ± 0.04	0.08 ± 0.02	0.16 ± 0.03	0.07 ± 0.01	0.80 ± 0.04	0.10 ± 0.01	0.98 ± 0.00	0.01 ± 0.00	0.05 ± 0.01	0.06 ± 0.01
DA: rot. + crop (sm)	0.20 ± 0.01	0.01 ± 0.01	0.03 ± 0.01	−0.00 ± 0.01	0.04 ± 0.01	−0.01 ± 0.00	−0.01 ± 0.00	0.99 ± 0.00	−0.01 ± 0.00	−0.01 ± 0.00	−0.00 ± 0.01
LT: change rot. + pos.	1.00 ± 0.00	0.55 ± 0.05	0.72 ± 0.02	0.44 ± 0.04	0.31 ± 0.08	0.76 ± 0.01	0.14 ± 0.01	0.99 ± 0.00	0.38 ± 0.01	0.35 ± 0.01	0.36 ± 0.02
DA: rot. + crop (lg) + col. dist.	1.00 ± 0.00	0.71 ± 0.01	0.69 ± 0.01	0.69 ± 0.00	0.84 ± 0.03	0.28 ± 0.00	−0.00 ± 0.00	0.07 ± 0.07	0.51 ± 0.01	0.50 ± 0.02	0.51 ± 0.01
DA: rot. + crop (sm) + col. dist.	1.00 ± 0.00	0.66 ± 0.00	0.69 ± 0.01	0.65 ± 0.02	0.83 ± 0.00	0.57 ± 0.03	0.18 ± 0.02	0.89 ± 0.01	0.46 ± 0.01	0.45 ± 0.02	0.44 ± 0.01
LT: change rot. + pos. + hues	1.00 ± 0.00	0.65 ± 0.04	0.75 ± 0.05	0.57 ± 0.03	0.49 ± 0.12	0.35 ± 0.01	0.02 ± 0.01	0.23 ± 0.04	0.48 ± 0.04	0.43 ± 0.01	0.43 ± 0.01

In [329], the value in evaluating an intermediate layer as opposed to a final layer is discussed, where the authors demonstrated that predicting the data augmentations applied during training is significantly more accurate from an intermediate layer as opposed to the final layer, implying that the intermediate layer contains much more information about the transformation applied. Our results suggest a distinct hypothesis, the value in using an intermediate layer as a representation for downstream tasks is not due to preservation of style information, as can be seen, R^2 scores on style variables are not significantly higher in Tab. G.7 relative to Tab. G.5. The value is in preservation of all content variables, as we can observe certain content variables are discarded in the final layer,

Table G.9: *BarlowTwins* $\lambda = 0.0051$ results: R^2 mean ± std. dev. over 3 random seeds. DA: data augmentation, LT: latent transformation, bold: $R^2 \geq 0.5$, red: $R^2 < 0.25$. Results for individual axes of object position & rotation are aggregated.

Views generated by	Class	Positions		Hues			Rotations
		object	spotlight	object	spotlight	background	
DA: colour distortion	0.48 ± 0.02	**0.51** ± 0.14	0.07 ± 0.01	0.08 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.21 ± 0.04
LT: change hues	**1.00** ± 0.00	**0.56** ± 0.20	**0.76** ± 0.07	0.30 ± 0.01	0.00 ± 0.00	0.01 ± 0.00	0.35 ± 0.01
DA: crop (large)	0.17 ± 0.02	0.10 ± 0.03	0.06 ± 0.02	0.29 ± 0.13	0.11 ± 0.05	**0.99** ± 0.00	0.02 ± 0.01
DA: crop (small)	0.15 ± 0.00	0.04 ± 0.02	0.05 ± 0.02	0.02 ± 0.01	0.00 ± 0.01	**1.00** ± 0.00	0.00 ± 0.01
LT: change positions	**0.88** ± 0.00	0.19 ± 0.20	0.05 ± 0.00	**0.50** ± 0.02	0.04 ± 0.01	**0.98** ± 0.00	0.27 ± 0.03
DA: crop (large) + colour distortion	**0.87** ± 0.02	0.49 ± 0.06	0.32 ± 0.03	0.25 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	**0.50** ± 0.02
DA: crop (small) + colour distortion	**0.81** ± 0.01	0.39 ± 0.07	0.42 ± 0.06	0.47 ± 0.04	0.03 ± 0.01	**0.85** ± 0.02	0.30 ± 0.02
LT: change positions + hues	**1.00** ± 0.00	0.28 ± 0.20	0.12 ± 0.05	0.31 ± 0.00	0.00 ± 0.00	0.01 ± 0.01	0.37 ± 0.06

but are preserved in an intermediate layer. With that being said, our theoretical result applies to the final layer, which is why said results were highlighted in the main paper. The discarding of certain content variables is an empirical phenomenon, likely a consequence of a limited number of negative samples in practice, leading to certain content variables being redundant, or unnecessary, for solving the contrastive objective.

The fact that we can recover certain content variables which appeared discarded in the output from the intermediate layer may suggest that we should be able to decode class. While scores are certainly increased, we do not see such drastic differences in R^2 scores, as was seen above. The drastic difference highlighted above was with regards to latent transformation, for which we always observed class encoded as a content variable. So, unfortunately, using an intermediate layer does not rectify the discrepancy between data augmentations and latent transformations. While latent transformations allow us to better interpret the effect of certain empirical techniques [329], as discussed in the main paper, we cannot make a one-to-one correspondence between data augmentations used in practice and latent transformations.

BarlowTwins: We repeat our analysis from § 8.5.2 using `BarlowTwins` [365] (instead of `SimCLR`) which, as discussed at the end of § 8.4.2, is also loosely related to Thm. 8.4.3. The `BarlowTwins` objective consists of an invariance term, which equates the diagonal elements of the cross-correlation matrix to 1, thereby making the embedding invariant to the distortions applied and a redundancy reduction term, which equates the off-diagonal elements of the cross-correlation matrix to 0, thereby decorrelating the different vector components of the embedding, reducing the redundancy between output units.

In Tab. G.9 we train `BarlowTwins` with $\lambda = 0.0051$, the default value for the hyperparameter which weights the redundancy reduction term relative to the invariance term. To confirm the insights are robust to the value of λ,in Tab. G.10, we report results with λ increased by an order of magnitude, $\lambda = 0.051$. We find that the results mirror Tab. 8.2, e.g. colour distortion yields a discarding of hue, crops isolate background hue where the larger the crop, the higher the identifiability of object hue, and crops & colour distortion yield high accuracy in inferring the object class variable.

Table G.10: *BarlowTwins* $\lambda = 0.051$ results: R^2 mean ± std. dev. over 3 random seeds. DA: data augmentation, LT: latent transformation, bold: $R^2 \geq 0.5$, red: $R^2 < 0.25$. Results for individual axes of object position & rotation are aggregated.

Views generated by	Class	Positions		Hues			Rotations
		object	spotlight	object	spotlight	background	
DA: colour distortion	**0.52** ± 0.07	0.43 ± 0.18	0.07 ± 0.02	0.10 ± 0.03	0.00 ± 0.00	0.00 ± 0.00	0.21 ± 0.05
LT: change hues	**1.00** ± 0.00	**0.55** ± 0.24	**0.74** ± 0.02	0.30 ± 0.00	0.00 ± 0.00	0.01 ± 0.01	0.33 ± 0.02
DA: crop (large)	0.19 ± 0.05	0.08 ± 0.02	0.05 ± 0.01	0.39 ± 0.36	0.08 ± 0.05	**0.96** ± 0.05	0.01 ± 0.02
DA: crop (small)	0.15 ± 0.00	0.05 ± 0.02	0.07 ± 0.02	0.00 ± 0.01	0.01 ± 0.01	**1.00** ± 0.00	0.00 ± 0.00
LT: change positions	**0.89** ± 0.01	0.19 ± 0.20	0.05 ± 0.01	0.48 ± 0.04	0.05 ± 0.02	**0.98** ± 0.00	0.25 ± 0.03
DA: crop (large) + colour distortion	**0.86** ± 0.03	0.40 ± 0.07	0.23 ± 0.02	0.24 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.47 ± 0.04
DA: crop (small) + colour distortion	**0.99** ± 0.01	**0.63** ± 0.03	**0.88** ± 0.01	0.32 ± 0.02	0.00 ± 0.00	0.16 ± 0.13	**0.52** ± 0.03
LT: change positions + hues	**1.00** ± 0.00	0.21 ± 0.22	0.07 ± 0.01	0.30 ± 0.00	0.00 ± 0.00	0.02 ± 0.01	0.46 ± 0.06

Table G.11: *MPI3D-real* results: R^2 mean ± std. dev. over 3 random seeds for dim($\hat{\mathbf{c}}$)= 5. DA: data augmentation, bold: $R^2 \geq 0.5$, red: $R^2 < 0.25$.

Views generated by	object color	object shape	object size	camera height	background color	horizontal axis	vertical axis
DA: colour distortion	0.39 ± 0.01	0.00 ± 0.00	0.16 ± 0.01	**1.00** ± 0.00	0.09 ± 0.15	**0.60** ± 0.06	0.42 ± 0.08
DA: crop (large)	**0.65** ± 0.17	0.01 ± 0.02	0.31 ± 0.03	**1.00** ± 0.00	**1.00** ± 0.00	0.37 ± 0.06	0.08 ± 0.03
DA: crop (small)	0.09 ± 0.02	0.03 ± 0.00	0.19 ± 0.01	**1.00** ± 0.00	**1.00** ± 0.00	0.21 ± 0.02	0.07 ± 0.00
DA: crop (large) + colour distortion	0.34 ± 0.00	0.00 ± 0.00	0.22 ± 0.03	**1.00** ± 0.00	0.39 ± 0.02	**0.54** ± 0.01	0.29 ± 0.01
DA: crop (small) + colour distortion	0.25 ± 0.02	0.00 ± 0.00	0.10 ± 0.01	**1.00** ± 0.00	**0.75** ± 0.16	**0.54** ± 0.01	0.29 ± 0.03

G.3.3 *MPI3D-real*

We ran the same experimental setup as in § 8.5.2 also on the *MPI3D-real* dataset [393] containing > 1 million *real* images with ground-truth annotations of 3D objects being moved by a robotic arm.

As *MPI3D-real* contains much lower resolution images (64×64) compared to ImageNet & Causal3DIdent (224×224), we used the standard convolutional encoder from the disentanglement literature [72], and ran a sanity check experiment to verify that by training the same backbone as in our unsupervised experiment with supervised learning, we can recover the ground-truth factors from the augmented views. In Tab. G.12, we observe that only five out of seven factors can be consistently inferred, object shape and size are somewhat ambiguous even when observing the original image. Note that while in the self-supervised case, we evaluate by training a nonlinear regression for each ground truth factor separately, in the supervised case, we train a network for all ground truth factors simultaneously from scratch for as many gradient steps as used for learning the self-supervised model.

In Tab. G.11, we report the evaluation results in the self-supervised scenario. Subject to the aforementioned caveats, the results show a similar trend as those on *Causal3DIdent*, i.e. with colour distortion, color factors of variation are decoded significantly worse than positional/rotational information.

Table G.12: Supervised *MPI3D-real* results: R^2 mean ± std. dev. over 3 random seeds. DA: data augmentation. bold: $R^2 \geq 0.5$, red: $R^2 < 0.25$.

Views generated by	object color	object shape	object size	camera height	background color	horizontal axis	vertical axis
Original	**0.90** ± 0.01	0.25 ± 0.02	**0.61** ± 0.02	**0.99** ± 0.00	**0.97** ± 0.01	**1.00** ± 0.00	**1.00** ± 0.00
DA: colour distortion	**0.61** ± 0.01	0.11 ± 0.00	0.47 ± 0.01	**0.98** ± 0.00	**0.93** ± 0.00	**0.99** ± 0.00	**1.00** ± 0.00
DA: crop (large)	**0.82** ± 0.01	0.05 ± 0.01	0.42 ± 0.02	**0.97** ± 0.01	**0.91** ± 0.00	**0.96** ± 0.00	**0.97** ± 0.01
DA: crop (small)	**0.71** ± 0.04	0.01 ± 0.00	0.32 ± 0.02	**0.95** ± 0.00	**0.85** ± 0.01	**0.79** ± 0.02	**0.90** ± 0.01
DA: crop (large) + colour distortion	0.45 ± 0.02	0.02 ± 0.00	0.22 ± 0.00	**0.95** ± 0.01	**0.67** ± 0.01	**0.91** ± 0.00	**0.94** ± 0.00
DA: crop (small) + colour distortion	0.45 ± 0.02	0.00 ± 0.00	0.17 ± 0.02	**0.91** ± 0.02	**0.55** ± 0.03	**0.69** ± 0.01	**0.79** ± 0.08

G.4 Experimental details

Ground-truth generative model. The generative process used in our numerical simulations (§ 8.5.1) is summarised by the following:

$$\begin{aligned}
\mathbf{c} \sim p(\mathbf{c}) = \mathcal{N}(0, \Sigma_{\mathbf{c}}), \quad \text{with} \quad \Sigma_{\mathbf{c}} \sim \text{Wishart}_{n_c}(\mathbf{I}, n_c), \\
\mathbf{s}|\mathbf{c} \sim p(\mathbf{s}|\mathbf{c}) = \mathcal{N}(\mathbf{a} + B\mathbf{c}, \Sigma_{\mathbf{s}}), \quad \text{with} \quad \Sigma_{\mathbf{s}} \sim \text{Wishart}_{n_s}(\mathbf{I}, n_s), \quad a_i, b_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), \\
\tilde{\mathbf{s}}_A|\mathbf{s}_A, A \sim p(\tilde{\mathbf{s}}_A|\mathbf{s}_A) = N(\mathbf{s}_A, \Sigma(A)) \quad \text{with} \quad \Sigma \sim \text{Wishart}_{n_s}(\mathbf{I}, n_s), \\
(\tilde{\mathbf{x}}, \mathbf{x}) = (\mathbf{f}_{\text{MLP}}(\tilde{\mathbf{z}}), \mathbf{f}_{\text{MLP}}(\mathbf{z})),
\end{aligned}$$

where the set of changing style vectors A is obtained by flipping a (biased) coin with p(chg.) = 0.75 for each style dimension independently, and where $\Sigma(A)$ denotes the submatrix of Σ defined by selecting the rows and columns corresponding to subset A.

When we do not allow for *statistical dependence* (Stat.) within blocks of content and style variables, we set the covariance matrices $\Sigma_{\mathbf{c}}$, $\Sigma_{\mathbf{s}}$, and Σ to the identity. When we do not allow for *causal dependence* (Cau.) of style on content, we set $a_i, b_{ij} = 0, \forall i, j$.

For $\mathbf{f}_{\text{MLP}}$, we use a 3-layer MLP with LeakyReLU ($\alpha = 0.2$) activation functions, specified using the same process as used in previous work [64, 65, 88]. For the square weight matrices, we draw $(n_c + n_s) \times (n_c + n_s)$ samples from $U(-1, 1)$, and perform l_2 column normalisation. In addition, to control for invertibility, we re-sample the weight matrices until their condition number is less than or equal to a threshold value. The threshold is pre-computed by sampling 24,975 weight matrices, and recording the minimum condition number.

Training encoder. Recall that the result of Thm. 8.4.3 corresponds to minimizing the following functional (8.5):

$$\mathscr{L}_{\text{AlignMaxEnt}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathbf{x},\tilde{\mathbf{x}}}}\left[\left(\mathbf{g}(\mathbf{x}) - \mathbf{g}(\tilde{\mathbf{x}})\right)^2\right] - H\left(\mathbf{g}(\mathbf{x})\right).$$

Note that InfoNCE [325, 329] (8.1) can be rewritten as:

$$\mathscr{L}_{\text{InfoNCE}}(\mathbf{g}; \tau, K) = \mathbb{E}_{\{\mathbf{x}_i,\tilde{\mathbf{x}}_i\}_{i=1}^K \sim p_{\mathbf{x},\tilde{\mathbf{x}}}}\left[-\sum_{i=1}^K \text{sim}(\mathbf{g}(\mathbf{x})_i, \mathbf{g}(\tilde{\mathbf{x}})_i)/\tau + \log \sum_{j=1}^K \exp\{\text{sim}(\mathbf{g}(\mathbf{x})_i, \mathbf{g}(\tilde{\mathbf{x}})_j)/\tau\}\right]. \tag{G.27}$$

Thus, if we consider $\tau = 1$, and $\text{sim}(u, v) = -(u - v)^2$,

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{g}; K) = \mathbb{E}_{\{\mathbf{x}_i, \tilde{\mathbf{x}}_i\}_{i=1}^K \sim p_{\mathbf{x},\tilde{\mathbf{x}}}} \left[\sum_{i=1}^K \left(\mathbf{g}(\mathbf{x})_i - \mathbf{g}(\tilde{\mathbf{x}})_i\right)^2 + \log \sum_{j=1}^K \exp\{-(\mathbf{g}(\mathbf{x})_i - \mathbf{g}(\tilde{\mathbf{x}})_j)^2\} \right] \tag{G.28}$$

we can approximately match the form of (8.5). In practice, we use $K = 6,144$.

For $\mathbf{g}$, as in [88], we use a 7-layer MLP with (default) LeakyReLU ($\alpha = 0.01$) activation functions. As the input dimensionality is $(n_c + n_s)$, we consider the following multipliers $[10, 50, 50, 50, 50, 10]$ for the number of hidden units per layer. In correspondence with Thm. 8.4.3, we set the output dimensionality to n_c.

We train our feature encoder for $300,000$ iterations, using Adam [528] with a learning rate of 10^{-4}.

Causal3DIdent. We here elaborate on details specific to the experiments in Subsection 8.5.2. We train the feature encoder for $200,000$ iterations using Adam with a learning rate of 10^{-4}. For the encoder we use a ResNet18 [392] architecture followed by a single hidden layer with dimensionality 100 and LeakyReLU activation function using the default (0.01) negative slope. The scores are evaluated on a test set consisting of $25,000$ samples not included in the training set.

Data augmentations. We here specify the parameters for the data augmentations we considered:

- colour distortion: see the paragraph labelled "Color distortion" in Appendix A of [329] for details. We use $s = 1.0$, the default value.
- crop: see the paragraph labelled "Random crop and resize to 224×224" in Appendix A of [329] for details. For small crops, a crop of random size (uniform from 0.08 to 1.0 in area) of the original size is made, which corresponds to what was used in the experiments reported in [329]. For large crops, a crop of random size (uniform from 0.8 to 1.0 in area) of the original size is made.
- rotation: as specified in the captions for Figure 4 & Table 3 in [329], we sample one of $\{0°, 90°, 180°, 270°\}$ uniformly. Note that for the pair, we sample two values without replacement.

A visual overview of the effect of these image-level data augmentations is shown in Fig. G.9.

Latent transformations. To generate views via latent transformations (LT) in our experiments on Causal3DIdent (§ 8.5.2), we proceed as follows.

Let $\mathbf{z}$ refer to the latent corresponding to the original image. For all latents specified to change, we sample $\hat{\mathbf{z}}'$ from a truncated normal distribution constrained to $[-1, 1]$, centered at $\mathbf{z}$, with $\sigma = 1.$. Then, we use nearest-neighbor matching to find the latent $\hat{\mathbf{z}}$ closest to $\hat{\mathbf{z}}'$ (in L^2 distance) for which there exists an image rendering.[1]

1: see [88] for further details

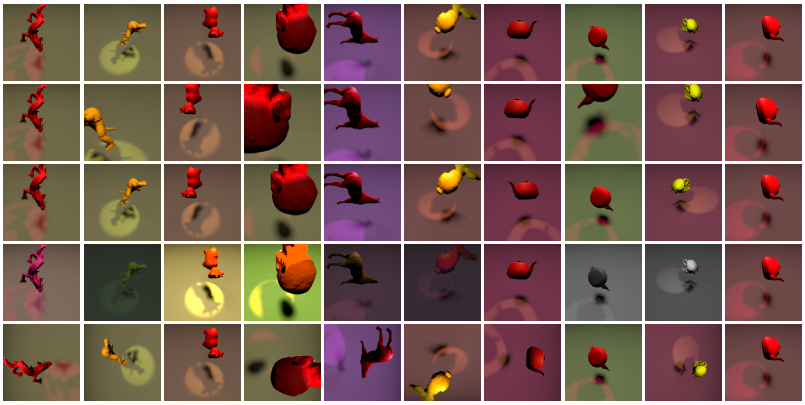


Figure G.9: Visual overview of the effect of different data augmentations (DA), applied to 10 representative samples. Rows correspond to *(top to bottom)*: original images, small random crop (+ random flip), large random crop (+ random flip), colour distortion (jitter & drop), and random rotation.

Evaluation. Recall that Thm. 8.4.3 states that $\mathbf{g}$ block-identifies the true content variables in the sense of Defn. 8.4.1, i.e., there exists an *invertible* function $\mathbf{h} : \mathbb{R}^{n_c} \to \mathbb{R}^{n_c}$ s.t. $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{c})$.

Since this is different from typical evaluation in disentanglement or ICA in that we do not assume independence and do not aim to find a one-to-one correspondence between inferred and ground truth latents, existing metrics, such as MCC [64, 65] or MIG [197], do not apply.

We therefore treat identifying $\mathbf{h}$ as a regression task, which we solve using kernel ridge regression with a Gaussian kernel [67]. Since the Gaussian kernel is universal, this constitutes a nonparametric regression technique with universal approximation capabilities, i.e., any nonlinear function can be approximated arbitrarily well given sufficient data.

We sample 4096×10 datapoints from the marginal for evaluation. For kernel ridge regression, we standardize the inputs and targets, and fit the regression model on 4096×5 (distinct) datapoints. We tune the regularization strength α and kernel variance γ by 3-fold cross-validated grid search over the following parameter grids: $\alpha \in [1, 0.1, 0.001, 0.0001]$, $\gamma \in [0.01, 0.22, 4.64, 100]$.

Compute. The experiments in Subsection 8.5.1 took on the order of 5-10 hours on a single GeForce RTX 2080 Ti GPU. The experiments in Subsection 8.5.2 on 3DIdent took 28 hours on four GeForce RTX 2080 Ti GPUs. The creation of the Causal3DIdent dataset additionally required approximately 150 hours of compute time on a GeForce RTX 2080 Ti.

Bibliography

Here are the references in citation order.

[1] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, Ltd, 2001 (cited on pages 2, 7–12, 125).

[2] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010 (cited on page 3).

[3] Cian Eastwood et al. 'On the DCI Framework for Evaluating Disentangled Representations: Extensions and Connections to Identifiability'. In: *UAI 2022 Workshop on Causal Representation Learning*. 2022 (cited on page 4).

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 'Representation learning: A review and new perspectives'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013), pp. 1798–1828 (cited on pages 4, 5, 40).

[5] Geoffrey Roeder, Luke Metz, and Durk Kingma. 'On linear identifiability of learned representations'. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 9030–9039 (cited on pages 4, 6, 19, 20, 22).

[6] Gaetano Kanizsa. *Vedere e pensare*. Società editrice il Mulino, 1991 (cited on page 4).

[7] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013 (cited on page 4).

[8] Ferenc Huszár. *Goals and Principles of Representation Learning*. Apr. 2018. URL: https://www.inference.vc/about/ (cited on page 4).

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 'Deep learning'. In: *nature* 521.7553 (2015), pp. 436–444 (cited on pages 4, 5).

[10] Geoffrey E Hinton. 'Learning multiple layers of representation'. In: *Trends in cognitive sciences* 11.10 (2007), pp. 428–434 (cited on pages 4, 116).

[11] Taco S Cohen and Max Welling. 'Transformation properties of learned visual representations'. In: *arXiv preprint arXiv:1412.7659* (2014) (cited on pages 5, 21).

[12] Irina Higgins et al. 'Towards a definition of disentangled representations'. In: *arXiv preprint arXiv:1812.02230* (2018) (cited on page 5).

[13] Bernhard Schölkopf et al. 'Toward causal representation learning'. In: *Proceedings of the IEEE* (2021) (cited on pages 5, 21, 29, 98, 111, 119).

[14] Aapo Hyvärinen, Jarmo Hurri, and Patrick O Hoyer. *Natural image statistics: A probabilistic approach to early computational vision.* Vol. 39. Springer Science & Business Media, 2009 (cited on page 5).

[15] Wei Ji Ma, Konrad Paul Kording, and Daniel Goldreich. *Bayesian Models of Perception and Action*. Online draft, Accessed 10th of September 2022. MIT Press, 2022 (cited on page 5).

[16] Aapo Hyvärinen. 'Painful intelligence: What AI can tell us about human suffering'. In: *arXiv preprint arXiv:2205.15409* (2022) (cited on page 5).

[17] Suzanna Becker and Geoffrey E Hinton. 'Self-organizing neural network that discovers surfaces in random-dot stereograms'. In: *Nature* 355.6356 (1992), pp. 161–163 (cited on pages 5, 97).

[18] Pascal Vincent et al. 'Extracting and composing robust features with denoising autoencoders'. In: *Proceedings of the 25th International Conference on Machine Learning*. 2008, pp. 1096–1103 (cited on pages 5, 97).

[19] Mark K Transtrum et al. 'Perspective: Sloppiness and emergent theories in physics, biology, and beyond'. In: *The Journal of chemical physics* 143.1 (2015), 07B201_1 (cited on page 5).

[20] Karl Pearson. 'LIII. On lines and planes of closest fit to systems of points in space'. In: *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2.11 (1901), pp. 559–572 (cited on page 5).

[21] Harold Hotelling. 'Analysis of a complex of statistical variables into principal components.' In: *Journal of educational psychology* 24.6 (1933), p. 417 (cited on page 5).

[22] Naftali Tishby and Noga Zaslavsky. 'Deep learning and the information bottleneck principle'. In: *2015 ieee information theory workshop (itw)*. IEEE. 2015, pp. 1–5 (cited on page 5).

[23] Ravid Shwartz-Ziv and Naftali Tishby. 'Opening the black box of deep neural networks via information'. In: *arXiv preprint arXiv:1703.00810* (2017) (cited on page 5).

[24] Alessio Ansuini et al. 'Intrinsic dimension of data representations in deep neural networks'. In: *Advances in Neural Information Processing Systems* 32 (2019) (cited on page 5).

[25] Matteo Marsili, Iacopo Mastromatteo, and Yasser Roudi. 'On sampling and modeling complex systems'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.09 (2013), P09003 (cited on page 5).

[26] Ryan John Cubero et al. 'Statistical criticality arises in most informative representations'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.6 (2019), p. 063402 (cited on page 5).

[27] O Duranthon, Matteo Marsili, and Rongrong Xie. 'Maximal relevance and optimal learning machines'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.3 (2021), p. 033409 (cited on page 5).

[28] Matteo Marsili and Yasser Roudi. 'Quantifying relevance in learning and inference'. In: *Physics Reports* 963 (2022), pp. 1–43 (cited on page 5).

[29] Rongrong Xie and Matteo Marsili. 'A random energy approach to deep learning'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2022.7 (2022), p. 073404 (cited on page 5).

[30] Andrew M Saxe et al. 'On the information bottleneck theory of deep learning'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12 (2019), p. 124020 (cited on pages 5, 6).

[31] Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. 'i-RevNet: Deep Invertible Networks'. In: *International Conference on Learning Representations*. 2018 (cited on pages 6, 62, 73, 105, 194, 201).

[32] Gemma E Moran et al. 'Identifiable Variational Autoencoders via Sparse Decoding'. In: *arXiv preprint arXiv:2110.10804* (2021) (cited on pages 6, 53).

[33] Aapo Hyvärinen and Erkki Oja. 'Independent component analysis: algorithms and applications'. In: *Neural networks* 13.4-5 (2000), pp. 411–430 (cited on page 7).

[34] Pierre Ablin. 'Exploration of multivariate EEG /MEG signals using non-stationary models'. 2019SACLT051. PhD thesis. 2019 (cited on pages 7, 11).

[35] George Darmois. 'Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire'. In: *Revue de l'Institut international de statistique* (1953), pp. 2–8 (cited on page 9).

[36] VP Skitović. 'On a property of a normal distribution'. In: *Doklady Akad. Nauk*. 1953 (cited on page 9).

[37] Pierre Comon. 'Independent component analysis, a new concept?' In: *Signal processing* 36.3 (1994), pp. 287–314 (cited on pages 9, 10, 33, 37, 222).

[38] Christian Jutten and Anisse Taleb. 'Source separation: from dusk till dawn'. In: () (cited on page 9).

[39] Flávio RM Pavan and Maria D Miranda. 'On the Darmois-Skitovich Theorem and Spatial Independence in Blind Source Separation'. In: *Journal of Communication and Information Systems* 33.1 (2018) (cited on pages 9, 10).

[40] James Clerk Maxwell. 'II. Illustrations of the dynamical theory of gases'. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 20.130 (1860), pp. 21–37 (cited on page 9).

[41] Dinh-Tuan Pham and J-F Cardoso. 'Blind separation of instantaneous mixtures of nonstationary sources'. In: *IEEE Transactions on Signal Processing* 49.9 (2001), pp. 1837–1848 (cited on page 10).

[42] Dinh Tuan Pham and Philippe Garat. 'Blind separation of mixture of independent sources through a quasi-maximum likelihood approach'. In: *IEEE Transactions on Signal Processing* 45.7 (1997), pp. 1712–1725 (cited on pages 10, 91).

[43] Jean-François Cardoso. 'The three easy routes to independent component analysis; contrasts and geometry'. In: *Proc. ICA*. Vol. 2001. 2001 (cited on page 10).

[44] Nicholas D Sidiropoulos, Rasmus Bro, and Georgios B Giannakis. 'Parallel factor analysis in sensor array processing'. In: *IEEE transactions on Signal Processing* 48.8 (2000), pp. 2377–2388 (cited on page 10).

[45] Jean-François Cardoso. 'Blind signal separation: statistical principles'. In: *Proceedings of the IEEE* 86.10 (1998), pp. 2009–2025 (cited on pages 11, 88).

[46] Anthony J Bell and Terrence J Sejnowski. 'An information-maximization approach to blind separation and blind deconvolution'. In: *Neural computation* 7.6 (1995), pp. 1129–1159 (cited on pages 11, 63, 86, 91, 100).

[47] Jean-François Cardoso. 'Infomax and maximum likelihood for blind source separation'. In: *IEEE Signal processing letters* 4.4 (1997), pp. 112–114 (cited on pages 11, 86, 100).

[48] Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. 'Faster ICA under orthogonal constraint'. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 4464–4468 (cited on pages 12, 86, 91, 151).

[49] Oliver J Maclaren and Ruanui Nicholson. 'What can be estimated? Identifiability, estimability, causal inference and ill-posed inverse problems'. In: *arXiv preprint arXiv:1904.02826* (2019) (cited on pages 12, 118).

[50] Oliver J Maclaren and Ruanui Nicholson. 'Models, identifiability, and estimability in causal inference'. In: *38th International Conference on Machine Learning. Workshop on the Neglected Assumptions in Causal Inference. ICML*. 2021 (cited on pages 12, 118).

[51] Hiroshi Sawada, Ryo Mukai, and Shoji Makino. 'Direction of arrival estimation for multiple source signals using independent component analysis'. In: *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings.* Vol. 2. IEEE. 2003, pp. 411–414 (cited on page 12).

[52] Timo Honkela, Aapo Hyvärinen, and Jaakko J Väyrynen. 'WordICA-emergence of linguistic representations for words by independent component analysis'. In: *Natural Language Engineering* 16.3 (2010), pp. 277–308 (cited on page 12).

[53] Erkki Oja, Kimmo Kiviluoto, and Simona Malaroiu. 'Independent component analysis for financial time series'. In: *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*. IEEE. 2000, pp. 111–116 (cited on page 12).

[54] Martin J McKeown and Terrence J Sejnowski. 'Independent component analysis of fMRI data: examining the assumptions'. In: *Human brain mapping* 6.5-6 (1998), pp. 368–372 (cited on pages 12, 83).

[55] Danielle Nuzillard and Albert Bijaoui. 'Blind source separation and analysis of multispectral astronomical images'. In: *Astronomy and Astrophysics Supplement Series* 147.1 (2000), pp. 129–138 (cited on page 12).

[56] Jean-François Cardoso et al. 'Component separation with flexible models. Application to the separation of astrophysical emissions'. In: *arXiv preprint arXiv:0803.1814* (2008) (cited on page 12).

[57] Jean-Francois Cardoso. 'Precision cosmology with the cosmic microwave background'. In: *IEEE Signal Processing Magazine* 27.1 (2009), pp. 55–66 (cited on page 12).

[58] Planck Collaboration et al. 'Planck 2013 results. XII. Component separation'. In: *A* 571 (2013) (cited on page 12).

[59] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Vol. 26. Springer, 2004 (cited on page 13).

[60] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006 (cited on page 13).

[61] George Casella and Roger L Berger. *Statistical inference*. Cengage Learning, 2021 (cited on page 13).

[62] Ilyes Khemakhem et al. 'Variational Autoencoders and Nonlinear ICA: A Unifying Framework'. In: *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS*. Vol. 108. 2020, pp. 2207–2217 (cited on pages 13, 14, 17–19, 21, 39, 52–54, 71, 98, 110, 150).

[63] L. Gresele* et al. 'Independent mechanisms analysis, a new concept?' In: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. *equal contribution. Curran Associates, Inc., Dec. 2021, pp. 28233–28248 (cited on pages 13, 23, 51, 53, 54, 98, 115, 184).

[64] Aapo Hyvärinen and Hiroshi Morioka. 'Unsupervised feature extraction by time-contrastive learning and nonlinear ICA'. In: *Advances in Neural Information Processing Systems*. 2016, pp. 3765–3773 (cited on pages 14, 18, 19, 21, 39, 51, 69, 98, 107, 110, 256, 258).

[65] Aapo Hyvärinen and Hiroshi Morioka. 'Nonlinear ICA of Temporally Dependent Stationary Sources'. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by Aarti Singh and Jerry Zhu. Vol. 54. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, Apr. 2017, pp. 460–469 (cited on pages 14, 18, 19, 21, 52, 69, 75, 107, 256, 258).

[66] Aapo Hyvärinen, Hiroaki Sasaki, and Richard Turner. 'Nonlinear ICA using auxiliary variables and generalized contrastive learning'. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 859–868 (cited on pages 14, 18–21, 52, 69, 73, 98, 110, 117, 120, 126, 210, 211, 213).

[67] Kevin P Murphy. *Machine learning: a probabilistic perspective*. Cambridge, MA: MIT Press, 2012 (cited on pages 15, 43, 47, 107, 258).

[68] Paul K. Rubenstein. *Variational Autoencoders are not autoencoders*. Jan. 2019. URL: http://paulrubenstein.co.uk/variational-autoencoders-are-not-autoencoders/ (cited on pages 15, 43).

[69] G Darmois. 'Analyse des liaisons de probabilité'. In: *Proc. Int. Stat. Conferences 1947*. 1951, p. 231 (cited on pages 16, 105, 244, 245).

[70] Aapo Hyvärinen and Petteri Pajunen. 'Nonlinear independent component analysis: Existence and uniqueness results'. In: *Neural Networks* 12.3 (1999), pp. 429–439 (cited on pages 16, 17, 33, 37, 41, 43, 54, 59, 98, 105, 110, 115, 120, 139, 199, 244, 245).

[71] George Papamakarios et al. 'Normalizing Flows for Probabilistic Modeling and Inference'. In: *Journal of Machine Learning Research* 22.57 (2021), pp. 1–64 (cited on pages 16, 38, 45, 52, 60, 135, 147, 148, 193).

[72] Francesco Locatello et al. 'Challenging common assumptions in the unsupervised learning of disentangled representations'. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4114–4124 (cited on pages 17, 21, 40, 53, 98, 110, 120, 255).

[73] Anisse Taleb and Christian Jutten. 'Source separation in post-nonlinear mixtures'. In: *IEEE Transactions on Signal Processing* 47.10 (1999), pp. 2807–2820 (cited on pages 17, 37, 40).

[74] Shubhangi Ghosh et al. 'On Pitfalls of Identifiability in Unsupervised Learning. A Note on:" Desiderata for Representation Learning: A Causal Perspective"'. In: *arXiv preprint arXiv:2202.06844* (2022) (cited on pages 17, 21).

[75] Stefan Harmeling et al. 'Kernel-based nonlinear blind source separation'. In: *Neural Computation* 15.5 (2003), pp. 1089–1124 (cited on page 19).

[76] Henning Sprekeler, Tiziano Zito, and Laurenz Wiskott. 'An extension of slow feature analysis for nonlinear blind source separation'. In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 921–947 (cited on page 19).

[77] Bernhard Schölkopf. 'Causality for machine learning'. In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 2022, pp. 765–804 (cited on page 19).

[78] Hermanni Hälvä and Aapo Hyvärinen. 'Hidden Markov Nonlinear ICA: Unsupervised Learning from Nonstationary Time Series'. In: *Conference on Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 939–948 (cited on pages 19, 21, 52, 150).

[79] Hermanni Hälvä et al. 'Disentangling identifiable features from noisy data with structured nonlinear ica'. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 1624–1633 (cited on pages 19, 21).

[80] Michael Gutmann and Aapo Hyvärinen. 'Noise-contrastive estimation: A new estimation principle for unnormalized statistical models'. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 2010, pp. 297–304 (cited on pages 19, 69, 100).

[81] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001 (cited on page 19).

[82] Ian Goodfellow et al. 'Generative adversarial nets'. In: *Advances in Neural Information Processing Systems*. 2014, pp. 2672–2680 (cited on pages 19, 40, 209).

[83] L. Gresele* et al. 'The Incomplete Rosetta Stone problem: Identifiability results for Multi-view Nonlinear ICA'. In: *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*. Vol. 115. Proceedings of Machine Learning Research. *equal contribution. PMLR, July 2019, pp. 217–227 (cited on pages 20, 24, 98, 110, 232).

[84] Ilyes Khemakhem, Diederik P Kingma, and Aapo Hyvärinen. 'Variational Autoencoders and Nonlinear ICA: A Unifying Framework'. In: *arXiv preprint arXiv:1907.04809* (2019) (cited on pages 20, 21).

[85] Ilyes Khemakhem et al. 'ICE-BeeM: Identifiable Conditional Energy-Based Deep Models Based on Nonlinear ICA'. In: *Advances in Neural Information Processing Systems 33*. 2020 (cited on pages 20–22, 102).

[86] L. Gresele* et al. 'Relative gradient optimization of the Jacobian term in unsupervised deep learning'. In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. *equal contribution. Curran Associates, Inc., Dec. 2020, pp. 16567–16578 (cited on pages 21, 24, 39, 50, 64, 149, 186, 188).

[87] David Klindt et al. 'Towards nonlinear disentanglement in natural data with temporal sparse coding'. In: *arXiv preprint arXiv:2007.10930* (2020) (cited on page 21).

[88] Roland S. Zimmermann et al. 'Contrastive Learning Inverts the Data Generating Process'. In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. 2021, pp. 12979–12990 (cited on pages 21, 51, 98, 99, 106, 107, 110, 244, 246, 256, 257).

[89] Yujia Zheng, Ignavier Ng, and Kun Zhang. 'On the Identifiability of Nonlinear ICA with Unconditional Priors'. In: *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*. 2022 (cited on pages 21, 115).

[90] Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. 'Function Classes for Identifiable Nonlinear Independent Component Analysis'. In: *UAI 2022 Workshop on Causal Representation Learning* (cited on pages 21, 115).

[91] Peter Sorrenson, Carsten Rother, and Ullrich Köthe. 'Disentanglement by Nonlinear ICA with General Incompressible-flow Networks (GIN)'. In: *International Conference on Learning Representations*. 2019 (cited on page 21).

[92] Frederik Träuble et al. *On Disentangled Representations Learned From Correlated Data*. ICML 2021. 2020 (cited on page 21).

[93] Phillip Lippe et al. 'CITRIS: Causal identifiability from temporal intervened sequences'. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 13557–13603 (cited on pages 21, 119).

[94] Johann Brehmer et al. 'Weakly supervised causal representation learning'. In: *arXiv preprint arXiv:2203.16437* (2022) (cited on pages 21, 23, 118, 119).

[95] Irina Higgins et al. *Towards a Definition of Disentangled Representations*. 2018 (cited on page 21).

[96] Hamza Keurti et al. 'Homomorphism Autoencoder—Learning Group Structured Representations from Interactions'. In: *UAI 2022 Workshop on Causal Representation Learning* (cited on page 21).

[97] Laurent Dinh, David Krueger, and Yoshua Bengio. 'NICE: Non-linear independent components estimation'. In: *arXiv preprint arXiv:1410.8516* (2014) (cited on pages 21, 55, 56, 59, 105, 193).

[98] Ivan Kobyzev, Simon Prince, and Marcus A Brubaker. 'Normalizing flows: Introduction and ideas'. In: *arXiv preprint arXiv:1908.09257* (2019) (cited on pages 21, 60, 193).

[99] George Papamakarios et al. 'Normalizing Flows for Probabilistic Modeling and Inference'. In: *Journal of Machine Learning Research* 22.57 (2021), pp. 1–64 (cited on pages 21, 105).

[100] Aapo Hyvärinen, Patrik O Hoyer, and Mika Inki. 'Topographic independent component analysis'. In: *Neural computation* 13.7 (2001), pp. 1527–1558 (cited on page 21).

[101] T Anderson Keller and Max Welling. 'Topographic VAEs learn equivariant capsules'. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 28585–28597 (cited on pages 21, 120).

[102] Shohei Shimizu et al. 'A linear non-Gaussian acyclic model for causal discovery.' In: *Journal of Machine Learning Research* 7.10 (2006) (cited on pages 21, 119).

[103] Kun Zhang and Aapo Hyvärinen. 'On the identifiability of the post-nonlinear causal model'. In: *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. AUAI Press. 2009, pp. 647–655 (cited on pages 21, 40).

[104] Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. 'Causal discovery with general non-linear relationships using non-linear ICA'. In: *Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 186–195 (cited on pages 21, 119).

[105] Ilyes Khemakhem et al. 'Causal autoregressive flows'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 3520–3528 (cited on pages 21, 119).

[106] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017 (cited on pages 22, 23, 29, 31, 35, 135).

[107] Frederick Eberhardt. 'Green and grue causal variables'. In: *Synthese* 193.4 (2016), pp. 1029–1046 (cited on page 22).

[108] P. K. Rubenstein* et al. 'Causal Consistency of Structural Equation Models'. In: *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*. *equal contribution. Aug. 2017, p. ID 11 (cited on page 22).

[109] Sander Beckers and Joseph Y Halpern. 'Abstracting causal models'. In: *Proceedings of the aaai conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 2678–2685 (cited on page 22).

[110] Fabio Massimo Zennaro. 'Abstraction between Structural Causal Models: A Review of Definitions and Properties'. In: *UAI 2022 Workshop on Causal Representation Learning* (cited on page 22).

[111] John Wu et al. 'Similarity Analysis of Contextual Word Representation Models'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 4638–4655 (cited on page 22).

[112] Luca Moschella et al. 'Relative representations enable zero-shot latent space communication'. In: *arXiv preprint arXiv:2209.15430* (2022) (cited on page 22).

[113] Rui Shu et al. 'Weakly Supervised Disentanglement with Guarantees'. In: *8th International Conference on Learning Representations*. 2020 (cited on page 23).

[114] Francesco Locatello et al. 'Weakly-Supervised Disentanglement Without Compromises'. In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. PMLR, 2020, pp. 6348–6359 (cited on pages 23, 98, 104, 110, 119).

[115] P. Reizinger* et al. 'Embrace the Gap: VAEs Perform Independent Mechanism Analysis'. In: *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. *equal first authorship. Curran Associates, Inc., Dec. 2022 (cited on page 24).

[116] H. Richard* et al. 'Modeling Shared responses in Neuroimaging Studies through MultiView ICA'. In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. *equal contribution. Red Hook, NY: Curran Associates, Inc., Dec. 2020, pp. 19149–19162 (cited on pages 24, 227).

[117] J. von Kügelgen* et al. 'Self-supervised learning with data augmentations provably isolates content from style'. In: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. *equal contribution. Curran Associates, Inc., Dec. 2021, pp. 16451–16467 (cited on pages 24, 29).

[118] Luigi Gresele and Matteo Marsili. 'On maximum entropy and inference'. In: *Entropy* 19.12 (2017), p. 642 (cited on page 25).

[119] G. Parascandolo* et al. 'Learning explanations that are hard to vary'. In: *9th International Conference on Learning Representations (ICLR)*. *equal contribution. May 2021 (cited on page 25).

[120] J. von Kügelgen*, L. Gresele*, and B. Schölkopf. 'Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects'. In: *IEEE Transactions on Artificial Intelligence* 2.1 (2021). *equal contribution, pp. 18–27. DOI: 10.1109/TAI.2021.3073088 (cited on page 25).

[121] L. Gresele* et al. 'Causal Inference Through the Structural Causal Marginal Problem'. In: *Proceedings of the 39th International Conference on Machine Learning (ICML)*. Vol. 162. *equal contribution. PMLR, July 2022, pp. 7793–7824 (cited on pages 25, 118).

[122] Judea Pearl. *Causality*. Cambridge university press, 2009 (cited on pages 29, 31, 102, 117, 118).

[123] B Schölkopf et al. 'On causal and anticausal learning'. In: *29th International Conference on Machine Learning (ICML 2012)*. International Machine Learning Society. 2012, pp. 1255–1262 (cited on pages 29, 31).

[124] Julius von Kügelgen et al. 'Semi-supervised learning, causality, and the conditional cluster assumption'. In: *Conference on Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 1–10 (cited on page 29).

[125] Julius von Kügelgen, Alexander Mey, and Marco Loog. 'Semi-generative modelling: Covariate-shift adaptation with cause and effect features'. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 1361–1369 (cited on page 29).

[126] Daniel Greenfeld and Uri Shalit. 'Robust learning with the hilbert-schmidt independence criterion'. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 3759–3768 (cited on page 29).

[127] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. 'Preventing failures due to dataset shift: Learning predictive models that transport'. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 3118–3127 (cited on page 29).

[128] Dominik Rothenhäusler et al. 'Anchor regression: Heterogeneous data meet causality'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 83.2 (2021), pp. 215–246 (cited on page 29).

[129] Martin Arjovsky et al. 'Invariant risk minimization'. In: *arXiv preprint arXiv:1907.02893* (2019) (cited on page 29).

[130] Christina Heinze-Deml and Nicolai Meinshausen. 'Conditional variance penalties and domain shift robustness'. In: *Machine Learning* 110.2 (2021), pp. 303–348 (cited on page 29).

[131] Sara Magliacane et al. 'Domain adaptation by using causal inference to predict invariant conditional distributions'. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018, pp. 10869–10879 (cited on page 29).

[132] Mingming Gong et al. 'Domain adaptation with conditional transferable components'. In: *International Conference on Machine Learning*. PMLR. 2016, pp. 2839–2848 (cited on page 29).

[133] Judea Pearl and Elias Bareinboim. 'External validity: From do-calculus to transportability across populations'. In: *Statistical Science* 29.4 (2014), pp. 579–595 (cited on page 29).

[134] Mateo Rojas-Carulla et al. 'Invariant models for causal transfer learning'. In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 1309–1342 (cited on pages 29, 110).

[135] K. Zhang et al. 'Domain adaptation under Target and Conditional Shift'. In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by S. Dasgupta and D. McAllester. Vol. 28. JMLR Workshop and Conference Proceedings. 2013, pp. 819–827 (cited on page 29).

[136] Elias Bareinboim, Andrew Forney, and Judea Pearl. 'Bandits with unobserved confounders: A causal approach'. In: *Advances in Neural Information Processing Systems* 28 (2015), pp. 1342–1350 (cited on page 29).

[137] Junzhe Zhang and Elias Bareinboim. 'Near-Optimal Reinforcement Learning in Dynamic Treatment Regimes'. In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 13401–13411 (cited on page 29).

[138] Chaochao Lu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. 'Deconfounding reinforcement learning in observational settings'. In: *arXiv preprint arXiv:1812.10576* (2018) (cited on page 29).

[139] Chaochao Lu et al. 'Sample-Efficient Reinforcement Learning via Counterfactual-Based Data Augmentation'. In: *arXiv preprint arXiv:2012.09092* (2020) (cited on page 29).

[140] Lars Buesing et al. 'Woulda, coulda, shoulda: Counterfactually-guided policy search'. In: *arXiv preprint arXiv:1811.06272* (2018) (cited on page 29).

[141] Sanghack Lee and Elias Bareinboim. 'Structural causal bandits: where to intervene?' In: *Advances in Neural Information Processing Systems 31* 31 (2018) (cited on page 29).

[142] Andrew Forney, Judea Pearl, and Elias Bareinboim. 'Counterfactual data-fusion for online reinforcement learners'. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 1156–1164 (cited on page 29).

[143] Anirudh Goyal et al. 'Recurrent Independent Mechanisms'. In: *9th International Conference on Learning Representations*. 2021 (cited on page 29).

[144] Giambattista Parascandolo et al. 'Learning independent causal mechanisms'. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 4036–4044 (cited on page 29).

[145] Michel Besserve et al. 'Group invariance principles for causal generative models'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pp. 557–565 (cited on pages 29, 31, 33).

[146] Julius von Kügelgen et al. 'Towards causal generative scene models via competition of experts'. In: *ICLR Workshop on "Causal Learning for Decision Making"*. 2020 (cited on pages 29, 98).

[147] M Besserve et al. 'Counterfactuals uncover the modular structure of deep generative models'. In: *Eighth International Conference on Learning Representations (ICLR 2020)*. 2020 (cited on page 29).

[148] Xinwei Shen et al. 'Disentangled Generative Causal Representation Learning'. In: *arXiv preprint arXiv:2010.02637* (2020) (cited on pages 29, 98, 104).

[149] Felix Leeb et al. 'Structural autoencoders improve representations for generation and transfer'. In: *arXiv preprint arXiv:2006.07796* (2020) (cited on pages 29, 98, 104).

[150] Dominik Janzing and Bernhard Schölkopf. 'Causal inference using the algorithmic Markov condition'. In: *IEEE Transactions on Information Theory* 56.10 (2010), pp. 5168–5194 (cited on pages 29, 31, 41).

[151] Dominik Janzing, Raphael Chaves, and Bernhard Schölkopf. 'Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference'. In: *New Journal of Physics* 18.9 (2016), p. 093052 (cited on page 29).

[152] Jan Lemeire and Dominik Janzing. 'Replacing causal faithfulness with algorithmic independence of conditionals'. In: *Minds and Machines* 23.2 (2013), pp. 227–249 (cited on page 29).

[153] Andrei N Kolmogorov. 'On tables of random numbers'. In: *Sankhyā: The Indian Journal of Statistics, Series A* (1963), pp. 369–376 (cited on pages 29, 31).

[154] B. Steudel, D. Janzing, and B. Schölkopf. 'Causal Markov condition for submodular information measures'. In: *Conference on Learning Theory (COLT)*. Ed. by A. Kalai and M. Mohri. Madison, WI, USA: OmniPress, 2010, pp. 464–476 (cited on page 29).

[155] Povilas Daniušis et al. 'Inferring deterministic causal relations'. In: *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. 2010, pp. 143–150 (cited on pages 30, 32, 41, 129).

[156] Dominik Janzing et al. 'Information-geometric approach to inferring causal directions'. In: *Artificial Intelligence* 182 (2012), pp. 1–31 (cited on pages 30, 32, 34, 41, 128–130).

[157] Dominik Janzing, Patrik O Hoyer, and Bernhard Schölkopf. 'Telling cause from effect based on high-dimensional observations'. In: *International Conference on Machine Learning*. 2010 (cited on pages 30–32, 34, 41, 127, 129, 134, 135).

[158] Jakob Zscheischler, Dominik Janzing, and Kun Zhang. 'Testing whether linear equations are causal: A free probability theory approach'. In: *27th Conference on Uncertainty in Artificial Intelligence*. 2011, pp. 839–847 (cited on pages 30–32, 41, 127).

[159] Jan Lemeire and Erik Dirkx. *Causal models as minimal descriptions of multivariate systems*. 2006 (cited on page 31).

[160] Patrik Hoyer et al. 'Nonlinear causal discovery with additive noise models'. In: *Advances in Neural Information Processing Systems* 21 (2008), pp. 689–696 (cited on page 31).

[161] Jonas Peters et al. 'Causal Discovery with Continuous Additive Noise Models'. In: *Journal of Machine Learning Research* 15 (2014), pp. 2009–2053 (cited on page 31).

[162] Jonas Peters and Peter Bühlmann. 'Identifiability of Gaussian structural equation models with equal error variances'. In: *Biometrika* 101.1 (2014), pp. 219–228 (cited on page 31).

[163] Naji Shajarisales et al. 'Telling cause from effect in deterministic linear dynamical systems'. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 285–294 (cited on page 31).

[164] Joris M Mooij et al. 'Distinguishing cause from effect using observational data: methods and benchmarks'. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1103–1204 (cited on page 31).

[165] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 'Causal inference by using invariant prediction: identification and confidence intervals'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 5.78 (2016), pp. 947–1012 (cited on page 31).

[166] Patrick Blöbaum et al. 'Cause-effect inference by comparing regression errors'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pp. 900–909 (cited on page 31).

[167] Dominik Janzing. 'Causal version of Principle of Insufficient Reason and MaxEnt'. In: *arXiv preprint arXiv:2102.03906* (2021) (cited on page 31).

[168] Dominik Janzing and Bernhard Schölkopf. 'Detecting confounding in multivariate linear models via spectral analysis'. In: *Journal of Causal Inference* 6.1 (2018) (cited on page 34).

[169] James Mahoney. 'Beyond correlational analysis: Recent innovations in theory and method'. In: *Sociological forum*. JSTOR. 2001, pp. 575–593 (cited on page 35).

[170] Wesley C Salmon. *Scientific explanation and the causal structure of the world*. Princeton University Press, 2020 (cited on page 35).

[171] Charles Tilly. 'Historical analysis of political processes'. In: *Handbook of sociological theory*. Springer, 2001, pp. 567–588 (cited on page 35).

[172] Khaled Alyani, Marco Congedo, and Maher Moakher. 'Diagonality measures of Hermitian positive-definite matrices with application to the approximate joint diagonalization problem'. In: *Linear Algebra and its Applications* 528 (2017), pp. 290–320 (cited on pages 35, 136, 184).

[173] Gabriel Lamé. *Leçons sur les coordonnées curvilignes et leurs diverses applications*. Mallet-Bachelier, 1859 (cited on page 36).

[174] Gaston Darboux. *Leçons sur les systemes orthogonaux et les coordonnées curvilignes*. Gauthier-Villars, 1910 (cited on page 36).

[175] Parry Moon and Domina Eberle Spencer. *Field theory handbook, including coordinate systems, differential equations and their solutions*. Springer, 1971 (cited on pages 36, 41).

[176] Robert Phillips. 'Liouville's theorem'. In: *Pacific Journal of Mathematics* 28.2 (1969), pp. 397–405 (cited on pages 38, 51, 185, 187).

[177] James Bradbury et al. *JAX: composable transformations of Python+NumPy programs*. Version 0.1.55. 2018 (cited on pages 38, 63, 147).

[178] Jake Bruce et al. *Distrax: Probability distributions in JAX*. Version 0.0.1. 2021 (cited on pages 38, 147).

[179] Chin-Wei Huang et al. 'Neural Autoregressive Flows'. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. PMLR, 2018, pp. 2083–2092 (cited on pages 38, 56, 59, 148, 193, 199).

[180] Ricky T. Q. Chen et al. 'Residual Flows for Invertible Generative Modeling'. In: *Advances in Neural Information Processing Systems*. 2019 (cited on pages 38, 148, 149).

[181] Shun-ichi Amari, Andrzej Cichocki, Howard Hua Yang, et al. 'A new learning algorithm for blind signal separation'. In: *Advances in Neural Information Processing Systems*. 1996, pp. 757–763 (cited on pages 40, 87, 150).

[182] Aapo Hyvärinen. 'Fast and robust fixed-point algorithms for independent component analysis'. In: *IEEE transactions on Neural Networks* 10.3 (1999), pp. 626–634 (cited on pages 40, 63, 89, 152).

[183] Kun Zhang and Laiwan Chan. 'Minimal nonlinear distortion principle for nonlinear independent component analysis'. In: *Journal of Machine Learning Research* 9.Nov (2008), pp. 2455–2487 (cited on pages 40, 43).

[184] Aditya Ramesh, Youngduck Choi, and Yann LeCun. 'A spectral regularizer for unsupervised disentanglement'. In: *arXiv preprint arXiv:1812.01161* (2018) (cited on page 40).

[185] William S. Peebles et al. 'The Hessian Penalty: A Weak Prior for Unsupervised Disentanglement'. In: *ECCV*. Vol. 12351. Springer, 2020, pp. 581–597 (cited on page 40).

[186] Petteri Pajunen. 'Blind source separation using algorithmic information theory'. In: *Neurocomputing* 22.1-3 (1998), pp. 35–48 (cited on page 41).

[187] Petteri Pajunen. 'Blind source separation of natural signals based on approximate complexity minimization'. In: *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*. Citeseer. 1999, p. 267 (cited on page 41).

[188] Geoffrey E Hinton and Lawrence M Parsons. 'Frames of reference and mental imagery'. In: *Attention and performance IX* (1981), pp. 261–277 (cited on page 41).

[189] Michael Mistry, Jonas Buchli, and Stefan Schaal. 'Inverse dynamics control of floating base systems using orthogonal decomposition'. In: *2010 IEEE International Conference on Robotics and Automation*. IEEE. 2010, pp. 3406–3412 (cited on page 41).

[190] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006 (cited on pages 43, 79).

[191] Diederik P. Kingma and Max Welling. 'Auto-Encoding Variational Bayes'. In: *2nd International Conference on Learning Representations (ICLR)*. 2014 (cited on pages 43, 45, 104, 161).

[192] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 'Stochastic Backpropagation and Approximate Inference in Deep Generative Models'. In: *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*. Vol. 32. JMLR Workshop and Conference Proceedings. JMLR.org, 2014, pp. 1278–1286 (cited on pages 43, 45).

[193] Carl Doersch. 'Tutorial on Variational Autoencoders'. In: *ArXiv preprint* abs/1606.05908 (2016) (cited on pages 43, 45, 162).

[194] Irina Higgins et al. 'beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework'. In: *5th International Conference on Learning Representations*. 2017 (cited on pages 43, 53, 120, 161).

[195] Alexander A. Alemi et al. 'Fixing a Broken ELBO'. In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*. Ed. by Jennifer G. Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 159–168 (cited on page 43).

[196] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. 'Variational Inference of Disentangled Latent Concepts from Unlabeled Observations'. In: *International Conference on Learning Representations*. 2018 (cited on page 43).

[197] Ricky TQ Chen et al. 'Isolating sources of disentanglement in VAEs'. In: *Advances in Neural Information Processing Systems*. 2018, pp. 2615–2625 (cited on pages 43, 161, 258).

[198] Hyunjik Kim and Andriy Mnih. 'Disentangling by factorising'. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2649–2658 (cited on pages 43, 161).

[199] Christopher P Burgess et al. 'Understanding disentangling in β-VAE'. In: *arXiv preprint arXiv:1804.03599* (2018) (cited on pages 43, 53, 120).

[200] K Zhang and A Hyvärinen. 'On the Identifiability of the Post-Nonlinear Causal Model'. In: *25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*. AUAI Press. 2009, pp. 647–655 (cited on page 43).

[201] Daniella Horan, Eitan Richardson, and Yair Weiss. 'When Is Unsupervised Disentanglement Possible?' In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 5150–5161 (cited on pages 43, 52, 53, 115, 188).

[202] Michal Rolinek, Dominik Zietlow, and Georg Martius. 'Variational Autoencoders Pursue PCA Directions (by Accident)'. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019, pp. 12406–12415. DOI: 10.1109/cvpr.2019.01269 (cited on pages 43, 44, 47, 49, 52, 53, 161, 182, 183, 188).

[203] Abhishek Kumar and Ben Poole. 'On Implicit Regularization in β-VAEs'. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5480–5490 (cited on pages 43, 44, 46, 47, 49, 52, 53, 161, 182, 183).

[204] Bin Dai, Ziyu Wang, and David Wipf. 'The Usual Suspects? Reassessing Blame for VAE Posterior Collapse'. In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. Vol. 119. PMLR, July 2020, pp. 2313–2322 (cited on page 43).

[205] Dominik Zietlow, Michal Rolinek, and Georg Martius. 'Demystifying Inductive Biases for (Beta-) VAE Based Architectures'. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12945–12954 (cited on pages 43, 44, 182).

[206] Didrik Nielsen et al. 'SurVAE Flows: Surjections to Bridge the Gap between VAEs and Flows'. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle et al. 2020 (cited on pages 44–47, 51–53, 183).

[207] James Lucas et al. 'Don't Blame the ELBO! A Linear VAE Perspective on Posterior Collapse'. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach et al. 2019, pp. 9403–9413 (cited on pages 44, 47, 49, 52, 53, 182–184).

[208] Michael Struwe. *Variational Methods*. Vol. 991. Springer Berlin Heidelberg, 2000 (cited on page 45).

[209] Diederik P. Kingma and Max Welling. 'An Introduction to Variational Autoencoders'. In: *Foundations and Trends® in Machine Learning* 12.4 (2019). arXiv: 1906.02691, pp. 307–392. DOI: 10.1561/2200000056 (cited on page 45).

[210] Danilo Rezende and Shakir Mohamed. 'Variational inference with normalizing flows'. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 1530–1538 (cited on pages 45, 55, 56, 59, 62, 73, 193, 194).

[211] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012 (cited on pages 47, 244).

[212] I. Csiszar and F. Matus. 'Information projections revisited'. In: *IEEE Trans. Inf. Theory* 49.6 (June 2003), pp. 1474–1490. DOI: 10.1109/tit.2003.810633 (cited on page 47).

[213] David A. Klindt et al. 'Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding'. In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021 (cited on pages 51, 52, 188).

[214] Nicholas Watters et al. *Spriteworld: A Flexible, Configurable Reinforcement Learning Environment*. https://github.com/deepmind/spriteworld/. 2019. URL: https://github.com/deepmind/spriteworld/ (cited on pages 51, 52, 188).

[215] Jack Brady and Geoffrey Roeder. 'iSprites: A Dataset for Identifiable Multi-Object representation Learning'. In: *ICML2020: Workshop on Object-Oriented Learning*. 2020 (cited on pages 52, 188).

[216] David L. Donoho and Carrie Grimes. 'Image Manifolds which are Isometric to Euclidean Space'. en. In: *J. Math. Imaging Vis.* 23.1 (July 2005), pp. 5–24. DOI: 10.1007/s10851-005-4965-4. (Visited on 04/05/2022) (cited on pages 52, 53, 188).

[217] Maximilian Seitzer et al. 'On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks'. en. In: 2021. (Visited on 03/19/2022) (cited on page 52).

[218] Oleh Rybkin, Kostas Daniilidis, and Sergey Levine. 'Simple and Effective VAE Training with Calibrated Decoders'. In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 9179–9189 (cited on pages 52, 162).

[219] Edmond Cunningham and Madalina Fiterau. 'A change of variables method for rectangular matrix-vector products'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 2755–2763 (cited on page 52).

[220] Anthony L Caterini et al. 'Rectangular flows for manifold learning'. In: *Advances in Neural Information Processing Systems* 34 (2021) (cited on page 52).

[221] Rui Shu et al. 'Amortized inference regularization'. In: *Advances in Neural Information Processing Systems* 31 (2018) (cited on page 53).

[222] Esteban G Tabak, Eric Vanden-Eijnden, et al. 'Density estimation by dual ascent of the log-likelihood'. In: *Communications in Mathematical Sciences* 8.1 (2010), pp. 217–233 (cited on page 55).

[223] Diederik P Kingma and Prafulla Dhariwal. 'Glow: generative flow with invertible 1× 1 convolutions'. In: *Advances in Neural Information Processing Systems*. 2018, pp. 10236–10245 (cited on pages 55, 105, 193, 194).

[224] Will Grathwohl et al. 'FFJORD: Free-form continuous dynamics for scalable reversible generative models'. In: *arXiv preprint arXiv:1810.01367* (2018) (cited on pages 55, 56, 194).

[225] Esteban G Tabak and Cristina V Turner. 'A family of nonparametric density estimation algorithms'. In: *Communications on Pure and Applied Mathematics* 66.2 (2013), pp. 145–164 (cited on pages 55, 193, 194).

[226] Oren Rippel and Ryan Prescott Adams. 'High-dimensional probability estimation with deep density models'. In: *arXiv preprint arXiv:1302.5125* (2013) (cited on pages 55, 193).

[227] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 'Density estimation using Real NVP'. In: *5th International Conference on Learning Representations*. 2017 (cited on pages 56, 59, 105, 193).

[228] Durk P Kingma et al. 'Improved variational inference with inverse autoregressive flow'. In: *Advances in neural information processing systems*. 2016, pp. 4743–4751 (cited on pages 56, 59, 193).

[229] Tian Qi Chen and David K Duvenaud. 'Neural networks with cheap differential operators'. In: *Advances in Neural Information Processing Systems*. 2019, pp. 9961–9971 (cited on page 56).

[230] Jens Behrmann et al. 'Invertible residual networks'. In: *arXiv preprint arXiv:1811.00995* (2018) (cited on pages 56, 59, 194).

[231] Tian Qi Chen et al. 'Neural ordinary differential equations'. In: *Advances in Neural Information Processing Systems*. 2018, pp. 6572–6583 (cited on pages 56, 73, 194).

[232] J-F Cardoso and Beate H Laheld. 'Equivariant adaptive source separation'. In: *IEEE Transactions on signal processing* 44.12 (1996), pp. 3017–3030 (cited on pages 56, 61, 63, 87).

[233] Shun-Ichi Amari. 'Natural gradient works efficiently in learning'. In: *Neural computation* 10.2 (1998), pp. 251–276 (cited on pages 56, 61, 64).

[234] Matthias W Seeger and Hannes Nickisch. 'Large scale variational inference and experimental design for sparse generalized linear models'. In: *arXiv preprint arXiv:0810.0901* (2008) (cited on page 56).

[235] Martin J Wainwright and Michael I Jordan. 'Graphical models, exponential families, and variational inference'. In: *Foundations and Trends® in Machine Learning* 1.1-2 (2008), pp. 1–305 (cited on page 56).

[236] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 'Variational inference: A review for statisticians'. In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877 (cited on page 56).

[237] Yann LeCun et al. 'Backpropagation applied to handwritten zip code recognition'. In: *Neural computation* 1.4 (1989), pp. 541–551 (cited on pages 58, 62).

[238] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 'Learning representations by back-propagating errors'. In: *Nature* 323.6088 (1986), pp. 533–536 (cited on page 58).

[239] Aidan N Gomez et al. 'The reversible residual network: Backpropagation without storing activations'. In: *Advances in neural information processing systems*. 2017, pp. 2214–2224 (cited on pages 58, 194).

[240] Atilim Gunes Baydin et al. 'Automatic differentiation in machine learning: a survey'. In: *Journal of machine learning research* 18.153 (2018) (cited on pages 59, 196).

[241] K. Hornik, M. Stinchcombe, and H. White. 'Multilayer Feedforward Networks Are Universal Approximators'. In: *Neural Netw.* 2.5 (July 1989), pp. 359–366 (cited on page 59).

[242] Kurt Hornik. 'Approximation capabilities of multilayer feedforward networks'. In: *Neural networks* 4.2 (1991), pp. 251–257 (cited on page 59).

[243] Stefano Squartini, Francesco Piazza, and Ali Shawker. 'New riemannian metrics for improvement of convergence speed in ICA based learning algorithms'. In: *2005 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. 2005, pp. 3603–3606 (cited on page 61).

[244] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009 (cited on page 61).

[245] Silvere Bonnabel. 'Stochastic gradient descent on Riemannian manifolds'. In: *IEEE Transactions on Automatic Control* 58.9 (2013), pp. 2217–2229 (cited on page 61).

[246] Fangchang Ma, Ulas Ayaz, and Sertac Karaman. 'Invertibility of convolutional generative networks from partial measurements'. In: *Advances in Neural Information Processing Systems*. 2018, pp. 9628–9637 (cited on pages 62, 201).

[247] George Papamakarios, Theo Pavlakou, and Iain Murray. 'Masked autoregressive flow for density estimation'. In: *Advances in Neural Information Processing Systems*. 2017, pp. 2338–2347 (cited on pages 65, 105, 193, 205, 206).

[248] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml (cited on page 65).

[249] David Martin et al. 'A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics'. In: *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001.* Vol. 2. IEEE. 2001, pp. 416–423 (cited on page 65).

[250] Yann LeCun, Corinna Cortes, and Christopher JC Burges. 'The MNIST database of handwritten digits, 1998'. In: *URL http://yann. lecun. com/exdb/mnist* 10 (1998), p. 34 (cited on page 65).

[251] Naomi Allen et al. 'UK Biobank: Current status and what it means for epidemiology'. In: *Health Policy and Technology* 1.3 (2012), pp. 123–126 (cited on page 70).

[252] Karla L Miller et al. 'Multimodal population brain imaging in the UK Biobank prospective epidemiological study'. In: *Nature neuroscience* 19.11 (2016), p. 1523 (cited on page 70).

[253] David C Van Essen et al. 'The WU-Minn human connectome project: an overview'. In: *Neuroimage* 80 (2013), pp. 62–79 (cited on pages 70, 83).

[254] Meredith A Shafto et al. 'The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing'. In: *BMC neurology* 14.1 (2014), p. 204 (cited on page 70).

[255] Jean-François Champollion. *Précis du systeme hiéroglyphique des anciens Egyptiens, ou Recherches sur les élémens premiers de cette écriture sacrée, sur leurs diverses combinaisons, et sur les rapports de ce systeme avec les autres méthodes graphiques égyptiennes avec un volume de planches*. Imprimerie royale, 1828 (cited on page 70).

[256] Wikipedia contributors. *Rosetta Stone — Wikipedia, The Free Encyclopedia*. [Online; accessed 18-September-2022]. 2022. URL: https://en.wikipedia.org/w/index.php?title=Rosetta_Stone&oldid=1108751825 (cited on page 70).

[257] H Hotelling. 'Relations between two sets of variates.' In: *Biometrika* (1936) (cited on page 79).

[258] Francis R Bach and Michael I Jordan. 'A probabilistic interpretation of canonical correlation analysis'. In: (2005) (cited on page 79).

[259] Pei Ling Lai and Colin Fyfe. 'Kernel and nonlinear canonical correlation analysis'. In: *International Journal of Neural Systems* 10.05 (2000), pp. 365–377 (cited on page 79).

[260] Kenji Fukumizu, Francis R Bach, and Arthur Gretton. 'Statistical consistency of kernel canonical correlation analysis'. In: *Journal of Machine Learning Research* 8.Feb (2007), pp. 361–383 (cited on page 79).

[261] Galen Andrew et al. 'Deep canonical correlation analysis'. In: *International conference on machine learning*. 2013, pp. 1247–1255 (cited on page 79).

[262] Tomer Michaeli, Weiran Wang, and Karen Livescu. 'Nonparametric canonical correlation analysis'. In: *International Conference on Machine Learning*. 2016, pp. 1967–1976 (cited on page 79).

[263] Bernhard Schölkopf et al. 'Modeling confounding by half-sibling regression'. In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7391–7398 (cited on pages 79, 80).

[264] Po-Hsuan Cameron Chen et al. 'A reduced-dimension fMRI shared response model'. In: *Advances in Neural Information Processing Systems*. 2015, pp. 460–468 (cited on pages 81, 90, 91, 93, 225, 231).

[265] James V Haxby et al. 'A common, high-dimensional model of the representational space in human ventral temporal cortex'. In: *Neuron* 72.2 (2011), pp. 404–416 (cited on pages 81, 90, 227).

[266] Jason R Taylor et al. 'The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample'. In: *Neuroimage* 144 (2017), pp. 262–269 (cited on pages 83, 94).

[267] Cathie Sudlow et al. 'UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age'. In: *PLoS medicine* 12.3 (2015) (cited on page 83).

[268] Saurabh Sonkusare, Michael Breakspear, and Christine Guo. 'Naturalistic Stimuli in Neuroscience: Critically Acclaimed'. In: *Trends in Cognitive Sciences* 23 (June 2019). DOI: 10.1016/j.tics.2019.05.004 (cited on page 83).

[269] Scott Makeig et al. 'Independent component analysis of electroencephalographic data'. In: *Advances in neural information processing systems*. 1996, pp. 145–151 (cited on page 83).

[270] Ricardo Vigário et al. 'Independent component analysis for identification of artifacts in magnetoencephalographic recordings'. In: *Advances in neural information processing systems*. 1998, pp. 229–235 (cited on page 83).

[271] Karl J Friston et al. 'Statistical parametric maps in functional imaging: a general linear approach'. In: *Human brain mapping* 2.4 (1994), pp. 189–210 (cited on page 83).

[272] Jean-Baptiste Poline and Matthew Brett. 'The general linear model and fMRI: does love last forever?' In: *Neuroimage* 62.2 (2012), pp. 871–880 (cited on page 83).

[273] Christian F Beckmann et al. 'Investigations into resting-state connectivity using independent component analysis'. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1457 (2005), pp. 1001–1013 (cited on page 83).

[274] Sanna Malinen, Yevhen Hlushchuk, and Riitta Hari. 'Towards natural stimulation in fMRI—issues of data analysis'. In: *Neuroimage* 35.1 (2007), pp. 131–139 (cited on page 84).

[275] Andreas Bartels and Semir Zeki. 'Brain dynamics during natural viewing conditions—a new guide for mapping connectivity in vivo'. In: *Neuroimage* 24.2 (2005), pp. 339–349 (cited on page 84).

[276] Vince D Calhoun et al. 'Different activation dynamics in multiple neural systems during simulated driving'. In: *Human brain mapping* 16.3 (2002), pp. 158–167 (cited on page 84).

[277] Tzyy-Ping Jung et al. 'Extended ICA removes artifacts from electroencephalographic recordings'. In: *Advances in neural information processing systems*. 1998, pp. 894–900 (cited on page 84).

[278] Ricardo Vigário et al. 'Independent component approach to the analysis of EEG and MEG recordings'. In: *IEEE transactions on biomedical engineering* 47.5 (2000), pp. 589–593 (cited on page 84).

[279] Arnaud Delorme et al. 'Independent EEG sources are dipolar'. In: *PloS one* 7.2 (2012) (cited on pages 84, 91).

[280] Aapo Hyvärinen. 'Independent component analysis: recent advances'. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371.1984 (2013), p. 20110534 (cited on pages 84, 119).

[281] Niklas Pfister et al. 'Robustifying independent component analysis by adjusting for group-wise stationary noise'. In: *Journal of Machine Learning Research* 20.147 (2019), pp. 1–50 (cited on pages 84, 232).

[282] Markus Svensén, Frithjof Kruggel, and Habib Benali. 'ICA of fMRI group study data'. In: *NeuroImage* 16.3 (2002), pp. 551–563 (cited on pages 84, 232).

[283] Vince D Calhoun, Jingyu Liu, and Tülay Adalı. 'A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data'. In: *Neuroimage* 45.1 (2009), S163–S172 (cited on pages 84, 89).

[284] Rene Huster, Sergey M Plis, and Vince D Calhoun. 'Group-level component analyses of EEG: validation and evaluation'. In: *Frontiers in neuroscience* 9 (2015), p. 254 (cited on page 84).

[285] Michael Zibulevsky. 'Blind source separation with relative newton method'. In: *Proc. ICA*. Vol. 2003. 2003, pp. 897–902 (cited on page 86).

[286] Pierre Ablin. 'Exploration of multivariate EEG/MEG signals using non-stationary models'. PhD thesis. Université Paris-Saclay (ComUE), 2019 (cited on page 87).

[287] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006 (cited on page 87).

[288] Vince D Calhoun et al. 'A method for making group inferences from functional MRI data using independent component analysis'. In: *Human brain mapping* 14.3 (2001), pp. 140–151 (cited on pages 89, 231).

[289] Gaël Varoquaux et al. 'CanICA: Model-based extraction of reproducible group-level ICA patterns from fMRI time series'. In: *arXiv preprint arXiv:0911.4650* (2009) (cited on pages 89, 232).

[290] Tom Eichele et al. 'EEGIFT: group independent component analysis for event-related EEG data'. In: *Computational intelligence and neuroscience* 2011 (2011) (cited on pages 89, 91, 231).

[291] Christian F Beckmann et al. 'Group comparison of resting-state fMRI data using multi-subject ICA and dual regression'. In: *Neuroimage* 47.Suppl 1 (2009), S148 (cited on page 89).

[292] Hejia Zhang et al. 'A searchlight factor model approach for locating shared information in multi-subject fMRI analysis'. In: *arXiv preprint arXiv:1609.09432* (2016) (cited on pages 89, 92, 93, 229–231).

[293] Ying Guo and Giuseppe Pagnoni. 'A unified framework for group independent component analysis for multi-subject fMRI data'. In: *NeuroImage* 42.3 (2008), pp. 1078–1093 (cited on pages 90, 231).

[294] Olivier Bermond and Jean-François Cardoso. 'Approximate likelihood for noisy mixtures'. In: *Proc. ICA*. Vol. 99. Citeseer. 1999, pp. 325–330 (cited on page 90).

[295] Kaare Brandt Petersen, Ole Winther, and Lars Kai Hansen. 'On the slow convergence of EM and VBEM in low-noise linear models'. In: *Neural computation* 17.9 (2005), pp. 1921–1926 (cited on page 90).

[296] Christian F Beckmann and Stephen M Smith. 'Tensorial extensions of independent component analysis for multisubject fMRI analysis'. In: *Neuroimage* 25.1 (2005), pp. 294–311 (cited on pages 90, 231).

[297] Fengyu Cong et al. 'Validating rationale of group-level component analysis based on estimating number of sources in EEG through model order selection'. In: *Journal of neuroscience methods* 212.1 (2013), pp. 165–172 (cited on pages 90, 232).

[298] Vera A Grin-Yatsenko et al. 'Independent component approach to the analysis of EEG recordings at early stages of depressive disorders'. In: *Clinical Neurophysiology* 121.3 (2010), pp. 281–289 (cited on page 90).

[299] Vince D Calhoun et al. 'fMRI activation in a visual-perception task: network of areas detected using the general linear model and independent components analysis'. In: *NeuroImage* 14.5 (2001), pp. 1080–1088 (cited on pages 90, 91).

[300] R. P. Monti and A. Hyvärinen. 'A unified probabilistic model for learning latent factors and their connectivities from high-dimensional data'. In: *Proc. 34th Conf. on Uncertainty in Artificial Intelligence (UAI2018)*. Monterey, California, 2018 (cited on page 90).

[301] Petr Tichavsky and Zbynek Koldovsky. 'Optimal pairing of signal components separated by blind techniques'. In: *IEEE Signal Processing Letters* 11.2 (2004), pp. 119–122 (cited on page 90).

[302] Fabrizio Esposito et al. 'Independent component analysis of fMRI group studies by self-organizing clustering'. In: *Neuroimage* 25.1 (2005), pp. 193–205 (cited on pages 90, 231).

[303] Nima Bigdely-Shamlo et al. 'Measure projection analysis: a probabilistic approach to EEG source comparison and multi-subject inference'. In: *Neuroimage* 72 (2013), pp. 287–303 (cited on pages 90, 231).

[304] Po-Hsuan Chen et al. 'A convolutional autoencoder for multi-subject fMRI data aggregation'. In: *arXiv preprint arXiv:1608.04846* (2016) (cited on pages 90, 229–231).

[305] Jacek P Dmochowski et al. 'Correlated components of ongoing EEG point to emotionally laden attention–a possible marker of engagement?' In: *Frontiers in human neuroscience* 6 (2012), p. 112 (cited on page 90).

[306] Simon Kamronn, Andreas Trier Poulsen, and Lars Kai Hansen. 'Multiview Bayesian correlated component analysis'. In: *Neural computation* 27.10 (2015), pp. 2207–2230 (cited on pages 90, 94).

[307] Lang Tong et al. 'Indeterminacy and identifiability of blind identification'. In: *IEEE Transactions on circuits and systems* 38.5 (1991), pp. 499–509 (cited on page 91).

[308] Adel Belouchrani et al. 'A blind source separation technique using second-order statistics'. In: *IEEE Transactions on signal processing* 45.2 (1997), pp. 434–444 (cited on page 91).

[309] Ana S Lukic et al. 'An ICA algorithm for analyzing multiple data sets'. In: *Proceedings. International Conference on Image Processing*. Vol. 2. IEEE. 2002, pp. II–II (cited on page 91).

[310] Marco Congedo et al. 'Group independent component analysis of resting state EEG in large normative samples'. In: *International Journal of Psychophysiology* 78.2 (2010), pp. 89–99 (cited on page 91).

[311] John D Hunter. 'Matplotlib: A 2D graphics environment'. In: *Computing in science & engineering* 9.3 (2007), pp. 90–95 (cited on page 91).

[312] Fabian Pedregosa et al. 'Scikit-learn: Machine learning in Python'. In: *The Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cited on pages 91, 152).

[313] Alexandre Gramfort et al. 'MEG and EEG data analysis with MNE-Python'. In: *Frontiers in neuroscience* 7 (2013), p. 267 (cited on page 91).

[314] Alexandre Abraham et al. 'Machine learning for neuroimaging with scikit-learn'. In: *Frontiers in neuroinformatics* 8 (2014), p. 14 (cited on page 91).

[315] Manoj Kumar et al. 'BrainIAK tutorials: User-friendly learning materials for advanced fMRI analysis'. In: *PLoS computational biology* 16.1 (2020), e1007549 (cited on page 91).

[316] Stephen M Smith et al. 'Group-PCA for very large fMRI datasets'. In: *Neuroimage* 101 (2014), pp. 738–749 (cited on pages 91, 231).

[317] Janice Chen et al. 'Shared memories reveal shared structure in neural activity across individuals'. In: *Nature neuroscience* 20.1 (2017), pp. 115–125 (cited on pages 92, 226).

[318] Michael Hanke et al. 'A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie'. In: *Scientific data* 1 (2014), p. 140003 (cited on pages 92, 227).

[319] Ana Luísa Pinho et al. 'Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping'. In: *Scientific data* 5 (2018) (cited on pages 92, 227).

[320] Hejia Zhang, Po-Hsuan Chen, and Peter Ramadge. 'Transfer learning on fMRI datasets'. In: *International Conference on Artificial Intelligence and Statistics*. 2018, pp. 595–603 (cited on page 92).

[321] J Swaroop Guntupalli, Ma Feilong, and James V Haxby. 'A computational model of shared fine-scale structure in the human connectome'. In: *PLoS computational biology* 14.4 (2018), e1006120 (cited on page 93).

[322] Samuel A Nastase et al. 'Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space'. In: *bioRxiv* (2019). DOI: 10.1101/741975 (cited on page 93).

[323] Roberto Domingo Pascual-Marqui et al. 'Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details'. In: *Methods Find Exp Clin Pharmacol* 24.Suppl D (2002), pp. 5–12 (cited on page 94).

[324] Zhirong Wu et al. 'Unsupervised Feature Learning via Non-Parametric Instance Discrimination'. In: *Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, 2018, pp. 3733–3742 (cited on pages 97, 99, 102).

[325] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 'Representation learning with contrastive predictive coding'. In: *arXiv preprint arXiv:1807.03748* (2018) (cited on pages 97, 99, 100, 256).

[326] Olivier J. Hénaff. 'Data-Efficient Image Recognition with Contrastive Predictive Coding'. In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 4182–4192 (cited on pages 97, 99).

[327] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 'Contrastive Multiview Coding'. In: *Computer Vision - ECCV 2020*. Vol. 12356. Springer, 2020, pp. 776–794 (cited on pages 97, 99).

[328] Kaiming He et al. 'Momentum Contrast for Unsupervised Visual Representation Learning'. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9726–9735 (cited on pages 97, 99).

[329] Ting Chen et al. 'A simple framework for contrastive learning of visual representations'. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1597–1607 (cited on pages 97, 99, 100, 105, 106, 108–110, 120, 253, 254, 256, 257).

[330] Jean-Bastien Grill et al. 'Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning'. In: *Advances in Neural Information Processing Systems 33*. 2020 (cited on pages 97, 99, 100, 105, 110).

[331] Xinlei Chen and Kaiming He. 'Exploring Simple Siamese Representation Learning'. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2021, pp. 15750–15758 (cited on pages 97, 99, 100, 105, 110).

[332] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. 'Learning to see by moving'. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 37–45 (cited on page 97).

[333] Carl Doersch, Abhinav Gupta, and Alexei A Efros. 'Unsupervised visual representation learning by context prediction'. In: *IEEE International Conference on Computer Vision*. 2015, pp. 1422–1430 (cited on page 97).

[334] Xiaolong Wang and Abhinav Gupta. 'Unsupervised learning of visual representations using videos'. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2794–2802 (cited on page 97).

[335] Mehdi Noroozi and Paolo Favaro. 'Unsupervised learning of visual representations by solving jigsaw puzzles'. In: *European conference on computer vision*. Springer. 2016, pp. 69–84 (cited on page 97).

[336] Ronan Collobert et al. 'Natural Language Processing (Almost) from Scratch'. In: *Journal of Machine Learning Research* 12 (2011), pp. 2461–2505 (cited on page 97).

[337] Tomas Mikolov et al. 'Distributed representations of words and phrases and their compositionality'. In: *Advances in Neural Information Processing Systems*. 2013, pp. 3111–3119 (cited on page 97).

[338] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 'Glove: Global vectors for word representation'. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543 (cited on page 97).

[339] Lajanugen Logeswaran and Honglak Lee. 'An efficient framework for learning sentence representations'. In: *6th International Conference on Learning Representations*. 2018 (cited on page 97).

[340] Jacob Devlin et al. 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2019, pp. 4171–4186 (cited on page 97).

[341] Alec Radford et al. 'Improving language understanding by generative pre-training'. In: *Technical report, OpenAI* (2018) (cited on page 97).

[342] Yinhan Liu et al. 'Roberta: A robustly optimized bert pretraining approach'. In: *arXiv preprint arXiv:1907.11692* (2019) (cited on page 97).

[343] Tom Brown et al. 'Language Models are Few-Shot Learners'. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 1877–1901 (cited on page 97).

[344] Alexei Baevski et al. 'wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations'. In: *Advances in Neural Information Processing Systems 33*. 2020 (cited on page 97).

[345] Alexei Baevski, Steffen Schneider, and Michael Auli. 'vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations'. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020 (cited on page 97).

[346] Mirco Ravanelli et al. 'Multi-task self-supervised learning for robust speech recognition'. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 6989–6993 (cited on page 97).

[347] Steffen Schneider et al. 'wav2vec: Unsupervised Pre-Training for Speech Recognition'. In: *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*. 2019, pp. 3465–3469 (cited on page 97).

[348] R. Hadsell, S. Chopra, and Y. LeCun. 'Dimensionality Reduction by Learning an Invariant Mapping'. In: *IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 2. 2006, pp. 1735–1742 (cited on pages 97, 102).

[349] S. Chopra, R. Hadsell, and Y. LeCun. 'Learning a similarity metric discriminatively, with application to face verification'. In: *IEEE Conference on Computer Vision and Pattern Recognition*. Vol. 1. 2005, pp. 539–546 (cited on page 97).

[350] Jane Bromley et al. 'Signature verification using a "siamese" time delay neural network'. In: *Advances in Neural Information Processing Systems* 6 (1993), pp. 737–744 (cited on page 97).

[351] Olivier Chapelle and Bernhard Schölkopf. 'Incorporating invariances in nonlinear SVMs'. In: *Advances in Neural Information Processing Systems 14*. Cambridge, MA, USA: MIT Press, 2002, pp. 609–616 (cited on page 98).

[352] Tri Dao et al. 'A kernel theory of modern data augmentation'. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 1528–1537 (cited on page 98).

[353] Shuxiao Chen, Edgar Dobriban, and Jane H Lee. 'A group-theoretic framework for data augmentation'. In: *Journal of Machine Learning Research* 21.245 (2020), pp. 1–71 (cited on page 98).

[354] Bernhard Schölkopf. 'Causality for machine learning'. In: *arXiv preprint arXiv:1911.10500* (2019) (cited on page 98).

[355] Mengyue Yang et al. 'CausalVAE: Structured Causal Disentanglement in Variational Autoencoder'. In: *arXiv preprint arXiv:2004.08697* (2020) (cited on pages 98, 104).

[356] Raphael Suter et al. 'Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness'. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6056–6065 (cited on page 98).

[357] Jovana Mitrovic et al. 'Representation Learning via Invariant Causal Mechanisms'. In: *9th International Conference on Learning Representations*. 2021 (cited on pages 98, 102).

[358] Chaochao Lu et al. 'Nonlinear Invariant Risk Minimization: A Causal Approach'. In: *arXiv preprint arXiv:2102.12353* (2021) (cited on page 98).

[359] David Klindt et al. 'Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding'. In: *International Conference on Learning Representations (ICLR)* (2021) (cited on pages 98, 110, 119).

[360] Ekin D Cubuk et al. 'Autoaugment: Learning augmentation strategies from data'. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 113–123 (cited on page 99).

[361] Ekin D Cubuk et al. 'Randaugment: Practical automated data augmentation with a reduced search space'. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 702–703 (cited on page 99).

[362] Christian Szegedy et al. 'Going Deeper with Convolutions'. In: *Computer Vision and Pattern Recognition (CVPR)*. 2015 (cited on page 99).

[363] Andrew G. Howard. *Some Improvements on Deep Convolutional Neural Network Based Image Classification*. 2013 (cited on page 99).

[364] Terrance DeVries and Graham W Taylor. 'Improved Regularization of Convolutional Neural Networks with Cutout'. In: *arXiv preprint arXiv:1708.04552* (2017) (cited on page 99).

[365] Jure Zbontar et al. 'Barlow Twins: Self-Supervised Learning via Redundancy Reduction'. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. 2021, pp. 12310–12320 (cited on pages 99, 100, 105, 106, 109, 110, 254).

[366] Michael Gutmann and Aapo Hyvärinen. 'Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics.' In: *The Journal of Machine Learning Research* 13 (2012), pp. 307–361 (cited on page 100).

[367] Michael Tschannen et al. 'On Mutual Information Maximization for Representation Learning'. In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020 (cited on page 100).

[368] Ben Poole et al. 'On variational bounds of mutual information'. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 5171–5180 (cited on page 100).

[369] R. Devon Hjelm et al. 'Learning deep representations by mutual information estimation and maximization'. In: *7th International Conference on Learning Representations*. 2019 (cited on page 100).

[370] Philip Bachman, R. Devon Hjelm, and William Buchwalter. 'Learning Representations by Maximizing Mutual Information Across Views'. In: *Advances in Neural Information Processing Systems 32*. 2019, pp. 15509–15519 (cited on page 100).

[371] Ralph Linsker. 'Self-organization in a perceptual network'. In: *Computer* 21.3 (1988), pp. 105–117 (cited on page 100).

[372] Ralph Linsker. 'An application of the principle of maximum information preservation to linear systems'. In: *Advances in Neural Information Processing Systems*. 1989, pp. 186–194 (cited on page 100).

[373] Te-Won Lee, Mark Girolami, and Terrence J Sejnowski. 'Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources'. In: *Neural computation* 11.2 (1999), pp. 417–441 (cited on page 100).

[374] Tongzhou Wang and Phillip Isola. 'Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere'. In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 9929–9939 (cited on pages 100, 106, 110).

[375] Frederick Eberhardt and Richard Scheines. 'Interventions and causal inference'. In: *Philosophy of science* 74.5 (2007), pp. 981–995 (cited on page 102).

[376] Maximilian Ilse, Jakub M Tomczak, and Patrick Forré. 'Selecting data augmentation for simulating interventions'. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 4555–4562 (cited on page 102).

[377] Aapo Hyvärinen and Patrik Hoyer. 'Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces'. In: *Neural computation* 12.7 (2000), pp. 1705–1720 (cited on page 103).

[378] Quoc V Le et al. 'Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis'. In: *CVPR 2011*. IEEE. 2011, pp. 3361–3368 (cited on page 103).

[379] Fabian Theis. 'Towards a general independent subspace analysis'. In: *Advances in Neural Information Processing Systems* 19 (2006), pp. 1361–1368 (cited on page 103).

[380] Michael A Casey and Alex Westner. 'Separation of mixed audio sources by independent subspace analysis'. In: *ICMC*. 2000, pp. 154–161 (cited on page 103).

[381] Haruo Hosoya. 'Group-based Learning of Disentangled Representations with Generalizability for Novel Contents.' In: *IJCAI*. 2019, pp. 2506–2513 (cited on page 104).

[382] Jens Behrmann et al. 'Invertible residual networks'. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 573–582 (cited on page 105).

[383] Martin Edward Newell. 'The Utilization of Procedure Models in Digital Image Synthesis.' AAI7529894. PhD thesis. The University of Utah, 1975 (cited on pages 107, 247).

[384] Greg Turk and Marc Levoy. 'Zippered polygon meshes from range images'. In: *Proceedings of the 21th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1994, Orlando, FL, USA, July 24-29, 1994*. Ed. by Dino Schweitzer, Andrew S. Glassner, and Mike Keeler. ACM, 1994, pp. 311–318 (cited on pages 107, 247).

[385] *The Stanford 3D Scanning Repository*. http://graphics.stanford.edu/data/3Dscanrep/. 2021 (cited on pages 107, 247).

[386] *Keenan's 3D Model Repository*. https://www.cs.cmu.edu/ kmcrane/Projects/ModelRepository/. 2021 (cited on pages 107, 247).

[387] Venkat Krishnamurthy and Marc Levoy. 'Fitting Smooth Surfaces to Dense Polygon Meshes'. In: *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, New Orleans, LA, USA, August 4-9, 1996*. Ed. by John Fujii. ACM, 1996, pp. 313–324 (cited on pages 107, 247).

[388] Emil Praun, Adam Finkelstein, and Hugues Hoppe. 'Lapped textures'. In: *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2000, New Orleans, LA, USA, July 23-28, 2000*. Ed. by Judith R. Brown and Kurt Akeley. ACM, 2000, pp. 465–470 (cited on pages 107, 247).

[389] *Suggestive Contour Gallery*. https://gfx.cs.princeton.edu/proj/sugcon/models/. 2021 (cited on pages 107, 247).

[390] *Animal Diversity Web*. https://animaldiversity.org/accounts/Lepus_othus/. 2021 (cited on pages 107, 247).

[391] *Churchill Polar Bears*. https://churchillpolarbears.org/churchill/. 2021 (cited on pages 107, 247).

[392] Kaiming He et al. 'Deep Residual Learning for Image Recognition'. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2016 (cited on pages 108, 257).

[393] Muhammad Waleed Gondal et al. 'On the Transfer of Inductive Bias from Simulation to the Real World: a New Disentanglement Dataset'. In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 15740–15751 (cited on pages 109, 250, 255).

[394] Robert Geirhos et al. 'Shortcut learning in deep neural networks'. In: *Nature Machine Intelligence* 2.11 (Nov. 2020), pp. 665–673 (cited on page 109).

[395] Mohammad Pezeshki et al. 'Gradient starvation: A learning proclivity in neural networks'. In: *arXiv preprint arXiv:2011.09468* (2020) (cited on page 109).

[396] Robert Geirhos et al. 'ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness'. In: *International Conference on Learning Representations (ICLR)* (2019) (cited on page 109).

[397] Yuandong Tian, Xinlei Chen, and Surya Ganguli. 'Understanding self-supervised learning dynamics without contrastive pairs'. In: *Proceedings of the 38th International Conference on Machine Learning*, ed. by Marina Meila and Tong Zhang. Vol. 139. 2021, pp. 10268–10278 (cited on pages 110, 111).

[398] Sanjeev Arora et al. 'A theoretical analysis of contrastive unsupervised representation learning'. In: *36th International Conference on Machine Learning*. International Machine Learning Society (IMLS). 2019, pp. 9904–9923 (cited on page 111).

[399] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. 'Contrastive estimation reveals topic posterior information to linear models'. In: *arXiv preprint arXiv:2003.02234* (2020) (cited on page 111).

[400] Jason D Lee et al. 'Predicting what you already know helps: Provable self-supervised learning'. In: *arXiv preprint arXiv:2008.01064* (2020) (cited on page 111).

[401] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. 'Contrastive learning, multi-view redundancy, and linear models'. In: *Algorithmic Learning Theory*. PMLR. 2021, pp. 1179–1206 (cited on page 111).

[402] Yao-Hung Hubert Tsai et al. 'Self-supervised Learning from a Multi-view Perspective'. In: *International Conference on Learning Representations*. 2020 (cited on page 111).

[403] Yuxiang Wei et al. 'Orthogonal jacobian regularization for unsupervised disentanglement in image generation'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 6721–6730 (cited on page 115).

[404] Edmond Cunningham, Adam D Cobb, and Susmit Jha. 'Principal Component Flows'. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 4492–4519 (cited on page 115).

[405] Joanna Sliwa et al. 'Probing the Robustness of Independent Mechanism Analysis for Representation Learning'. In: *UAI 2022 Workshop on Causal Representation Learning* (cited on page 116).

[406] Dániel Csaba and Bálint Szoke. *Learning with misspecified models*. Tech. rep. (cited on page 116).

[407] Rhea Chowers and Yair Weiss. 'Why do CNNs Learn Consistent Representations in their First Layer Independent of Labels and Architecture?' In: *arXiv preprint arXiv:2206.02454* (2022) (cited on page 116).

[408] Alessandro Ingrosso and Sebastian Goldt. 'Data-driven emergence of convolutional structure in neural networks'. In: *arXiv preprint arXiv:2202.00565* (2022) (cited on page 116).

[409] Jason Yosinski et al. 'How transferable are features in deep neural networks?' In: *Advances in neural information processing systems* 27 (2014) (cited on page 116).

[410] Ilyes Khemakhem. 'Advances in identifiability of nonlinear probabilistic models'. PhD thesis. UCL (University College London), 2022 (cited on page 116).

[411] Casper Kaae Sønderby et al. 'Ladder variational autoencoders'. In: *Advances in neural information processing systems* 29 (2016) (cited on page 116).

[412] Shengjia Zhao, Jiaming Song, and Stefano Ermon. 'Learning hierarchical features from generative models'. In: *arXiv preprint arXiv:1702.08396* (2017) (cited on page 116).

[413] Philip Botros and Jakub M Tomczak. 'Hierarchical vampprior variational fair auto-encoder'. In: *arXiv preprint arXiv:1806.09918* (2018) (cited on page 116).

[414] Jakub Tomczak and Max Welling. 'VAE with a VampPrior'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pp. 1214–1223 (cited on page 116).

[415] Wenju Xu, Shawn Keshmiri, and Guanghui Wang. 'Stacked wasserstein autoencoder'. In: *Neurocomputing* 363 (2019), pp. 195–204 (cited on page 116).

[416] Arash Vahdat and Jan Kautz. 'NVAE: A deep hierarchical variational autoencoder'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 19667–19679 (cited on page 116).

[417] Qi Lyu and Xiao Fu. 'On Finite-Sample Identifiability of Contrastive Learning-Based Nonlinear Independent Component Analysis'. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 14582–14600 (cited on page 116).

[418] Omar Chehab, Alexandre Gramfort, and Aapo Hyvärinen. 'The optimal noise in noise-contrastive learning is not what you think'. In: *Uncertainty in Artificial Intelligence*. PMLR. 2022, pp. 307–316 (cited on page 116).

[419] Qi Lyu et al. 'Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective'. In: *International Conference on Learning Representations*. 2021 (cited on page 116).

[420] Elias Bareinboim et al. 'On Pearl's hierarchy and the foundations of causal inference'. In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 2022, pp. 507–556 (cited on pages 117, 118).

[421] Judea Pearl and Dana Mackenzie. *The book of Why: the new science of cause and effect*. Basic books, 2018 (cited on page 117).

[422] Judea Pearl and Elias Bareinboim. 'External validity: From do-calculus to transportability across populations'. In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 2022, pp. 451–482 (cited on page 117).

[423] Yimin Huang and Marco Valtorta. 'Pearl's calculus of intervention is complete'. In: *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. 2006, pp. 217–224 (cited on page 117).

[424] Ilya Shpitser and Judea Pearl. 'Identification of conditional interventional distributions'. In: *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. 2006, pp. 437–444 (cited on page 117).

[425] Judea Pearl. 'Interview with Judea Pearl'. In: *Observational Studies* 8.2 (2022), pp. 23–36 (cited on page 117).

[426] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015 (cited on page 118).

[427] Scott Cunningham. 'Causal inference'. In: *Causal Inference*. Yale University Press, 2021 (cited on page 118).

[428] M.A. Hernan and J.M. Robins. *Causal Inference*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press, 2023 (cited on page 118).

[429] Guido W Imbens. 'Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics'. In: *Journal of Economic Literature* 58.4 (2020), pp. 1129–79 (cited on page 118).

[430] James Robins. 'The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies'. In: *Health service research methodology: a focus on AIDS* (1989), pp. 113–159 (cited on page 118).

[431] Charles F Manski. 'Nonparametric bounds on treatment effects'. In: *The American Economic Review* 80.2 (1990), pp. 319–323 (cited on page 118).

[432] Alexander Balke and Judea Pearl. 'Bounds on treatment effects from studies with imperfect compliance'. In: *Journal of the American Statistical Association* 92.439 (1997), pp. 1171–1176 (cited on page 118).

[433] Jin Tian and Judea Pearl. 'Probabilities of causation: Bounds and identification'. In: *Annals of Mathematics and Artificial Intelligence* 28.1 (2000), pp. 287–313 (cited on page 118).

[434] Sébastien Lachapelle et al. 'Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA'. In: *Conference on Causal Learning and Reasoning*. PMLR. 2022, pp. 428–484 (cited on pages 118, 119).

[435] Sebastien Lachapelle and Simon Lacoste-Julien. 'Partial Disentanglement via Mechanism Sparsity'. In: *UAI 2022 Workshop on Causal Representation Learning* (cited on page 118).

[436] Clark Glymour, Kun Zhang, and Peter Spirtes. 'Review of causal discovery methods based on graphical models'. In: *Frontiers in genetics* 10 (2019), p. 524 (cited on page 119).

[437] Thomas S Verma and Judea Pearl. 'Equivalence and synthesis of causal models'. In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 2022, pp. 221–236 (cited on page 119).

[438] Aapo Hyvärinen. *Causal Discovery and Latent-Variable Models*. Talk at the First Workshop on Causal Representation Learning at UAI 2022. Aug. 2022. URL: https://crl-uai-2022.github.io/slides-recording/ (cited on page 119).

[439] Kartik Ahuja et al. 'Interventional Causal Representation Learning'. In: *arXiv preprint arXiv:2209.11924* (2022) (cited on page 119).

[440] Taco Cohen. 'Towards a Grounded Theory of Causation for Embodied AI'. In: *UAI 2022 Workshop on Causal Representation Learning* (cited on page 119).

[441] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018 (cited on page 119).

[442] Andrea Dittadi et al. 'The role of pretrained representations for the ood generalization of rl agents'. In: *arXiv preprint arXiv:2107.05686* (2021) (cited on page 119).

[443] Ossama Ahmed et al. 'CausalWorld: A Robotic Manipulation Benchmark for Causal Structure and Transfer Learning'. In: *International Conference on Learning Representations*. 2020 (cited on page 119).

[444] Tobias Leemann et al. 'Disentangling Embedding Spaces with Minimal Distributional Assumptions'. In: *arXiv preprint arXiv:2206.13872* (2022) (cited on page 119).

[445] Alexander D'Amour et al. 'Underspecification presents challenges for credibility in modern machine learning'. In: *Journal of Machine Learning Research* (2020) (cited on page 120).

[446] Relja Arandjelovic and Andrew Zisserman. 'Look, listen and learn'. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 609–617 (cited on page 120).

[447] Bruno Korbar. 'Co-training of audio and video representations from self-supervised temporal synchronization'. In: (2018) (cited on page 120).

[448] Gianluigi Silvestri, Daan Roos, and Luca Ambrogioni. 'Closing the gap: Exact maximum likelihood training of generative autoencoders using invertible layers'. In: *arXiv preprint arXiv:2205.09546* (2022) (cited on page 120).

[449] David Ha and Jürgen Schmidhuber. 'World Models'. In: *arXiv e-prints* (2018), arXiv–1803 (cited on page 120).

[450] Danijar Hafner et al. 'Dream to Control: Learning Behaviors by Latent Imagination'. In: *International Conference on Learning Representations*. 2019 (cited on page 120).

[451] Kaiming He et al. 'Masked autoencoders are scalable vision learners'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16000–16009 (cited on page 121).

[452] Aaron Van Den Oord, Oriol Vinyals, et al. 'Neural discrete representation learning'. In: *Advances in neural information processing systems* 30 (2017) (cited on page 121).

[453] Richard D Lange et al. 'Neural Networks as Paths through the Space of Representations'. In: *arXiv preprint arXiv:2206.10999* (2022).

[454] Cian Eastwood, Ian Mason, and Christopher KI Williams. 'Unit-level surprise in neural networks'. In: *I (Still) Can't Believe It's Not Better! Workshop at NeurIPS 2021*. PMLR. 2022, pp. 33–40.

[455] Bruno A Olshausen and David J Field. 'Sparse coding with an overcomplete basis set: A strategy employed by V1?' In: *Vision research* 37.23 (1997), pp. 3311–3325.

[456] Bruno A Olshausen and David J Field. 'Emergence of simple-cell receptive field properties by learning a sparse code for natural images'. In: *Nature* 381.6583 (1996), pp. 607–609.

[457] Giambattista Parascandolo. *Backprop, evolution, and the myth of Two dogs*. June 2021. URL: https://gibipara92.github.io/2021/06/09/backprop-evolution-two-dogs.html.

[458] Artem Kaznatcheev and Konrad Paul Kording. 'Nothing makes sense in deep learning, except in the light of evolution'. In: *arXiv preprint arXiv:2205.10320* (2022).

[459] Hiroshi Morioka, Vince Calhoun, and Aapo Hyvärinen. 'Nonlinear ICA of fMRI reveals primitive temporal structures linked to rest, task, and behavioral traits'. In: *NeuroImage* 218 (2020), p. 116989.

[460] Hubert Banville et al. 'Uncovering the structure of clinical EEG signals with self-supervised learning'. In: *Journal of Neural Engineering* 18.4 (2021), p. 046020.

[461] Imant Daunhawer et al. 'On the Limitations of Multimodal VAEs'. In: *arXiv preprint arXiv:2110.04121* (2021).

[462] Adrián Javaloy, Maryam Meghdadi, and Isabel Valera. 'Mitigating Modality Collapse in Multimodal VAEs via Impartial Optimization'. In: *arXiv preprint arXiv:2206.04496* (2022).

[463] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. 'The frontier of simulation-based inference'. In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30055–30062.

[464] Andreas Krämer, Jonas Köhler, and Frank Noé. 'Training invertible linear layers through rank-one perturbations'. In: *arXiv preprint arXiv:2010.07033* (2020).

[465] Ben Poole et al. 'Dreamfusion: Text-to-3d using 2d diffusion'. In: *arXiv preprint arXiv:2209.14988* (2022).

[466] Behrooz Ghorbani, Hamid Javadi, and Andrea Montanari. 'An instability in variational inference for topic models'. In: *International conference on machine learning*. PMLR. 2019, pp. 2221–2231.

[467] John P Barton et al. 'Large pseudocounts and l 2-norm penalties are necessary for the mean-field inference of Ising and Potts models'. In: *Physical Review E* 90.1 (2014), p. 012132.

[468] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[469] Jonathan S Yedidia, William T Freeman, and Yair Weiss. 'Understanding belief propagation and its generalizations'. In: *Exploring artificial intelligence in the new millennium*. 2003, pp. 239–269.

[470] Luca Peliti. 'Statistical mechanics in a nutshell'. In: *Statistical Mechanics in a Nutshell*. Princeton University Press, 2011.

[471] Paul Rubenstein et al. 'Practical and Consistent Estimation of f-Divergences'. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

[472] Ben Baker, Benjamin Lansdell, and Konrad P Kording. 'Three aspects of representation in neuroscience'. In: *Trends in Cognitive Sciences* (2022).

[473] James CR Whittington et al. 'Disentangling with Biological Constraints: A Theory of Functional Cell Types'. In: *arXiv preprint arXiv:2210.01768* (2022).

[474] Irina Higgins et al. 'Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons'. In: *Nature communications* 12.1 (2021), pp. 1–14.

[475] Finbarr O'Sullivan. 'A statistical perspective on ill-posed inverse problems'. In: *Statistical science* (1986), pp. 502–518.

[476] Nelson Elhage et al. 'Toy Models of Superposition'. In: *Transformer Circuits Thread* (2022).

[477] Jin Tian and Judea Pearl. *A general identification condition for causal effects*. eScholarship, University of California, 2002.

[478] Judea Pearl. 'The logic of counterfactuals in causal inference'. In: (2011).

[479] A Philip Dawid. 'Causal inference without counterfactuals'. In: *Journal of the American statistical Association* 95.450 (2000), pp. 407–424.

[480] Kurt Koffka. *Principles of Gestalt psychology*. Routledge, 2013.

[481] Gaetano Kanizsa. *Grammatica del vedere: Saggi su percezione e gestalt*. Società editrice il Mulino, 1980.

[482] Gaetano Kanizsa. 'Subjective contours'. In: *Scientific American* 234.4 (1976), pp. 48–53.

[483] Geoffrey E Hinton. 'Connectionist learning procedures'. In: *Machine learning*. Elsevier, 1990, pp. 555–610.

[484] Geoffrey E Hinton. 'Preface to the special issue on connectionist symbol processing'. In: *Artificial Intelligence* 46.1-2 (1990), pp. 1–4.

[485] David S Touretzky and Geoffrey E Hinton. 'Symbols among the neurons: Details of a connectionist inference architecture'. In: *IJCAI*. Vol. 85. 1985, pp. 238–243.

[486] Gary Marcus. 'The next decade in ai: four steps towards robust artificial intelligence'. In: *arXiv preprint arXiv:2002.06177* (2020).

[487] Kareem Ahmed et al. 'Semantic Probabilistic Layers for Neuro-Symbolic Learning'. In: *arXiv preprint arXiv:2206.00426* (2022).

[488] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*. Vol. 191. American Mathematical Soc., 2007 (cited on page 128).

[489] Shun-ichi Amari. 'Information geometry'. In: *Japanese Journal of Mathematics* 16.1 (2021), pp. 1–48 (cited on page 128).

[490] Dominik Janzing et al. 'Justifying information-geometric causal inference'. In: *Measures of complexity*. Springer, 2015, pp. 253–265 (cited on page 129).

[491] Alexander N Gorban and Ivan Yu Tyukin. 'Blessing of dimensionality: mathematical foundations of the statistical physics of data'. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2118 (2018), p. 20170237 (cited on page 130).

[492] Manfred R Schroeder. 'Listening with two ears'. In: *Music perception* 10.3 (1993), pp. 255–280 (cited on page 135).

[493] Danilo Jimenez Rezende. *Short Notes on Divergence Measures*. 2018 (cited on page 138).

[494] Tom Hennigan et al. *Haiku: Sonnet for JAX*. Version 0.0.3. 2020 (cited on page 147).

[495] Pauli Virtanen et al. 'SciPy 1.0: fundamental algorithms for scientific computing in Python'. In: *Nature methods* 17.3 (2020), pp. 261–272 (cited on page 147).

[496] James R Smart. *Modern geometries*. Brooks/Cole Pacific Grove, CA, 1998 (cited on page 155).

[497] Mirjam Soeten. 'Conformal maps and the theorem of Liouville'. PhD thesis. Faculty of Science and Engineering, 2011 (cited on page 155).

[498] Ruy Tojeiro. 'Liouville's theorem revisited'. In: *Enseignement Mathematique* 53.1/2 (2007), p. 67 (cited on page 156).

[499] Bin Dai and David Wipf. 'Diagnosing and Enhancing VAE Models'. In: *International Conference on Learning Representations*. 2018 (cited on page 161).

[500] Partha Ghosh et al. 'From Variational to Deterministic Autoencoders'. In: *International Conference on Learning Representations*. 2019 (cited on pages 161, 183).

[501] Jerrold E Marsden and Anthony Tromba. *Vector calculus*. Macmillan, 2003 (cited on page 180).

[502] Michael E Tipping and Christopher M Bishop. 'Probabilistic principal component analysis'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3 (1999), pp. 611–622 (cited on page 182).

[503] Abhishek Kumar, Ben Poole, and Kevin Murphy. 'Regularized autoencoders via relaxed injective probability flow'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 4292–4301 (cited on page 183).

[504] Lutz Prechelt. 'Early stopping-but when?' In: *Neural Networks: Tricks of the trade*. Springer, 1998, pp. 55–69 (cited on pages 186, 188).

[505] Raúl Rojas. *Neural networks: a systematic introduction*. Springer Science & Business Media, 2013 (cited on pages 191, 200).

[506] Emiel Hoogeboom, Rianne van den Berg, and Max Welling. 'Emerging convolutions for generative normalizing flows'. In: *arXiv preprint arXiv:1901.11137* (2019) (cited on pages 193, 194).

[507] Jakub M Tomczak and Max Welling. 'Improving variational auto-encoders using householder flow'. In: *arXiv preprint arXiv:1611.09630* (2016) (cited on page 193).

[508] Rianne van den Berg et al. 'Sylvester normalizing flows for variational inference'. In: *arXiv preprint arXiv:1803.05649* (2018) (cited on page 194).

[509] Mahdi Karami et al. 'Invertible Convolutional Flow'. In: *Advances in Neural Information Processing Systems*. 2019, pp. 5636–5646 (cited on page 194).

[510] Marc Finzi et al. 'Invertible Convolutional Networks'. In: 2019 (cited on pages 194, 201).

[511] Leemon Baird, David Smalenberger, and Shawn Ingkiriwang. 'One-step neural network inversion with PDF learning and emulation'. In: *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* Vol. 2. IEEE. 2005, pp. 966–971 (cited on page 194).

[512] Wikipedia. *Computational complexity of mathematical operations — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Computational%20complexity%20of%20mathematical%20operations&oldid=958179308. [Online; accessed 11-June-2020]. 2020 (cited on page 194).

[513] Léon Bottou. 'Large-scale machine learning with stochastic gradient descent'. In: *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186 (cited on page 196).

[514] Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. Vol. 105. Siam, 2008 (cited on page 196).

[515] Charles C Margossian. 'A review of automatic differentiation and its efficient implementation'. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.4 (2019), e1305 (cited on page 197).

[516] Moshe Leshno et al. 'Multilayer feedforward networks with a nonpolynomial activation function can approximate any function'. In: *Neural networks* 6.6 (1993), pp. 861–867 (cited on page 199).

[517] Jianxin Wu. 'Introduction to convolutional neural networks'. In: (2017) (cited on page 201).

[518] Roy R Lederman and Ronen Talmon. 'Learning the geometry of common latent variables using alternating-diffusion'. In: *Applied and Computational Harmonic Analysis* 44.3 (2018), pp. 509–536 (cited on page 220).

[519] Le Song et al. 'Nonparametric estimation of multi-view latent variable models'. In: *International Conference on Machine Learning*. 2014, pp. 640–648 (cited on page 220).

[520] Animashree Anandkumar et al. 'Tensor decompositions for learning latent variable models'. In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 2773–2832 (cited on page 220).

[521] Virginia R De Sa. 'Spectral clustering with two views'. In: *ICML workshop on learning with multiple views*. 2005, pp. 20–27 (cited on page 220).

[522] Abhishek Kumar, Piyush Rai, and Hal Daume. 'Co-regularized multi-view spectral clustering'. In: *Advances in neural information processing systems*. 2011, pp. 1413–1421 (cited on page 220).

[523] Russell A Poldrack et al. 'Toward open sharing of task-based fMRI data: the OpenfMRI project'. In: *Frontiers in neuroinformatics* 7 (2013), p. 12 (cited on page 227).

[524] Shinji Nishimoto et al. 'Reconstructing visual experiences from brain activity evoked by natural movies'. In: *Current biology* 21.19 (2011), pp. 1641–1646 (cited on page 227).

[525] Alexander G Huth et al. 'A continuous semantic space describes the representation of thousands of object and action categories across the human brain'. In: *Neuron* 76.6 (2012), pp. 1210–1224 (cited on page 227).

[526] Edwin T Jaynes. 'On the rationale of maximum-entropy methods'. In: *Proceedings of the IEEE* 70.9 (1982), pp. 939–952 (cited on page 244).

[527] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation. Blender Institute, Amsterdam, 2021 (cited on page 246).

[528] Diederik P. Kingma and Jimmy Ba. 'Adam: A Method for Stochastic Optimization'. In: *3rd International Conference on Learning Representations*. 2015 (cited on page 257).