3

An application of the EM algorithm to archaeological data analysis

W. Andrew Scott*

Simon W. Hillson†

3.1 Introduction

The association between archaeology and statistics has a very long history. Simple descriptive statistics are found in the earliest excavation reports and collaboration between archaeologists and statisticians has taken place for almost as long. It seems strange therefore that despite such an auspicious start, only comparatively recently has there been any major use of statistics by archaeologists. Only in the early 1970's did the use of statistics really take off and even today use of the more powerful multivariate techniques is both relatively infrequent and tends to be confined to specific areas of archaeology.

There are many possible reasons for this. One suggestion is that, due to its historic development as one of the humanities, there is an inbuilt resistance to mathematics in archaeology. Another is that archaeological training doesn't place sufficient emphasis on quantitative methods. While a certain amount of support for both of these views can be found in the literature, it seems more likely that the answer is a simpler one—namely that statistics, or rather statistical computing, doesn't offer archaeology the facilities it actually needs. The upsurge of interest in statistics in the 1970's was generated by the increasing availability of computing power and took place in many other disciplines as well as archaeology. It serves to illustrate the fact that archaeologists, far from being resistant to change, are ready to adopt new techniques when these prove to be useful. Computing has been accepted so completely that today many excavations even have microcomputers on site. The reason statistical analysis has lagged behind is, we believe, due to the nature of archaeological data. The complex structure of many archaeological objects, skulls, dentition even ceramics, can only adequately be described by multiple measurements and hence requires the use of multivariate statistical methods. However, in their standard form, all such methods require complete data sets and this is very rarely the case in real life. The ensuing practical difficulties with analysis are common

* Centre for Applied Statistics,
Lancaster University
† Institute of Archaeology,
31–34 Gordon Square,
London

Variable		Missing values	
		No.	%
maximum cranial length	L	0	0.0
maximum breadth	B	1	0.8
basibregmatic height	H'	4	3.3
upper facial height	G'H	3	2.4
bimaxillary breadth	GB	7	5.8
basialveolar length	GL	5	4.1
nasal height	NH'	3	2.4
nasal breadth	NB	2	1.7
bidacryonic chord	DC	16	13.2
palatal length	G'1	13	10.7
palatal breadth	G2	22	18.2
orbital breadth	O'1	11	9.1
orbital height	O2	0	0.0
sex		48	39.7

Table 3.1

to many subjects but are particularly acute in archaeology where material is often damaged or poorly preserved. The result is often the use of simpler but less appropriate techniques. In this paper the problems involved are illustrated by the application of discriminant analysis to a small archaeological data set.

3.2 Material

The data used in this investigation comprised a set of thirteen measurements on a total of 121 human skulls from two Egyptian cemeteries: Badari, a predynastic site from Upper Egypt, and Sedment, a IXth Dynasty site from Middle Egypt. The data were originally published by Stoessiger 1927 and Woo 1930 and the variables used here are a selection from over seventy recorded for each collection. Only skulls from adult individuals have been included. Apart from sex differences the data show little evidence of grouping within collections. They therefore make useful homogeneous collections for experiments with statistical methods. All the skulls had been sexed for the original publications, although 27 males and 21 females were marked as questionable. Further examination by one of the authors for a previous study (Hillson 1978) confirmed the majority of the original decisions but reclassified five of the questionable female skulls from Sedment as male. For the purposes of this paper, however, these 48 skulls are treated as being of unknown sex.

Table 3.1 lists the variables used, together with the number of values missing for each. For the continuous variables this ranges from 0% to almost 20% with an average of 5%

Table 3.2 shows the pattern of missing values within cases. Even if sex were known over 30% of the cases would be incomplete. When sex is included this proportion rises to almost 60%.

Apart from the missing information this can be regarded statistically as a 'nice' dataset with no apparent peculiarities. Since discriminant analysis is an accepted

	Missing values per case								
	0	1	2	3	4	5	6	7	8
a) excluding sex									
No.	83	1	9 6	6	3	1	1	1	1
%	68.6	15.7	5.0	5.0	2.5	0.8	0.8	0.8	0.8
b) all variables									
No.	49	46	9	8	5	1	1	0	2
%	40.5	38.0	7.4	6.6	4.1	0.8	0.8	0.0	1.7

Table 3.2

technique with a long history it might be expected that performing the analysis with standard software should be straightforward. Unfortunately, however, this is not the case.

3.3 Treatment of missing data

The problem of missing data has received much attention in the statistical literature and solutions for certain special cases have existed for many years. In more general situations however, no completely satisfactory treatment has been available, although a number of arbitrary procedures have traditionally been used to overcome the problem.

3.3.1 Removal of cases and/or variables

This is the oldest and most popular solution. Cases with missing values are dropped to leave a complete data set which is readily analysed. If missing values are particularly common for some variables it may be more economic to drop these before removing cases. This process is not very helpful in an archaeological context. Of the 121 skulls in this study 72 had missing values so that over half would be eliminated. Yet 46 of these had only one variable missing (but not the same variable in each case). If these skulls were removed the perfectly good values for the remaining thirteen variables would also be lost. This procedure can also introduce bias into the results. For example, less heavily built skulls are more prone to damage and consequently are more likely to be removed. If the average degree of damage differs markedly with location, then skulls from one site could appear more robust than those from another simply because the fragile examples have been taken out of the analysis.

3.3.2 Use of only those measurements which are present

The calculations needed for many complex statistical methods can often be couched in terms of univariate or bivariate statistics. For example discriminant analysis can be performed using just the within-group means and covariances of the measurements and each of these statistics can be estimated using only one or two variables at a time. So this process will utilise a higher proportion of the available information than the previous one. However it can lead to anomalous or inconsistent results particularly when missing values are not missing at random.

3.3.3 Replacement of missing values

A third alternative is to fill in the data set by replacing missing values with some reasonable estimate. The most commonly used replacement is the arithmetic mean. With discriminant analysis the within-group mean may be used. Whilst the mean may be a good representation of a group, individual skulls can vary considerably from it. Replacement by a mean thus diminishes the effect of normal variation in size and shape and leads to an exaggerated impression of group differences. This effect can be partially alleviated by using multivariate regression to produce replacement values. However variation in the data can still be considerably underestimated when the proportion of missing values is large.

3.4 The EM algorithm

For the last decade a new efficient statistical method for coping with missing data problems, known as the expectation-maximisation or EM algorithm, has been available. The mathematical basis of this approach, also referred to as self-consistency (Efron 1967) or the missing information principle (Orchard & Woodbury 1972), was first presented in its most general form by Dempster *et al.* 1977, although special applications of the algorithm had been derived and used many times before this. The algorithm is a means of obtaining maximum-likelihood parameter estimates for incomplete data by using the conditional expectation of the likelihood of the unobserved complete data. This leads to an iterative process which successively alters parameter estimates until convergence occurs. This method of dealing with missing information has several advantages. It does not underestimate the variation in the data, it utilises all the available information and it avoids the possibility of inconsistent estimates. It can also reduce bias and can cope with both continuous and discrete missing values. The disadvantage is that the algorithm is computationally expensive, both in terms of storage and time. Specific implementations of the algorithm can be highly complex, but for many common statistical techniques it reduces to a relatively straightforward computational routine. Little & Schlucter 1985 consider the case where the data are a mixture of discrete variables and normal continuous variables. They give a detailed formulation of the algorithm which can be adapted to perform a variety of statistical analyses, including a form of discriminant analysis for mixed categorical and continuous data due to Krzanowski 1980.

3.5 Available computer facilities

In practice the first of these procedures, deletion of cases, has usually been the standard method of dealing with missing data in statistical computer programs although some (e.g. SPSS) also allowed the second, use of available measurements. In recent years a few major packages have implemented procedures for replacing missing values by estimates, but none have yet introduced methods based on the EM algorithm. Discriminant analysis techniques are also particularly poorly covered in standard packages. Of the five general purpose packages supported at Lancaster, two, MINITAB and GENSTAT, provide no discriminant procedures at all, although it would of course be possible to program the techniques in GENSTAT. The remaining three, SAS, SPSS and BMDP, do provide facilities for discriminant analysis, but only the standard linear and quadratic forms. The newer efficient methods of dealing with mixed discrete and continuous variables are not implemented. Both SAS and SPSS restrict analysis to complete cases although SPSS allows mean substitution in the classification stage. Only BMDP provides a reasonable range of missing value treatments, but to employ these requires the use of a completely separate program prior to the discriminant program. None of these packages can deal with missing discrete data except by pretending that it is continuous. A 'correct' analysis of the present data would therefore be difficult, if not impossible, with any of these programs without extensive programming and statistical expertise. Over the last two years a new statistical package has been under development at Lancaster which utilises the methods of Little and Schlucter to provide those statistical techniques most useful to archaeology, even when the data is incomplete. This has

a) Sex unknown

		actual site		
		Badari	*Sedment*	*Total*
predicted	Badari	16	2	18
site	Sedment	0	30	30
	Total	16	32	48

95.8% correctly classified

b) Male

		actual site		
		Badari	*Sedment*	*Total*
predicted	Badari	18	1	19
site	Sedment	2	24	26
	Total	20	25	45

93.3% correctly classified

c) Female

		actual site		
		Badari	*Sedment*	*Total*
predicted	Badari	14	0	14
site	Sedment	3	11	14
	Total	17	11	28

89.3% correctly classified

d) All cases

		actual site		
		Badari	*Sedment*	*Total*
predicted	Badari	48	3	51
site	Sedment	5	65	70
	Total	53	68	121

93.4% correctly classified

Table 3.3: Comparison of predicted and actual site

been used to analyse the present data.

3.6 Results

Table 3.3 shows the classification results table obtained from the analysis. Overall 93% of the cases were correctly assigned to site.

Previous results from the analysis of this data with all skulls coded for sex (Hillson 1985) gave overall misclassification rates of 20.7% using mean substitution and 20.3% using a method designed to utilise as much information as possible. Use of the EM

algorithm has therefore produced a considerable improvement in this case.

Of course these misclassification rates are all underestimated since the same data was used in both the analysis and classification stages. In order to obtain a more reliable estimate of predictive ability the data was divided in two and the discriminant functions from the separate analyses of each half were used to classify the other half. Table 3.4 gives the results. The overall misclassification rate has now increased to 16.5%. Females appear to be more accurately classified than males as were skulls from Sedment compared to Badari.

Interestingly lack of knowledge of sex does not seem to affect the misclassification rate. Many of the continuous variables showed strong sex differences and it is likely that these stood proxy for the sex effect when sex was unknown.

As a byproduct of the analysis it is possible to obtain predictions of the sex of those skulls with sex coded unknown. Table 3.5 shows these predicted values compared to those originally given. They verify the proposed changes for a number of Sedment females but also suggest the possible misclassification of a number of Badari females as male

3.7 Conclusions

The main theoretical problems preventing the efficient analysis of incomplete data have now been solved and the methodology appears to work well in practice. Statisticians have been using these new techniques for the last ten years but unfortunately standard software has yet to implement them in such a way that they can be readily employed by users without considerable statistical and computing knowledge. Completion of the statistical package currently under development at Lancaster, designed primarily for archaeologists, should help to remedy this.

References

DEMPSTER, A. P., N. M. LAIRD, & D. E. RUBIN 1977. "Maximum likelihood from incomplete data via the EM algorithm", *J. R. Statist. Soc. B.*, 39: 5–38.

EFRON, B. 1967. "The two sample problem with censored data", *Proc. 5th Berkeley Symp.*, 4: 831–853.

HILLSON, S. W. 1978. *Human biological variation in the Nile valley.* PhD thesis, London University.

HILLSON, S. W. 1985. "Missing information and collections of skeletal material". *in* Fieller, N. R. J. *et al.*, (eds.), *Palaeoenvironmental Investigations,* International Series. B. A. R., Oxford.

KRZANOWSKI, W. J. 1980. "Mixtures of continuous and categorical variables in discriminant analysis", *Biometrics,* 36: 493–499.

LITTLE, R. J. A. & M. D. SCHLUCTER 1985. "Maximum likelihood estimation for mixed continuous and categorical data with missing values", *Biometrika,* 72: 497–512.

ORCHARD, T. & M. A. WOODBURY 1972. "A missing information principle: theory and applications", *Proc. 6th Berkeley Symp.*, 1: 697–715.

a) Sex unknown

		actual site		
		Badari	*Sedment*	*Total*
predicted	Badari	15	4	19
site	Sedment	1	28	29
	Total	16	32	48

89.6% correctly classified

b) Male

		actual site		
		Badari	*Sedment*	*Total*
predicted	Badari	13	4	17
site	Sedment	7	21	28
	Total	20	25	45

75.6% correctly classified

c) Female

		actual site		
		Badari	*Sedment*	*Total*
predicted	Badari	13	0	13
site	Sedment	4	11	15
	Total	17	11	28

85.7% correctly classified

d) All cases

		actual site		
		Badari	*Sedment*	*Total*
predicted	Badari	41	8	49
site	Sedment	12	60	72
	Total	53	68	121

83.5% correctly classified

Table 3.4: Comparison of predicted and actual site

a) Badari

		original category		
		Male	*Female*	*Total*
predicted category	Male	4	1	5
	Female	9	2	11
	Total	13	3	16

b) Sedment

		original category		
		Male	Female	Total
predicted category	Male	12	7	19
	Female	2	11	13
	Total	14	18	32

Table 3.5: Prediction of sex

STOESSIGER, B. N. 1927. "A study of the Badarian crania recently excavated by the British School of Archaeology in Egypt", *Biometrika*, 22: 65–83.

WOO, T. L. 1930. "A study of seventy-one 9th Dynasty skulls from Sedment", *Biometrika*, 22: 65–83.