34 A parametric approach to seriation

Andy Scott

34.1 INTRODUCTION

Seriation is the oldest and probably the most used statistical technique in archaeology although its importance has greatly declined since the introduction of modern dating methods. A wide range of approaches to the solution of the seriation problem have been suggested. Many methods start from a matrix of measures of correlation or similarity between assemblages. These can be manipulated or permutated directly. Alternatively most multivariate statistical techniques can be used to produce unidimensional orderings of archaeological material and, theoretically at least, such sequences may well represent time. There are, however, a number of drawbacks to all these approaches which makes their use questionable and throws doubt on any conclusions obtained:

1) Choice of a similarity coefficient is usually an entirely arbitary procedure and frequently little or no consideration is given to the appropriateness of the chosen coefficient or the degree to which it effectively represents closeness in time.
2) The unit of analysis is invariably the assemblage rather than the artifact. Consequently each assemblage is given equal weight in the analysis regardless of its size or the nature of its constituents. Since larger assemblages contain more information, or at least may convey qualitatively different information, a more natural procedure would be to treat the artifact as the unit of analysis. Essentially the data has a hierarchical structure with two levels, yet this is ignored in all current seriation methods.
3) All current techniques are used just to provide an ordering of archaeological material, although some, for example multidimensional scaling, could be used to scale instead. Estimates of the accuracy of the positioning of any individual assemblage within a sequence are not provided. Thus there is no means of judging the accuracy of a derived sequence.

In addition to these difficulties there is another aspect of seriation which is usually ignored in descriptions of the technique. It may be that a chronological ordering for a given set of material is desired solely in order to clarify events for the particular site from which the material came. Sometimes however, once a sequence is formed, it is used as a means of classifying other material. The most obvious example here is Petrie's Egyptian sequence which is still in use today after nearly a century. The question then arises, as a problem distinct from the original seriation problem, as to the best way of matching extraneous material to an already created sequence. At present this would usually be done by hand since none of the existing techniques deals with the problem.

In one sense seriation can be considered as being a missing data problem since its primary purpose is to estimate the value of a variable, time, which is completely missing. Thus statistical methods for dealing with incomplete data should form an appropriate approach to seriation. In this paper, incomplete data techniques are used to develop a parametric method of seriation which goes some way towards overcoming the drawbacks associated with standard methods. It should be pointed out that this method is not intended to replace standard methodology but rather to complement it. In many instances the necessary distributional assumptions involved in parametric modelling will not be appropriate for the archaeological material involved. A parametric form of seriation was previously developed by Kendall (1963) based on an analysis of the methods of Petrie (1899) but this was limited to the study of artifacts classified into distinct varieties and made no allowance for the possibility of utilising continuous measurements.

34.2 A PARAMETRIC SERIATION MODEL

In order to develop parametric seriation methods it is necessary to derive a distributional model for the variation within and between assemblages. Suppose that the artifacts making up the assemblages can be divided into W distinct types and that for each type a set of descriptive measurements, continuous and discrete, is made. Denote this set of variables by x and the age or date of an artifact by t. The variable vector x is allowed to differ in composition and dimension with the type of artifact. The measurements used to describe pottery urns, for example, being unrelated to those needed to adequately describe handaxes or broaches. Represent the distribution of these variables for type w by $f(t,x|w)$, where w is a vector of W binary 0,1 elements each element indexing a particular artifact type, and suppose that the overall prevalence of type w is given by $p(w)$. Thus the overall distribution of artifacts is given by

$$[1] \qquad f(w,x,t) = p(w)f(t,x|w)$$

This distribution can be thought of as in some sense representing an imaginary infinite population of artifacts from which assemblages are constructed. In some cases it might be possible to identify the distribution with the population of artifacts existing over the time span under consideration. In most situations, however, the problems of site and artifact preservation make such identification dangerous.

Consider now the "statistical" process by which assemblages are created. For a set of N assemblages let the date or age, t_s, of assemblage s be randomly selected from some distribution $g(t)$. Similarly let the number of artifacts, n_s, in assemblage s be randomly selected from the conditional distribution $h(n_s|t_s)$ and let each artifact and its associated measurements be randomly chosen from the conditional distribution $f(w_{si},x_{si}|t_s), i=1\ldots n_s$, where

$$[2] \qquad f(w_{si},x_{si}|t_s) = f(w_{si},x_{si},t_s)/f(t_s)$$

and $f(t_s)$ is the marginal distribution of t_s obtained from $f(w_{si},x_{si},t_s)$. The likelihood of an observed set of N assemblages is therefore

$$[3] \qquad l = \prod_s \left\{ g(t_s)h(n_s|t_s)\prod_i f(w_{si},x_{si}|t_s) \right\} \quad s=1,N,\ i=1,n_s$$

This is a very general expression which, through suitable choice and interpretation of the component densities, can be used to represent a wide range of seriation data and the processes by which they are obtained. Suitable choices for the density forms are now considered briefly before the development of a specific model.

In most archaeological discussions of the seriation problem, it is assumed that each artifact type gradually increases in prevalence until it reaches a peak and then declines. This unimodal distribution over time has been thought of as applying to the proportion of assemblages containing the specified type of artifact, to the number of such artifacts within assemblages, or to both. Manipulation of assemblage order to achieve such "battleship" shaped curves formed the basis of an early form of seriation (Ford, 1962). A natural choice for $f(t|w)$ is therefore the normal distribution. This choice is open to criticism on the basis that the distribution of particular types of artifacts over time may well be skewed rather than symmetrical. Kendall (1963) was careful to allow for this in his analysis. The normal distribution, however, has two advantages. Analytically it is more tractable than most other candidate distributions, and its multivariate form can be used to encompass any continuous measurements made on the artifacts. Suppose therefore, that the descriptive variables x are all continuous and that the distribution of x and t for each artifact type, $f(x,t|w)$, is multivariate normal with mean of $x=\mu_w$, mean of $t=\tau_w$, variance of $x=Y_w$, variance of $t=\sigma_w^2$ and covariance of x and $t=\Delta_w$. τ_w can be thought of as the time at which artifact type w was most prevalent and σ_w^2 as measuring the range of time over which it was used. If x does in fact contain discrete variables these can be taken as defining new types of artifact and hence absorbed into the variable w.

The normal distribution is also a natural choice for $g(t)$, since it implies that the set of assemblages cluster round some central date and become less frequent as the time away from that point increases. An alternative is to use $f(t)$, the marginal distribution of t obtained from $f(w,x,t)$, and this is the choice used here. If the average number of artifacts in an assemblage is independent of time then, ignoring the obvious problems of oversimplistic interpretation, use of $f(t)$ implies that $f(w,x,t)$ can be interpreted as the distribution of artifacts within the population of assemblages as well as the population of artifacts available for incorporation within assemblages.

The distribution of the number of artifacts in an assemblage, $h(n|t)$, can be modelled in a variety of ways. One of the most convenient is to assume a Poisson distribution. Two alternatives

suggest themselves for the mean of this distribution. The first is to assume that the mean is proportional to $f(t)$. This is equivalent to assuming that the average number of artifacts in an assemblage is proportional to the number of artifacts available at any given time. The second and simplest alternative is to assume that the mean is independent of time i.e. to use $h(n)$ instead of $h(n|t)$. Since assemblages containing no artifacts are in general not included in seriation data sets, it will usually be more appropriate to take $h(n\text{-}1)$ as Poisson rather than $h(n)$. Suppose therfore that $h(n\text{-}1)$ is Poisson with mean m.

34.2.1 Model fitting

Once an appropriate choice of distributions is made the likelihood given above is fully defined. With the options chosen above

$$[4] \quad l = \prod_s \left\{ f(t_s) h(n_s - 1) \prod_i f(t_s, x_{si} | w_{si}) p(w_{si}) / f(t_s) \right\} \quad s = 1, N, i = 1, n_s$$

where w_{si} is the type of the ith artifact in assemblage s and x_{si} is the associated vector of continuous measurements. The expressions w_s and x_s will be used to represent the complete set of artifact types and measurements within assemblage s.

If t were known for each assemblage then the parameters of the various distributions could in theory be found by maximising l or its log. Equally if the parameters were known then they could be used to predict the values of t. Since both t and the parameters are in general completely unknown, neither of these procedures is possible, but an alternative approach known as the EM algorithm can be used (Dempster *et al.* 1977). Instead of maximising the complete log–likelihood, the EM approach is to maximise the expected value of the log–likelihood given the observed information. In many situations, the algorithm reduces to a simple two stage iterative process. Starting with an initial guess at the parameter values, the algorithm would use these to estimate the assemblage ages (more accurately the sufficient statistics of the parametric model) and then use the estimated ages to give improved values for the parameter estimates. This process is repeated until the values no longer change, at which point the algorithm has converged and we have estimates of both the parameters and the assemblage ages. In the present case, the expected value of the log–likelihood is not analytically tractable so that implementing the algorithm, while possible, is computationally difficult.

One final problem remains, that of "anchoring" the time distribution. If t is completely unobserved, then the parameters τ_w and σ_w^2 are not strictly speaking identifiable. Similar problems are encountered in factor analysis and other latent variable techniques, where it is often assumed that the latent variables have zero mean, variance 1 and no correlation. In the present situation a variety of ways can be used to solve this problem. If the age of some assemblages are known, possibly through other dating methods, then the known dates can be included in the data and serve to fix the time distribution. Alternatively the mean age and range of a specific artifact type may be known from other studies and again this serves to fix the time distribution. When t is completely unobserved, other methods must be used. Two dissimilar assemblages can be chosen and given arbitary dates such as 0 and 1. Alternatively the mean age and variance of one chosen artifact type can be fixed at 0 and 1. In both these cases, the resulting parameter estimates will be relative to the chosen values. A third alternative is to fix the overall mean and variance of t in $f(t)$ at suitable values and then to maximise l subject to these constraints.

Once the model has been fitted to a data set, the estimated parameters can be used to provide an ordering of the assemblages and hence a solution to the seriation problem. The conditional distribution of t_s, $f(t_s|w_s, x_s)$, can be used to obtain the variance of t_s for each assemblage as well as its expected value. Hence, unlike standard seriation methods, the proposed technique provides an indication of the accuracy and degree of overlap of the predicted sequence. The problem of dating or sequencing additional material is also solved by this method since estimates can be provided in the same way for assemblages that were not included in the analysis.

34.2.2 Model simplification

The seriation procedure proposed in the previous section is, in practice, computationally expensive since it involves a considerable amount of numerical integration and maximisation. In the present section a simpler method that approximates to this procedure while avoiding numerical integration is considered. Examination of the relevant expressions shows that the need for numerical integration arises from two sources. Firstly and most importantly is the need to form the expected value of the log-likelihood with respect to the "awkward" conditional distribution $f(t_s|x_s, w_s)$. Secondly is the presence of a term involving $\ln\{f(t_s)\}$ in the expression for the ex-

pected value of the log-likelihood. This term leads directly to the analytically intractable components in the solution to the likelihood equations. This aspect of the problem is most easily dealt with, by omitting the term from the likelihood. This would appear to be a rather drastic step but can be justified to some extent since the intractable components are all expected values of weighted averages of quantities that have expectation zero with respect to the population of artifacts as a whole. Thus in large samples they would be expected to be small anyway. For complete data (i.e. for known t) the parameter estimates from such a reduced likelihood are in fact consistent although not maximum likelihood (ML). In effect therefore omitting this term from the likelihood converts the EM algorithm from ML estimation to the estimation of a different consistent estimator. Alternatively the reduced likelihood can be thought of as corresponding to a model where each artifact is chosen separately from $f(t,x,w)$ but with certain groups of artifacts known to have the same value of t. The parameter estimates are now

$$
\begin{aligned}
m &= \textstyle\sum_s (n_s - 1)/N \\
p(w) &= \textstyle\sum_s n_{sw} / \sum_s n_s \\
\mu_w &= \textstyle\sum_s \sum_{i=w} x_{si} / n_w \\
\Upsilon_w &= \textstyle\sum_s \sum_{i=w} (x_{si} - \mu_w)(x_{si} - \mu_w)^T \\
\tau_w &= \textstyle\sum_s n_{sw} E(t_s) / n_w \\
\sigma_w^2 &= \textstyle\sum_s n_{sw} E(t_s - \tau_w)^2 / \sum_s n_{sw} \\
\Delta_w &= \textstyle\sum_s \sum_{i=w} (x_{si} - \mu_w)(E(t_s) - \tau_w)
\end{aligned}
\qquad [5]
$$

where n_{sw} is the number of artifacts of type w in assemblage s and n_w is the total number of artifacts of type w.

These estimates are simple to calculate apart from the last three expressions which involve the expectation of t_s and t_s^2 with respect to the conditional distribution of t_s. Analytic simplification of these expectations is not feasible but an approximation to them is. The marginal distribution of t over the population of artifacts, $f(t)$, is a mixture of normal distributions with overall mean

$$\tau = \textstyle\sum_w p(w) \tau_w \qquad [6]$$

and variance

$$\sigma^2 = \textstyle\sum_w p(w)(\sigma_w^2 + \tau_w^2) - \tau^2 \qquad [7]$$

Approximating $f(t)$ by a normal distribution with this mean and variance gives

$$f(t_2 | x_s, w_s, n_s) \approx N(M_s, S_s^2) \qquad [8]$$

where

$$M_s = \left[\frac{\tau(1-n_s)}{\sigma^2} + \sum_i \frac{\tau_i + \Delta_i^T \Omega_i^{-1} (x_{si} - \mu_i)}{\sigma_i^2 - \Delta_i^T \Omega_i^{-1} \Delta_i} \right] S_s^2 \qquad [9]$$

and

$$S_s^2 = \left[\frac{(1-n_s)}{\sigma^2} + \sum_i \frac{1}{\sigma_i^2 - \Delta_i^T \Omega_i^{-1} \Delta_i} \right]^{-1} \qquad [10]$$

so that

$$E(t_s) \approx M_s \quad \text{and} \quad E(t_s^2) \approx S_s^2 + M_s^2 \qquad [11]$$

Thus use of a normal approximation to $f(t)$ results in a very simple seriation technique in the form of an EM type procedure which alternates between 5 and 11 until convergence is reached.

The approximations made in deriving these expressions are fairly crude. However some justification for their use can be made. Empirically the procedure can be justified on the grounds that, as shown in the following section, it appears to work, producing reasonable parameter estimates and accurate seriation sequences. Theoretically further investigation of their adequacy is required.

The number of model parameters is proportional to the number of artifact types and as this increases, the instability of the estimative process can increase markedly. This is largely due to the proliferation of variance parameters. The problem can be prevented by dating sufficient assemblages to ensure that each artifact type has at least two different observed dates. An alternative is to use a model of constant within-type variance. Such a model, as well as solving the problem, is also likely to be more robust in general.

A special case to be considered is that of incidence data where only the presence or absence of each artifact type in each assemblage is recorded. Three different approaches to this form of data suggest themselves. Firstly it should be possible to reformulate the parametric model to allow for this type of data. Secondly incidence data could be treated as an incomplete form of abundance

		Artifact type							
		1	2	3	4	5	6	7	8
Mean Age	True	86.0	90.0	94.0	98.0	102.0	106.0	110.0	114.0
	Predicted	87.6	81.6	93.0	96.6	101.0	109.8	112.7	113.2
s.d	True	3.0	10.0	7.0	9.0	4.0	5.0	8.0	6.0
	Predicted	2.0	10.0	7.9	11.7	6.3	4.2	8.1	7.2
Correlation	True	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
	Predicted	0.1	0.6	-0.1	0.4	0.1	0.1	0.0	0.2

Table 34.1: Comparison of true and estimated parameter values.

Assemblage No.		Predicted Age	
	True Age		s.e.
1	69.5	76.3	3.9
2	81.9	87.5	1.9
3	89.8	87.7	2.0
4	89.9	81.6	3.5
5	91.2	88.3	2.6
6	95.5	106.2	3.7
7	99.1	87.2	6.7
8	100.4	95.2	2.0
9	100.6	99.8	3.2
10	102.2	98.1	3.1
11	104.9	100.5	4.2
12	105.8	110.5	4.2
13	107.3	106.2	2.8
14	108.1	109.7	2.4
15	108.9	107.9	4.7
16	109.6	113.3	4.1
17	113.8	112.5	3.9
18	115.4	114.1	3.7
19	118.4	117.5	4.9
20	125.0	119.4	5.5

Table 34.2: comparison of true and predicted assemblage ages for simulated data.

data where the missing values are the numbers of each artifact type. This would entail taking expectations with respect to these values as well as to the missing age values, but in principle this should be relatively straightforward. Neither of these approaches is developed further here. The final alternative is to treat the data as if it were abundance data and apply the model as before.

34.3 EXAMPLE

The utility of the new seriation technique was examined through its application to a simulated data set.

34.3.1 Simulation

An artificial data set was constructed, through random number generation, consisting of 105 artifacts divided into 20 assemblages and 8 artifact types. The artifact types were assumed to have equal overall prevalence but age distributions with differing means and variances. One continuous measurement, x, was generated for each artifact such that the joint distribution of age and x within each artifact type was bivariate normal with correlation coefficient equal to 0.3 and the marginal distribution of x a standard normal distribution. The data for each assemblage was constructed by first generating an age value for the assemblage from the mixture of normal distributions formed by amalgamating the individual age distributions of the artifact types. Next the number of artifacts in the assemblage was generated from a Poisson distribution with mean 5, and the type of each artifact determined by random selection from the conditional distribution of artifact types at the appropriate age. Finally the value of x for each artifact was generated from the appropriate conditional normal distribution given the artifact type and assemblage age.

34.3.2 Results

The simplified technique developed above was applied to this data set. The true age values of assemblages 1 and 20 were taken as observed in or-

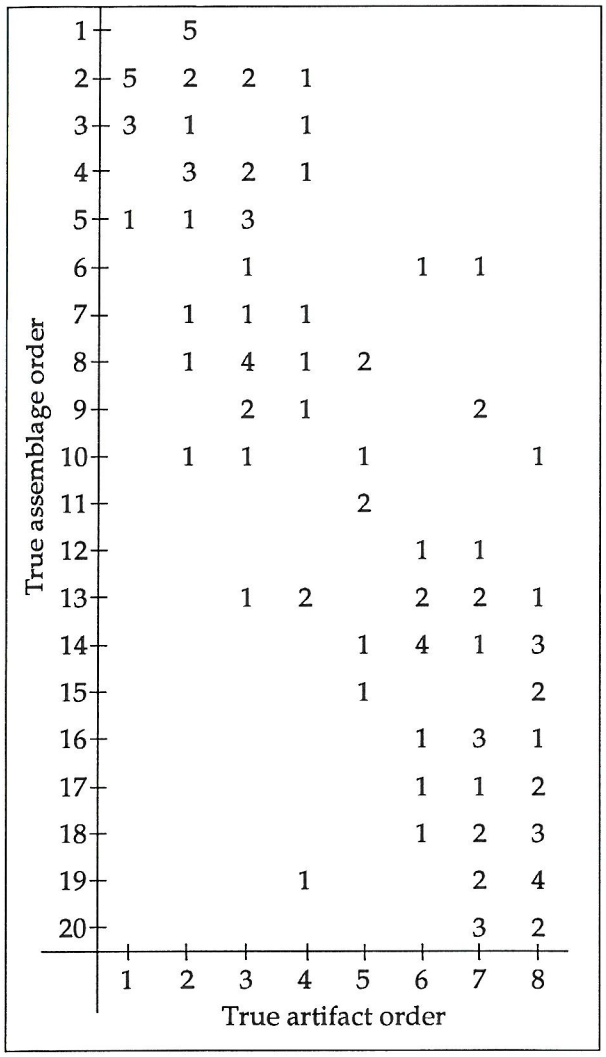

Figure 34.1: Simulated data — true seriation sequence.

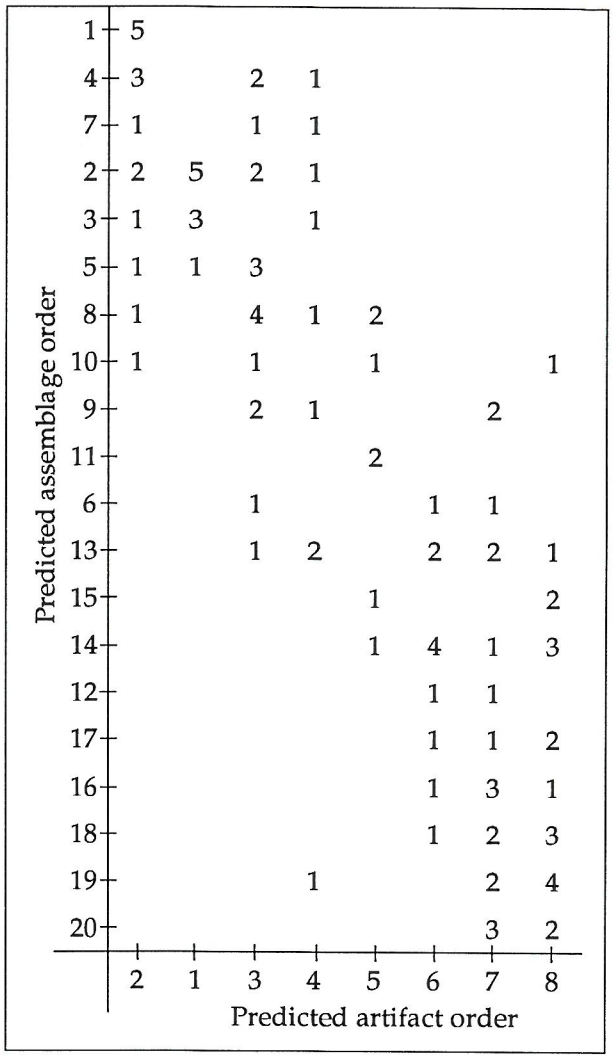

Figure 34.2: Simulated data — predicted seriation sequence.

der to fix the overall age distribution. The procedure coverged fairly rapidly with the age means stabilising much more quickly than the variances. Table 34.1 shows the true and estimated parameter values, while Table 34.2 gives the true and predicted assemblage ages, together with their standard errors. All are surprisingly accurate considering the proportion of missing data involved. Figure 34.1 shows the data in abundance form arranged in the correct seriation sequence. It appears very similar to real data sets, indicating that the underlying distribution used to generate the data may not be unreasonable. Figure 34.2 gives the corresponding predicted seriation sequence. The two are remarkably alike as is shown by Figure 34.3 which plots the predicted against the true sequence. The advantage of this new technique over standard methods is reflected in Figure 34.4 which plots the predicted assemblage ages with their standard errors against the true ages. Because the technique provides a scaling rather than just an ordering of the assemblages it is possible to pick out clusters of assemblages and assemblages which are close together. The standard errors do not reflect the uncertainty in the parameter estimates and hence are to some extent underestimated. Figure 34.4 would suggest however that the extent of this underestimation is not great.

34.4 DISCUSSION

The above example shows that a parametric form of seriation based on incomplete data methodology is not only feasible but in practice performs reasonable well. Previous attempts to derive parametric forms of seriation (e.g. Kendall 1963)

were generally based on the idea of permutating the assemblage sequence in order to maximise some specified criteria. In contrast the present method more realistically regards the age of each assemblage as a quantity to be estimated. This has a number of advantages over previous methods:

1) It is easily possible to incorporate additional information and measurements, both continuous and discrete, into the procedure. Previously this was only possible by categorising continuous variables and regarding all such additional variables as defining new artifact types. Such a process can rapidly produce so many types that very few are represented in more than a handful of assemblages. In addition the ordering inherent in continuous variables is lost when they are categorised.
2) Because the method produces an estimated age and standard error for each assemblage it is possible to judge whether any two assemblages are significantly different from each other. Thus it is possible to examine the accuracy of the predicted sequence. More importantly perhaps it is possible to pick out those assemblages which have the largest standard errors and hence are least accurately dated.
3) Since the method produces a mean age and variance for each artifact type, it automatically seriates types as well as assemblages and makes it possible to compare their utility in age discrimination.
4) Known dates can be incorporated into the seriation procedure, where they act to anchor and scale the time dimension. Thus unlike previous techniques the new procedure can be fully integrated with modern dating methods. This could be very useful for large data sets where it is too expensive to date all assemblages. The dates from a representative sample, could be used to stabilise the seriation procedure, providing more accurate age estimates for the remainder. Alternatively, if a reasonable number of dated assemblages are available and there is doubt as to the validity of the distributional assumptions of the technique, the known dates could be used to calibrate the time scale produced by the seriation procedure. In any case the predicted values for known ages act as a validity check.

The main drawback to the proposed technique is the distributional assumptions used in its derivation. In any specific archaeological situation these may be far from correct and in such cases the estimated assemblage ages and standard errors may

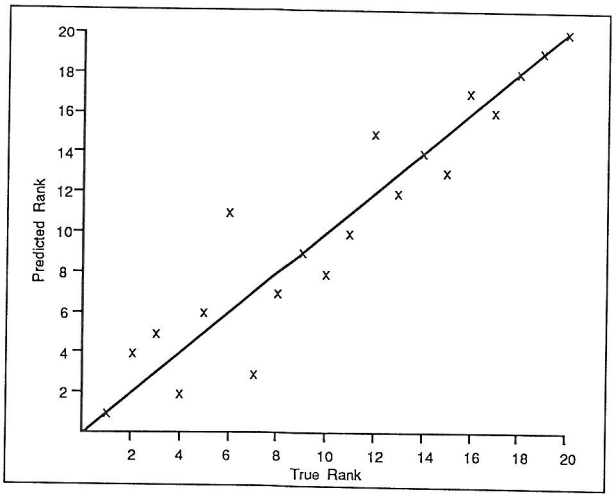

Figure 34.3: Simulated data — Predicted vs True assemblage rank

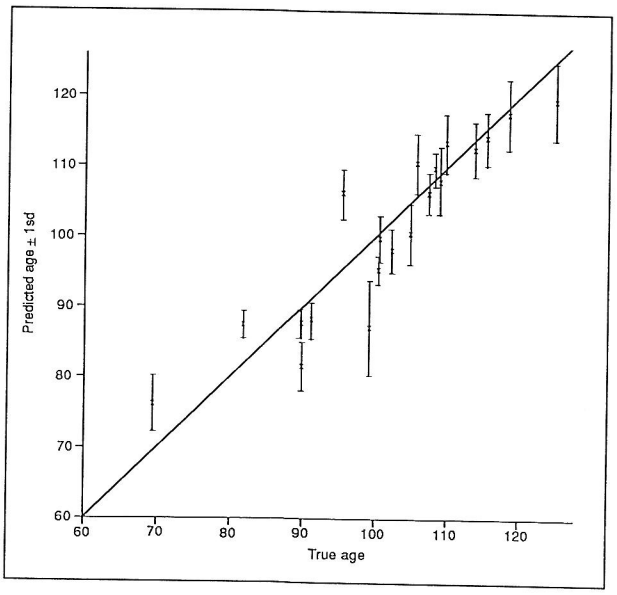

Figure 34.4: Simulated data — Predicted vs True ages

be unreliable. However even if this is felt to be the case, the technique can still be seen as an empirical method for producing a seriation and provides an alternative to current methods. Preliminary investigation using simulated data generated under different distributional assumptions suggests that the technique is robust to the distributions used, although further investigation is needed.

References

Dempster, A.P., Laird, N.M., & Rubin, D.B.
1977 Maximum likelihood from incomplete data via the E.M. Algorithm. *Journal of the Royal Statistical Society B.* 39:1–38.

Ford, J.A.
1962 *A quantitative method for deriving cultural chronology.* Pan American Union, Technical Manual 1, Washington D.C.
Kendall, D.G.
1963 A statistical approach to Flinders Petrie's sequence dating. *Bulletin of the Internationa Statistical Institute* 40:657–680
Petrie, W.M.F.
1899 Sequences in prehistoric remains. *Journal of the Royal Anthropological Institute* 29:295–301

Author's address
Andy Scott,
Dept. of Applied Statistics,
University of Reading,
Reading, U.K.