Biological Journal of the Linnean Society, 2006, **88**, 309–328. With 10 figures

Determining best complete subsets of specimens and characters for multivariate morphometric studies in the presence of large amounts of missing data

RICHARD E. STRAUSS* and MOMCHIL N. ATANASSOV

Department of Biological Sciences, Texas Tech University, Lubbock, Texas 79409-3131, USA

Received 2 April 2004; accepted for publication 18 April 2006

Missing data are frequent in morphometric studies of both fossil and recent material. A common method of addressing the problem of missing data is to omit combinations of characters and specimens from subsequent analyses; however, omitting different subsets of characters and specimens can affect both the statistical robustness of the analyses and the resulting biological interpretations. We describe a method of examining all possible subsets of complete data and of scoring each subset by the 'condition' (ratio of first eigenvalue to second, or of second to first, depending on context) of the corresponding covariance or correlation matrix, and subsequently choosing the submatrix that either optimizes one of these criteria or matches the estimated condition of the original data matrix. We then describe an extension of this method that can be used to choose the 'best' characters and specimens for which some specified proportion of missing data can be estimated using standard imputation techniques such as the expectation-maximization algorithm or multiple imputation. The methods are illustrated with published and unpublished data sets on fossil and extant vertebrates. Although these problems and methods are discussed in the context of conventional morphometric data, they are applicable to many other kinds of data matrices. *Biological Journal of the Linnean Society*, 2006, **88**, 309–328.

ADDITIONAL KEYWORDS: covariance – discriminant analysis – eigenanalysis – expectation-maximization – fossils – imputation – matrix condition – principal component analysis – morphometrics – simulation.

INTRODUCTION

The problem of missing data is relatively common in both observational and experimental studies in the biological, archaeological, and palaeontological sciences. Morphometric data sets, whether of fossil or extant organisms, usually consist of many characters (mensural variables) measured on each specimen (observation), often with complex patterns of missing, uncertain, or indeterminate values.

Because multivariate morphometric procedures generally require complete data matrices, with all character values present for all specimens, there are two possible solutions (Fomby, 1998). The first and better-studied solution is to estimate missing values from the available data (Beale & Little, 1975; Rubin, 1976; Little & Rubin, 1987). A number of statistical techniques have been developed over the past 30 years for estimating or imputing missing data (Allison, 2001). These vary from the simplest (and highly inadvisable) approach of replacing missing values by the univariate character means (Wilks, 1932), to sophisticated multivariate methods based on maximum-likelihood or Bayesian probabilities. The most commonly used multivariate method is the expectation-maximization (EM) imputation method (Dempster, Laird & Rubin, 1977; Strauss, Atanassov, & Oliveira, 2003), which estimates all missing values in the data matrix as a set in an iterative fashion via the covariance or correlation matrix. We do not address multiple imputation in this discussion because that method is specific to a particular kind of analysis and does not produce a single 'best' complete data matrix, as fully parametric 'single-imputation' procedures do (Rubin, 1996; Schafer & Olsen, 1998). Missing values can be estimated if they do not comprise too large a proportion of the data matrix.

The second, much more common solution to the problem of incomplete data is to omit the specimens or

*Corresponding author. E-mail: rich.strauss@ttu.edu

characters having missing values (marginalization), which can seriously reduce the sample size available for analysis (Gauthier, Landry, & Lapointe, 2003). Although omitting missing data reduces statistical power and can potentially lead to bias of results (Millis, 2003), it may be necessary if the proportion of missing data is large. Many alternate complete subsets may be possible (Table 1), and the choice of a particular complete subset can significantly influence the conclusions of a study (Proschan *et al*., 2001). The question thus arises as to which particular subsets of specimens or characters should be used, which in turn raises the issue of the criterion that should be used to judge the 'best' subset. If the purpose is a multivariate morphometric study, then the 'best' complete submatrix might be based on the adequacy of the dataset for subsequent statistical analysis, or might be selected to approximate the statistical properties of the original matrix.

The present study concerns data matrices that have relatively large proportions of missing data, as, for example, are common in palaeontological and

Table 1. The ten best subsets of complete data (of 203 possible) from a data matrix of log-transformed morphometric measurements of specimens of the pterosaur *Rhamphorhynchus* (Wellnhofer, 1975), ranked by decreasing condition number; by decreasing reciprocal condition factor; and by increasing difference between the condition number of the original matrix and that of the complete subset

Factor value	Characters												Number of specimens
By condition number													
6.337	8	10	11	12	–	–	–	–	–	–	–	–	62
6.278	8	10	11	–	–	–	–	–	–	–	–	–	70
6.207	7	8	10	11	12	15	–	–	–	–	–	–	41
6.075	7	8	10	11	12	–	–	–	–	–	–	–	59
6.033	7	8	10	11	15	–	–	–	–	–	–	–	46
5.953	7	8	10	11	–	–	–	–	–	–	–	–	67
5.906	10	11	12	13	–	–	–	–	–	–	–	–	62
5.887	8	10	11	12	13	–	–	–	–	–	–	–	59
5.882	7	8	9	10	11	12	13	15	–	–	–	–	37
5.846	6	7	8	9	10	11	12	13	15	–	–	–	28
By reciprocal condition number													
–3.904	1	7	9	–	–	–	–	–	–	–	–	–	52
–3.929	1	8	9	–	–	–	–	–	–	–	–	–	52
–4.175	1	7	8	9	–	–	–	–	–	–	–	–	50
–4.182	7	8	9	–	–	–	–	–	–	–	–	–	70
–4.578	1	9	10	–	–	–	–	–	–	–	–	–	51
–4.613	1	2	7	8	9	10	–	–	–	–	–	–	33
–4.653	7	9	10	–	–	–	–	–	–	–	–	–	69
–4.660	8	9	10	–	–	–	–	–	–	–	–	–	70
–4.780	1	7	9	10	–	–	–	–	–	–	–	–	49
–4.787	1	2	4	5	8	9	10	11	12	13	14	15	13
By best matching condition number													
0.535	1	2	4	5	8	9	10	11	12	13	14	15	13
0.560	1	2	4	7	8	9	10	11	12	13	14	15	14
0.806	1	8	9	10	–	–	–	–	–	–	–	–	48
0.813	4	5	7	8	9	10	11	12	13	–	–	–	23
1.209	1	2	7	8	9	10	11	14	15	–	–	–	22
1.243	4	5	8	9	10	11	12	13	14	15	–	–	19
1.284	1	7	8	–	–	–	–	–	–	–	–	–	52
1.291	1	4	5	8	9	10	11	12	13	14	15	–	16
1.411	4	6	7	8	9	10	11	12	13	–	–	–	23
1.417	4	5	6	7	8	9	10	11	12	13	–	–	19

The original matrix contains 96 specimens and 17 characters; complete subsets were constrained to have at least 10 specimens and three characters. Given for each subset are the identities of the particular characters comprising it and the number of specimens having complete data for those characters.

archaeological studies. The objectives are: (1) to describe and evaluate a procedure for examining all possible subsets of complete characters and specimens and selecting the 'best' subset, in terms of the statistical properties (particularly 'condition') of the resulting submatrix; and (2) to suggest how this submatrix might then be augmented by the 'best' subset of characters and specimens for which missing data can be estimated. Although the emphasis is on morphometric studies, the problems and solutions discussed here extend to other multivariate contexts.

SELECTING BEST COMPLETE SUBSETS

Enumerating complete submatrices

The procedure for finding all possible subsets of complete characters and specimens is straight-forward but computationally intensive: examine all possible combinations of characters and, for each possible combination, find all the specimens having complete data. Unless the distributional properties of the missing data within the matrix are known precisely, the problem is NP-complete and no method other than brute force is possible. For morphometric studies, a minimum of three characters is generally useful, so the number of characters examined at a time is varied from three to P, the total number of characters (although the lower limit can be increased). For p ($\leq P$) characters, the number of possible subsets is the number of combinations of P characters sampled p at a time: ${}_PC_p$. The total number of subsets S is then the sum of these combinations for $p = 3, \ldots, P$:

$$S = \sum_{p=3}^{P} {}_PC_p = \sum_{p=3}^{P} \frac{P!}{p!(P-p)!}$$

S becomes relatively large for numbers of characters greater than 20 or so (e.g. $P = 15$, $S = 32\,647$; $P = 20$, $S = 1048\,365$; $P = 25$, $S = 33\,554\,106$), and so this procedure is impractical for more than approximately 25 characters. For larger numbers, the characters having the greatest numbers of missing values can first be omitted. For data sets having more characters than specimens (e.g. the *Archaeopteryx* data described below), the roles of characters and specimens can be inverted for this purpose.

Numbers of characters and specimens for all possible complete submatrices are illustrated in Figure 1 for four important palaeontological data sets having substantial amounts of missing data: *Rhamphorhynchus*, with 96 specimens, 17 characters, and 35.3% missing values (Wellnhofer, 1975); *Pterodactylus*, with 64 specimens, 13 characters, and 11.8% missing values (Wellnhofer, 1970); *Pteranodon*, with 511 specimens, 14 characters, and 83.2% missing values (Bennett, 1991); and *Archaeopteryx*, with seven specimens, 132 characters, and 50.5% missing values (Wellnhofer, 1974, 1988, 1993). *Rhamphorhynchus* and *Pterodactylus* are Late Jurassic pterosaurs (from Solnhofen, Germany) and *Pteranodon* is a Late Cretaceous pterosaur (from North America). Twelve specimens (Wellnhofer's numbers 88, 89, 91, 93, 96–98, 100, 102–104, and 107) were omitted from the *Rhamphorhynchus* data set because of excessive numbers of missing data. For the *Archaeopteryx* data set, some of the missing data are unpublished but are potentially available from known specimens. The published *Archaeopteryx* data were used previously for a multivariate analysis of allometric patterns by Houck, Gauthier & Strauss (1990).

As the number of characters in the subset increases, the number of specimens decreases on average (Fig. 1), reflecting the trade-off in omitting characters vs. specimens from the original data sets. The structure of the patterns of the scatterplots of numbers of specimens vs. characters reflects the unevenness of the distribution of missing values in the data sets. For example, in the *Rhamphorhynchus* data set, most of the complete values are concentrated in the wing whereas, in the other data sets, the missing values are more randomly scattered among characters and specimens.

Different subsets can produce different results

The differences in combinations of specimens and characters can produce differing results in a multivariate analysis (Proschan *et al*., 2001), even if the specimens having missing data are representative of the entire sample. For example, discriminant analyses of three species of *Rhamphorhynchus* based on two different complete subsets (Figs 1A, 2) suggest different conclusions about interspecific relationships and degrees of distinctiveness. The species are distinguished primarily by differences in body size: *Rhamphorhynchus longicaudus* and *Rhamphorhynchus intermedius* are small-bodied whereas *Rhamphorhynchus muensteri* is larger. The analysis based on just five characters indicates that almost all (99%) of the discrimination among species is due to size variation, whereas that based on fewer specimens but ten characters suggests that relatively less discrimination among species is due to size variation (82%), and the remaining 18% is due to substantial shape differences between *Rhamphorhynchus intermedius* and the other two species. The Mahalanobis distance D^2 between the *R. longicaudus* and *R. intermedius* samples increases from a nonsignificant 16.4 (in units of variance; $P = 0.08$) for the first subset to a highly significant 36.9 ($P = 0.01$) for the second. In contrast, the Mahalanobis distance between the *R. longicaudus* and

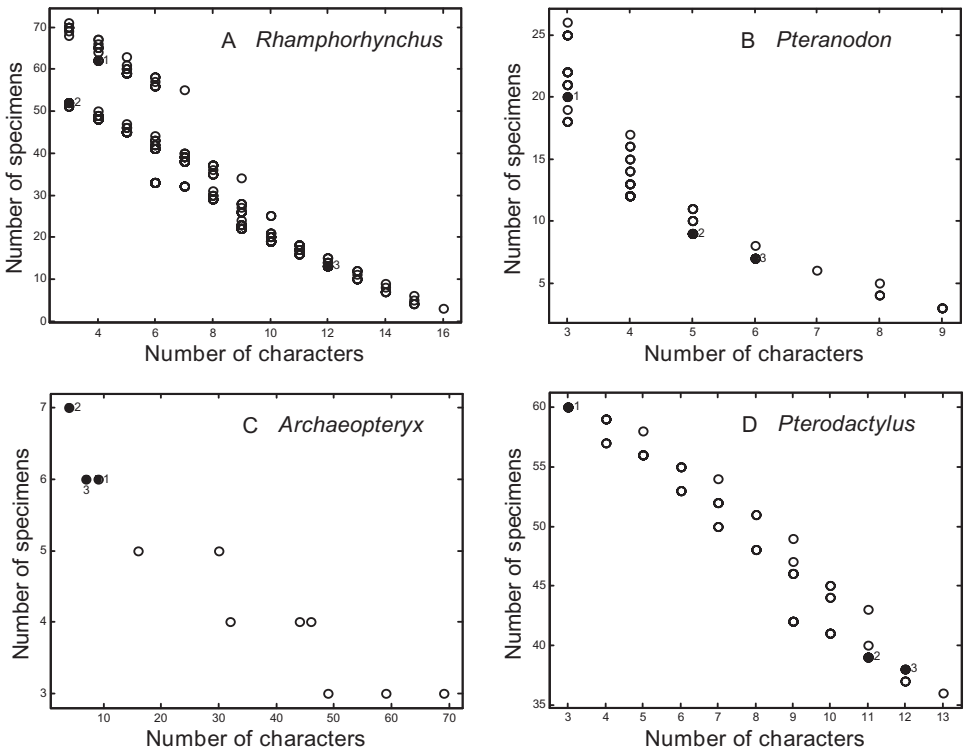

Figure 1. Scatterplot of possible combinations of specimens and characters having complete data, for specimens of (A) *Rhamphorhynchus*, (B) *Pteranodon*, (C) *Archaeopteryx*, and (D) *Pterodactylus*. Each point represents one or more unique combinations of specimens and characters. Numbered solid circles represent complete subsets based on: (1) the maximum condition index; (2) the maximum reciprocal condition index; and (3) the best matching condition index.

R. muensteri samples increases from a highly significant 100.6 ($P < 0.001$) to a larger but nonsignificant 146.1 ($P = 0.07$). The differences in statistical resolution and power are, of course, a consequence of the differences in numbers of variables and observations, but the results observed are also a function of the particular specimens and characters included in the analysis.

Differences in results thus can substantially affect the conclusions of a study. For example, Bennett (1995, 1996), on the basis of only a few characters, concluded that the different forms of *Rhamphorhynchus* are conspecific. In contrast, our own studies of the same taxa using more characters suggest that most of the forms are sufficiently distinctive to warrant formal recognition (Atanassov & Strauss, 1999; Atanassov & Strauss, 2000).

Subset-selection criteria

There are many different criteria that might be used for choosing among all possible submatrices of complete data. If the data are to be used for a multivariate analysis, then the submatrix having the best statistical properties in some sense might be selected. (This is probably a better policy than selecting the best subset based on the biological results because we are likely to be biased by the latter.) Here, we describe three statistical criteria that are based on measures of matrix condition.

Measuring matrix condition

Many standard multivariate methods (e.g. principal component analysis, discriminant analysis) are variants of an eigenanalysis or singular-value decomposition of the covariance or correlation matrix derived from the data matrix (Tabachnick & Fidell, 2006). Simple functions of the spread of the eigenvalues of a matrix provide useful diagnostics for the numerical stability of a matrix. Measures of condition of a matrix $\mathbf{X}$ are based on the eigenvalues (characteristic roots) of $\mathbf{X}'\mathbf{X}$ and measure the degree to which small relative changes in $\mathbf{X}'\mathbf{X}$ produce large relative changes in the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ (Groß, 2003). The spectral-norm condition number C, the square-root of the ratio of the largest eigenvalue to the smallest, is commonly used to measure the degree to which the covariances capture linear dependencies, the information that all variables have most in common. Its value is undefined for a sin-

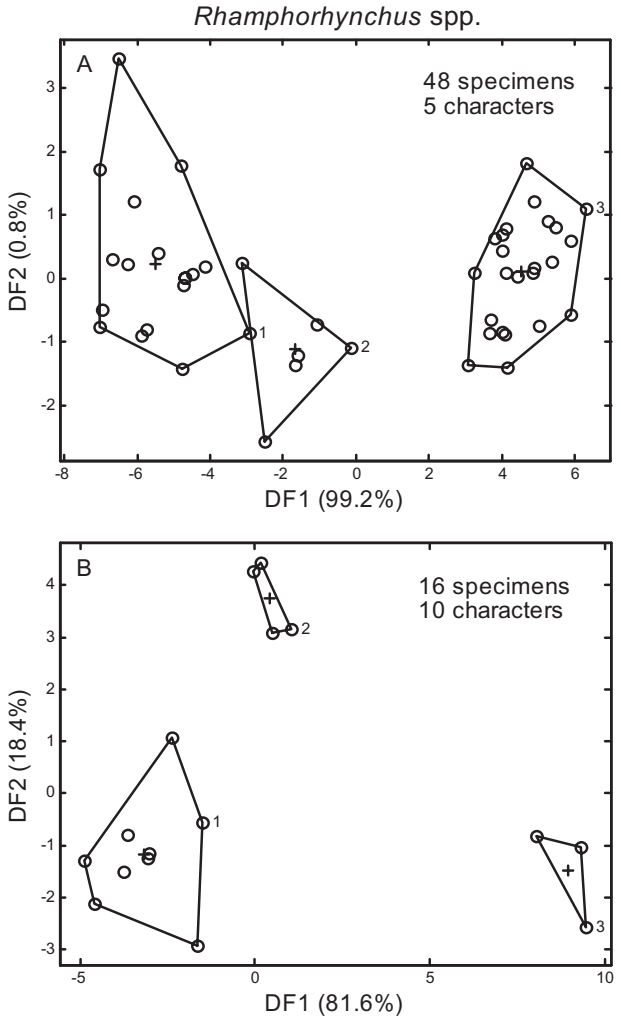

Figure 2. Discriminant analyses among species of *Rhamphorhynchus* based on two different complete submatrices of data. Species 1, *Rhamphorhynchus longicaudus*; 2, *Rhamphorhynchus intermedius*; 3, *Rhamphorhynchus muensteri*. A, discriminant function analysis (DFA) based on 48 specimens and five characters. B, DFA based on 16 specimens and 10 characters.

gular matrix (i.e. one containing linear dependencies among the variables), and in fact the base-b logarithm of C is an estimate of how many base-b digits are lost in solving a linear system with that matrix; that is, C estimates the worst-case loss of precision due to commonality of structure.

Condition numbers and their reciprocals can be expressed in log-ratio form, ignoring the square roots:

$$C = \ln\left(\frac{\lambda_1}{\lambda_P}\right) = \ln(\lambda_1) - \ln(\lambda_P),$$

$$C_r = \ln\left(\frac{\lambda_P}{\lambda_1}\right) = \ln(\lambda_p) - \ln(\lambda_1),$$

where λ_1 and λ_P are the first and last ordered eigenvalues (sorted high to low), respectively, for P variables, and C_r is the reciprocal of C. Values of C_r on this logarithmic scale are the negative values of C, and vice versa. Both C and C_r are undefined for singular covariance matrices, for which one or more of the last ordered eigenvalues are zero.

We have made a further modification of the condition and reciprocal-condition numbers, using instead the log-ratios of the first and second eigenvalues rather than the first and last:

$$C' = \ln\left(\frac{\lambda_1}{\lambda_2}\right) = \ln(\lambda_1) - \ln(\lambda_2),$$

$$C'_r = \ln\left(\frac{\lambda_2}{\lambda_1}\right) = \ln(\lambda_2) - \ln(\lambda_1).$$

These versions are log-transformed condition indices for the second eigenvalue (Groß, 2003), and have several advantages over the conventional condition numbers. First, C' and C'_r are defined for nonpositive-definite or positive semidefinite covariance matrices (i.e. having one or more negative or zero eigenvalues), both singular and nonsingular, whereas C and C_r are defined only for positive definite covariance matrices (having all positive eigenvalues). Second, C' and C'_r are nearly linearly correlated with the skewness of the eigenvalue distribution and with the mean correlation among the variables (Fig. 3C, D), both of which are measures of covariance structure. In contrast to this approximately linear behaviour of C' and C'_r, the conventional measures C and C_r are disproportionately sensitive to high levels of correlation and skewness of the eigenvalues (Fig. 3A, B). These relationships were determined by simulation for 30 observations and 12 variables constituting a homogeneous, multivariate normal sample by: (1) systematically varying the population correlation from 0.01 to 0.99, in increments of 0.01; (2) generating repeated random samples, 1000 per correlation level, from each population using the method of Kaiser & Dickman (1962); (3) calculating the eigenvalues of the correlation matrix of each random sample and corresponding condition indices and eigenvalue skewness values; and (4) plotting the means among replicates as a function of level of correlation and eigenvalue skewness.

Selecting subsets based on matrix condition

Our initial assumption was that the 'best' complete submatrix of data should be the one that corresponds in some way to the intended method of data analysis. For example, when the purpose of the study is to characterize patterns of allometric size variation or of size-invariant variation in shape among individuals via some form of eigenanalysis (e.g. principal component

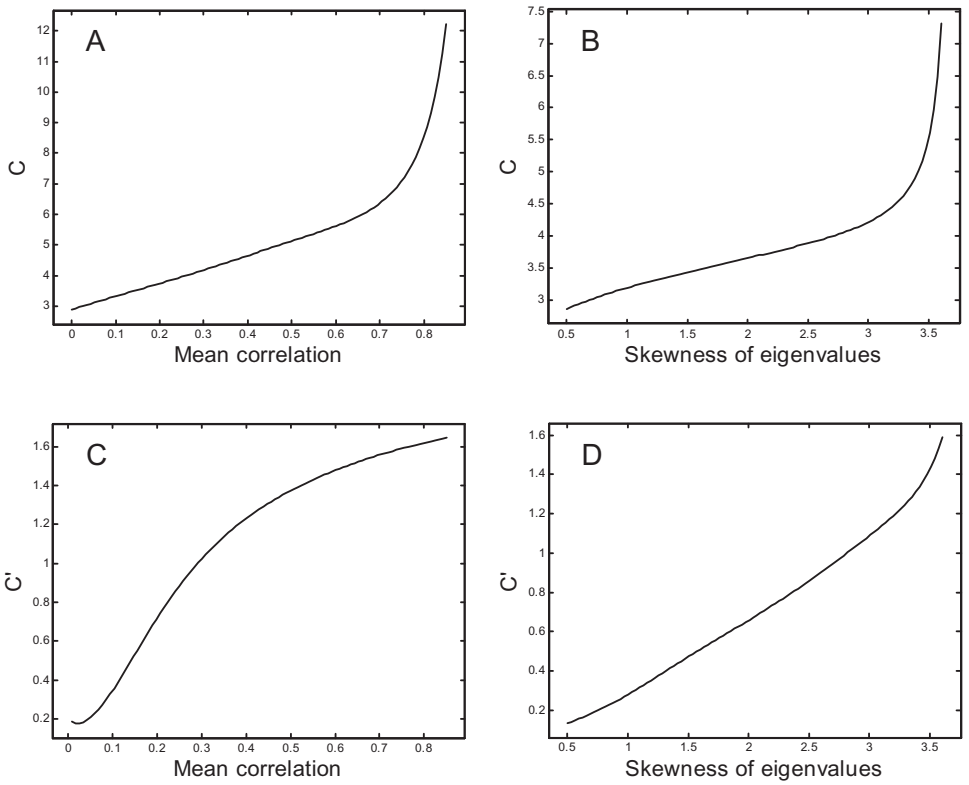

Figure 3. Relationships of the condition index C and the modified condition index C', respectively (both in logarithmic form) to (A, C) the mean level of correlation within the data matrix and (B, D) the skewness of the eigenvalue distribution of the covariance matrix.

analysis or common principal component analysis; Flury, 1988), the condition index C', which measures the ability of the covariances to capture the information that all variables have in common, should measure the stability of the analysis with respect to the data. In this case, the complete submatrix yielding the covariance matrix with the greatest condition index, max C', might be expected to have the best statistical properties for such an analysis.

However, if the purpose of the analysis is to use a method that involves a matrix inversion (e.g. multivariate analysis of variance, discriminant or canonical-variate analysis, estimation of Mahalanobis distances) or if the smaller eigenvectors are otherwise important (Jolliffe, 1982), then the greatest reciprocal condition index, max C'_r, which measures the stability of the covariance matrix to inversion, should characterize the complete submatrix having the best corresponding statistical properties.

Rather than matching the analytic method, a third possibility is that, for any kind of analysis, we identify the complete submatrix having statistical properties that best match those of the original data matrix. In this case, we might estimate the condition index associated with the original matrix and choose the complete submatrix that best matches it:

$$\min \Delta C' = \min(C'_o - C'_s)$$

where C'_o is the estimated condition index of the original matrix and C'_s is the condition index of the complete submatrix.

Matching matrix condition of the original matrix

Finding the complete submatrix of the data having a covariance matrix that best matches the covariances of the original (target) matrix might appear to be the best overall solution to the problem. However, if a submatrix is produced by omitting variables, the covariance matrices of the original and submatrices will be of different dimensions and a direct comparison is not possible. Because the first few eigenvectors and corresponding eigenvalues capture the most important structural aspects of a covariance matrix, comparison of the first few of these elements appears to be a reasonable alternative. Here, we concentrate on the first two eigenvalues (via the modified condition index),

although comparison of a larger number of eigenvalues could provide a more resolved measure of congruence.

An important problem is how to estimate the eigenvalues of the original matrix containing missing values. Very little work has been done concerning methods of directly obtaining eigenvalues and eigenvectors in the presence of missing data. More commonly, estimates of the covariance or correlation matrix are obtained using missing-data imputation techniques. For small numbers of missing values, this can be performed using maximum-likelihood methods. However, for relatively large amounts of missing data, the problem is more serious, given the requirements that a nonsingular covariance matrix must be not only real-symmetric, but also positive-definite (i.e. with all eigenvalues greater than zero). If the covariances among pairs of variables are estimated by pairwise deletion (Wilks, 1932), using only observations for which both variables have values, the resulting matrix will be only an approximate covariance matrix because the individual covariances will have been constructed from different, and possibly inconsistent, sets of observations (Hill & Thompson, 1978; Arbuckle, 1996). Such a matrix will be real-symmetric but, in the worst case, may have negative eigenvalues. The condition index of such a matrix is likely to be biased.

Because little empirical literature exists on this problem, the probability of obtaining a nonpositive-definite covariance matrix by pairwise deletion was estimated by simulation (Fig. 4). Number of variables was set at 5, 10, 15, and 20; number of observations was varied from 10 to 100 in increments of 5; and proportion of missing data was varied from 0% to 50% in increments of 1%. For each combination: (1) a random data matrix was generated using the method of Kaiser

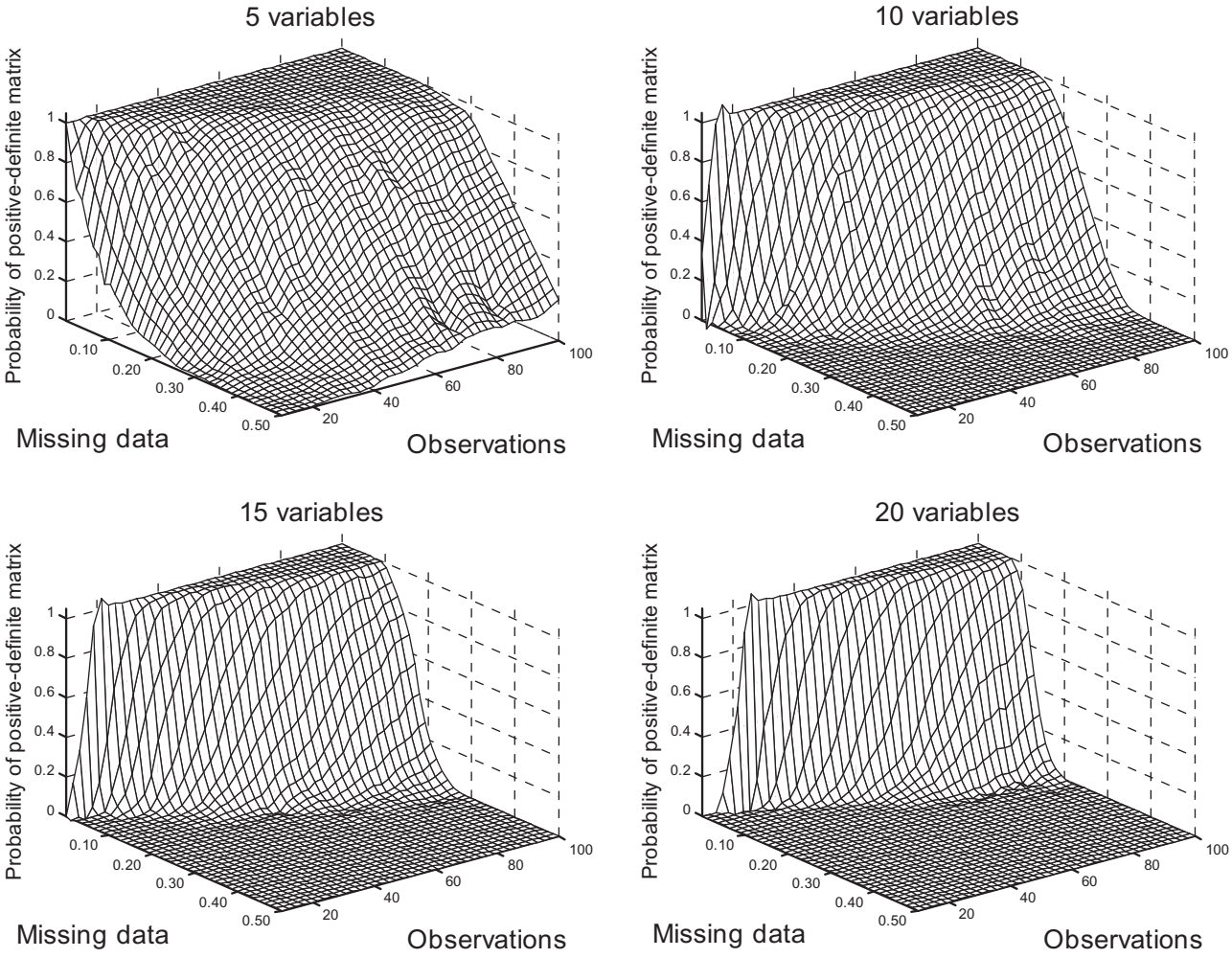

Figure 4. The estimated probability of a sample covariance matrix being positive-definite, as a function of the properties of the sample data matrix: number of variables, number of observations, and proportion of missing completely at random values. Interpolated probability surfaces were estimated by simulation, as described in the text.

& Dickman (1962) with a population correlation of 0.7 for all pairs of variables; (2) the required number of missing values was randomly scattered into the data matrix; (3) the covariance matrix was estimated by pairwise deletion; (4) a singular-value decomposition was performed on the estimated covariance matrix; and (5) the matrix was scored as either positive-definite or not, based on the occurrence of zero or negative eigenvalues. This procedure was repeated 1000 times for each combination of variables, observations and proportion of missing data, and the proportion of non-positive-definite matrices out of the total was taken to be an estimate of the probability of obtaining such a matrix under these conditions. The interpolated surfaces of Figure 4 suggest that, for reasonable numbers of specimens and characters, the probability of obtaining a proper, positive-definite covariance matrix by pairwise deletion of missing values is generally very low.

Given that this is likely to be a common problem, the challenge is then to find a positive-definite covariance matrix that is a close approximation of the indefinite one. There have been two basic approaches to this problem. The first is to estimate the missing values within the data matrix so as to stabilize the covariances; this is generally carried out using the expectation-maximization method (Dempster *et al*., 1977). Because the context of the present study is to reduce data matrices having large proportions of missing values, and imputation methods often fail to converge or to provide realistic estimates of missing values for large amounts of missing data, this alternative might not be feasible.

The second approach is to directly estimate the 'closest' proper covariance matrix, altering the improperly structured covariance matrix directly to approximate a positive-definite matrix by altering the eigenvalues until the minimum is greater than zero. The 'bending' method of Hayes & Hill (1980, 1981) has most commonly been used to constrain such estimated covariance matrices, particularly in quantitative genetics (Essl, 1991) and econometrics (Rebonato, 1999). The main disadvantages of the bending method are: (1) that it requires as a starting point and 'anchor' a positive-semidefinite matrix, the choice of which in this case is not obvious; (2) that the method does not preserve the covariances for pairs of variables having no missing data; and (3) that there is no way of determining to what extent the resulting matrix is optimal in any easily quantifiable sense. Modifications of the bending method have been proposed by Essl (1991) and Jorjani, Klei & Emanuelson (2002, 2003), whereas other methods for covariance or correlation matrices have been described by Frane (1976), Hu (1995), Kupiec (1998), Lucas (2001), Chen & McInroy (2002), and Higham (1989, 2002), amongst others. Knol & Ten Berge (1989) proposed a least-squares approximation for correlation matrices employing oblique Procrustes rotation. This technique allows the correlations for pairs of variables having complete data to remain unchanged. For the present study, we implemented their least-squares algorithm, secondarily estimating the corresponding covariance matrix as $\mathbf{C} = \mathbf{SRS}$, where $\mathbf{R}$ is the estimated correlation matrix and $\mathbf{S}$ is a diagonal matrix of the standard deviations of each variable obtained from all data available for that variable.

Examples

In addition to exemplifying the trade-off between numbers of specimens and numbers of characters for several empirical data sets, Figure 1 also indicates the particular complete subsets identified by the three optimization criteria. The three criteria can produce substantial differences in the size and structure of the best complete subset. Table 1 shows the ten best submatrices from the log-transformed *Rhamphorhynchus* data set of Figure 1(A), ranked by decreasing condition, C', by decreasing reciprocal condition, C'_r, and by increasing difference between the condition of the original matrix and that of the complete subset, $\Delta C'$. Table 1 illustrates that the particular complete subsets identified by each criterion can differ markedly not only in the numbers of characters selected, but also in their particular identities. Even when the characters included in two or more subsets are nearly identical, the numbers and combinations of specimens included can vary considerably.

The condition (or reciprocal condition) of a submatrix depends on the structure of the covariances among the particular characters sampled (Cole *et al*., 1994), and therefore is not a simple function only of numbers of characters and specimens. For example, for the *Rhamphorhynchus*, *Pteranodon*, and *Archaeopteryx* data sets, the best-conditioned submatrices by the C'_r criterion always involve few characters, but not always the greatest number of specimens (Fig. 1).

A Matlab (Mathworks, 1997) function *varcomb*, which finds and ranks all possible submatrices based on the optimization criteria used in the present study, is available at: http://www.biol.ttu.edu/Strauss/Matlab/Matlab.htm.

Assessment of criteria for multivariate analyses

Given that these optimization criteria appear to be reasonable for selecting particular complete submatrices of data to be used for subsequent analyses, the question remains as to how they might perform in real studies. Our initial working assumption was

that the max C' criterion would be best for eigenanalysis problems, max C'_r would be best for methods involving matrix inversions, and the 'best matching' criterion min $\Delta C'$ would be a reasonable compromise that would be suboptimal for each but that would preserve the original structure of the data to the greatest extent. To pursue this question, we simulated their performance for two multivariate methods commonly used in morphometric studies, principal component analysis (PCA) and discriminant function analysis (DFA = canonical variate analysis), using two complete multiple-group data sets differing in the amount of discrimination among the groups. The *Canis* data set (Mammalia, Canidae: Atanassov, 1996; Strauss *et al*., 2003) comprises three species (*Canis aureus*, $N = 30$; *Canis lupus*, $N = 27$; *Canis familiaris*, $N = 24$), with a total of 81 specimens and 15 cranial morphometric characters. The *Cottus* data set (Pisces, Cottidae: Strauss, 1989; Strauss, 1991) also comprises three species (*Cottus cognatus*, $N = 17$; *Cottus bairdi* (cf. *Cottus caeruleomentum*; Kinziger, Raesly, & Neely, 2000), $N = 19$; *Cottus carolinae*, $N = 16$) with 52 total specimens and 15 morphometric characters; specimens and characters were selected by random draw from larger sets to be comparable in number to those of the *Canis* data set. Basic PCA and DFA results are shown in Figure 5. The *Canis* species are quite distinguishable

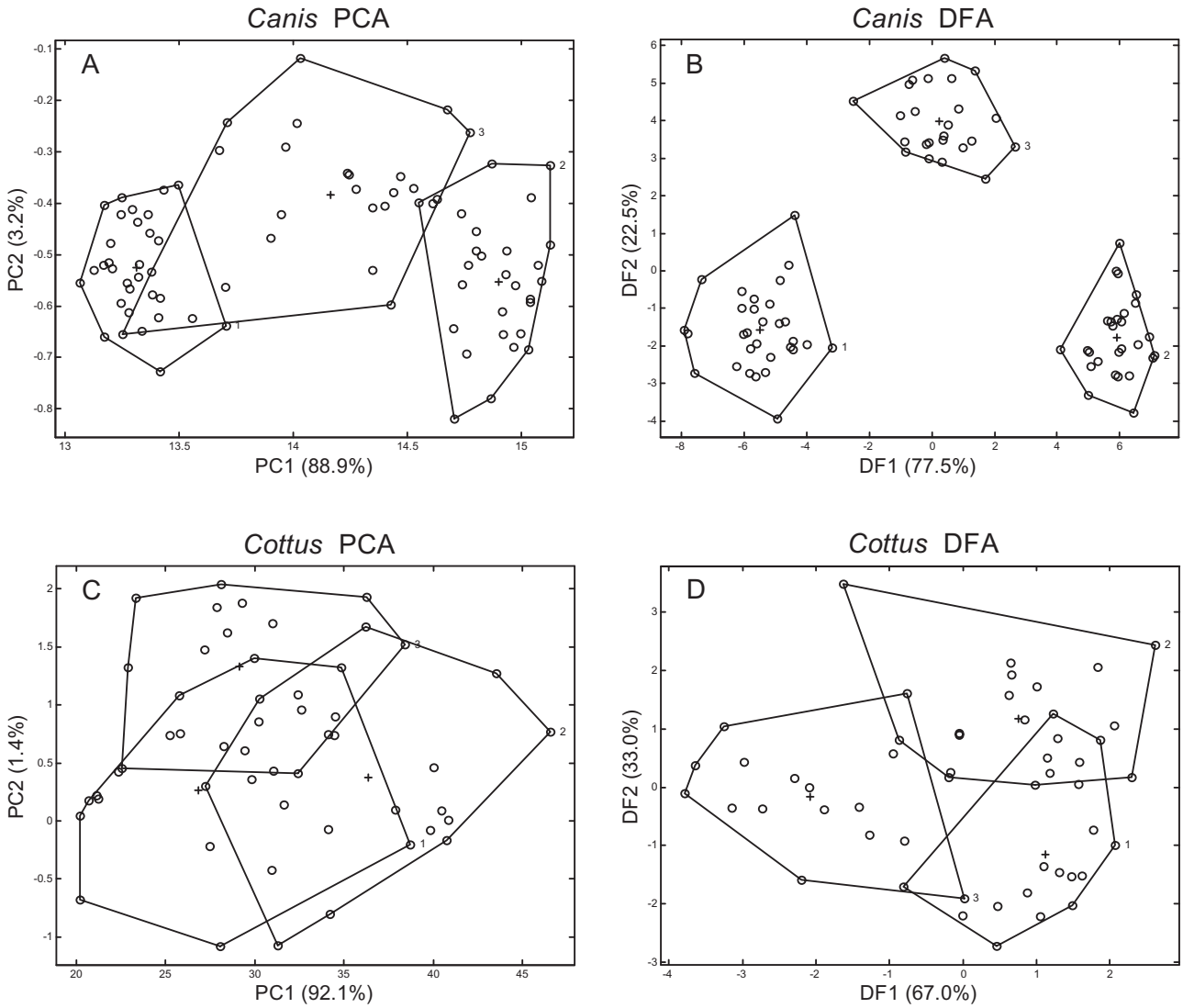

Figure 5. Four analyses of two complete empirical data sets used for missing-data simulations. A, principal component analysis (PCA) of data from three species of *Canis*, with 81 specimens and 15 characters. B, discriminant function analysis (DFA) of the *Canis* data. C, PCA of data from three species of *Cottus*, with 52 specimens and 15 characters. D, DFA of the *Cottus* data.

in terms of the first two principal components and the discriminant functions (with pairwise Mahalanobis distances D^2 in the range 97–216), whereas the *Cottus* species overlap widely on the first two components and marginally on the discriminant functions (D^2 in the range 5.6–11.2).

For each of the four analyses (two methods × two empirical data sets), we assessed the behaviour of the three optimization criteria (C', C'_r, and $\Delta C'$) with respect to a null hypothesis that the optimal complete submatrix performs no better than a randomly selected complete submatrix. Because we began with complete data matrices, we could simulate the occurrence of varying amounts of missing data and compare the results from any complete submatrix to the original results based on the complete data. For this purpose, we used two performance criteria, one based on congruence of the loadings of variables (function coefficients), and the second on congruence of the projection scores of observations. Both were measured as product-moment correlations. For PCAs, character congruence was measured as the correlation between the corresponding loadings on the first two components from the reduced submatrix and original complete-matrix results, whereas specimen congruence was measured as the correlation between the corresponding scores on the first two components. For DFAs, correlations were between corresponding loadings and scores on the (only) two discriminant functions.

The amount of missing data introduced into the complete data matrix was varied from 5% to 50% in increments of 5%. For each iteration of the simulation, the required number of missing values was scattered randomly throughout the matrix, with each value equally likely to be assigned as missing. All possible complete submatrices were found, and the submatrices corresponding to the three criterion values (max C', max C'_r, min $\Delta C'$) were identified; in addition, a random complete submatrix was selected, corresponding to the null hypothesis. Character congruence and specimen congruence with the complete-matrix analysis were determined for each and saved. This procedure was repeated for 1000 iterations to accumulate randomized sampling distributions of the performance criteria. Resulting distributions for the two methods × two empirical data sets are portrayed with box plots in Figures 6, 7, 8, 9; the four figures correspond to Figure 5A–D.

Several generalizations can be inferred from the results of these simulations.

First, the distributions of the character-congruence and specimen-congruence statistics were surprisingly broad for all three optimization criteria. Even in the best cases (Fig. 7E, F), some 'optimal' complete subsets performed poorly and, for most simulations, the median correlation levels were typically in the range of 0.6–0.9.

Second, increasing the amount of missing data has an inconsistent effect on the performance of the complete subsets. In most cases, increasing the amount of missing data from 5% to 10% to 15% produced a consistent but relatively minor decrease in performance. For some simulations, such as the PCAs of the *Cottus* data (Fig. 8), increasing amounts of missing data produced a progressive decrease in performance up to approximately 35–40%, with no further decrease beyond that level. For most other simulations the effect was less progressive or nonexistent.

Third, in general, none of the three optimization criteria performs much better than random for the *Canis* analyses (Figs 6, 7), with two exceptions: (1) complete subsets for min $\Delta C'$ (best matching condition) criterion consistently resulted in significantly better congruence of scores, but not loadings, for both PCA and DFA and (2) max C'_r produced significantly and consistently better congruence of loadings for DFA, but not of scores. Presumably the three *Canis* species are sufficiently distinctive that any complete subset of data will characterize the underlying trends and differences almost as well as any other subset.

Finally, for the *Cottus* analyses (Figs 7, 8), the min $\Delta C'$ (best matching condition) criterion consistently performs as well or better than the other criteria for both PCA and DFA, and much better than random. The max C' criterion performs slightly better than random for the PCA and max C'_r performs much better than random for the DFA, although neither is consistently better than min $\Delta C'$. It is evident that, because of the broad overlap of the *Cottus* species in the multivariate spaces, different complete subsets can produce widely varying PCA and DFA results, and that choice among them can be critical in recovering the underlying structure of the data.

In general, the results from the simulations suggest that the best overall strategy is to choose the complete subset that best matches the condition index, and thus the underlying covariance structure, of the full matrix. If the structure in the data is obvious (as in the *Canis* analyses), the best-matching criterion performs as well as, and sometimes better than, a randomly selected complete subset of data. But if the structure is more subtle (as in the *Cottus* analyses), then the performance of the best-matching criterion is much better than random for both kinds of analyses.

Estimating portions of missing data

If the proportion of missing data in the original matrix is not too large, all missing values can be estimated as

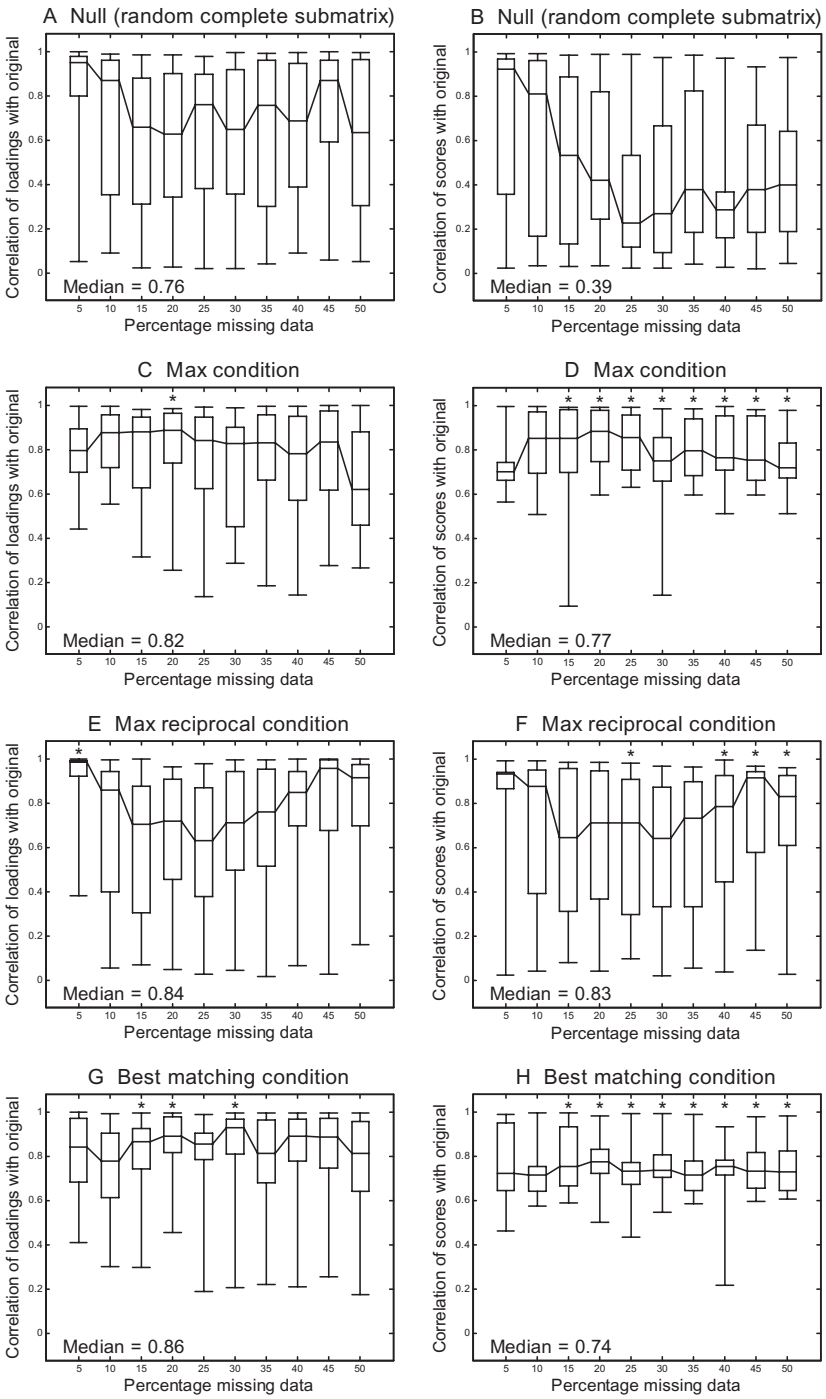

Figure 6. Results from simulations of missing completely at random data, based on principal component analyses of random subsets of the *Canis* data. Shown are distributions of correlations of character loadings (first column) and specimen scores (second column) from principal component analyses of random subsets with the original loadings and scores from the analysis of complete data (Fig. 5A). Box plots indicate the median, first and third quartiles, and range of each distribution. Asterisks indicate sampling distributions of correlations having medians significantly greater than the corresponding null distributions, based on a one-tailed Mann-Whitney test with $\alpha = 0.05$. See text for details. A, B, correlations of loadings and scores for random complete submatrices of data. C, D, correlations for complete submatrices having the maximum modified condition index. E, F, correlations for complete submatrices having the maximum modified reciprocal condition index. G, H, correlations for complete submatrices having the modified condition index best matching that of the original complete matrix.

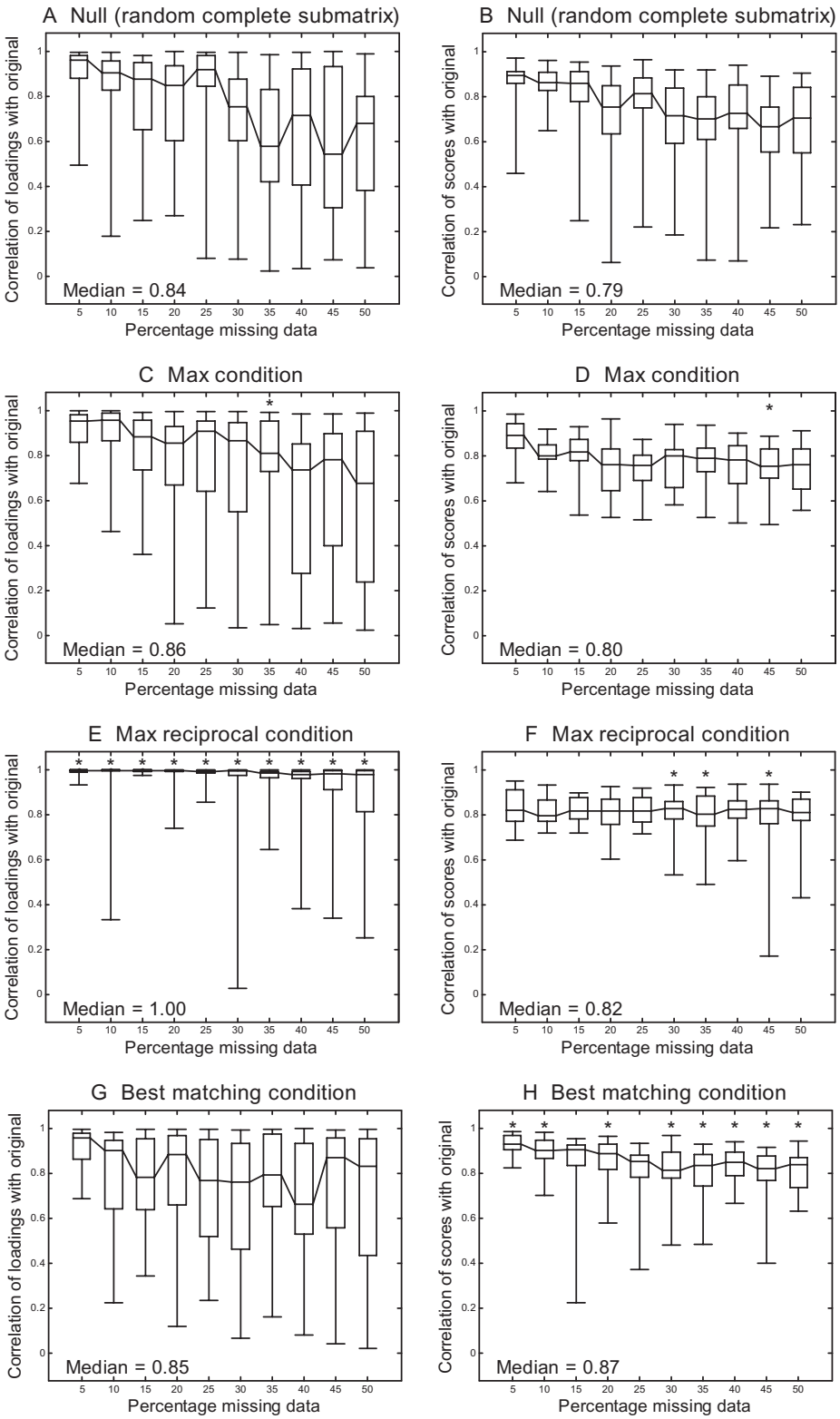

Figure 7. Results from simulations of missing completely at random data, based on discriminant analyses of random subsets of the *Canis* data. Shown are distributions of correlations of character loadings (first column) and specimen scores (second column) from discriminant function analyses of random subsets with the original loadings and scores from the analysis of complete data (Fig. 5B). Panels are as indicated in Figure 6.

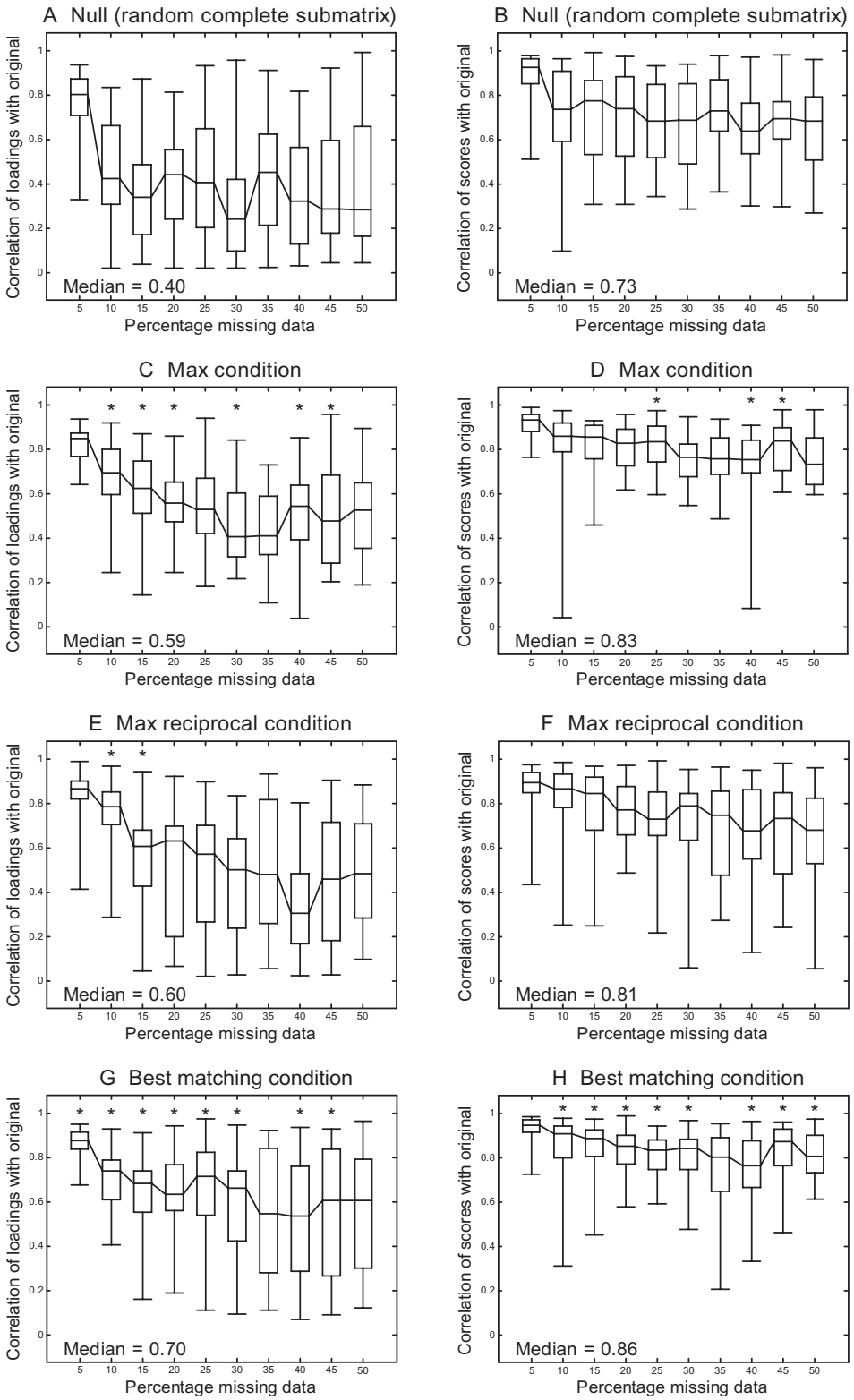

Figure 8. Results from simulations of missing completely at random data, based on principal component analyses of random subsets of the *Cottus* data. Shown are distributions of correlations of character loadings (first column) and specimen scores (second column) from principal component analyses of random subsets with the original loadings and scores from the analysis of complete data (Fig. 5C). Panels are as indicated in Figure 6.

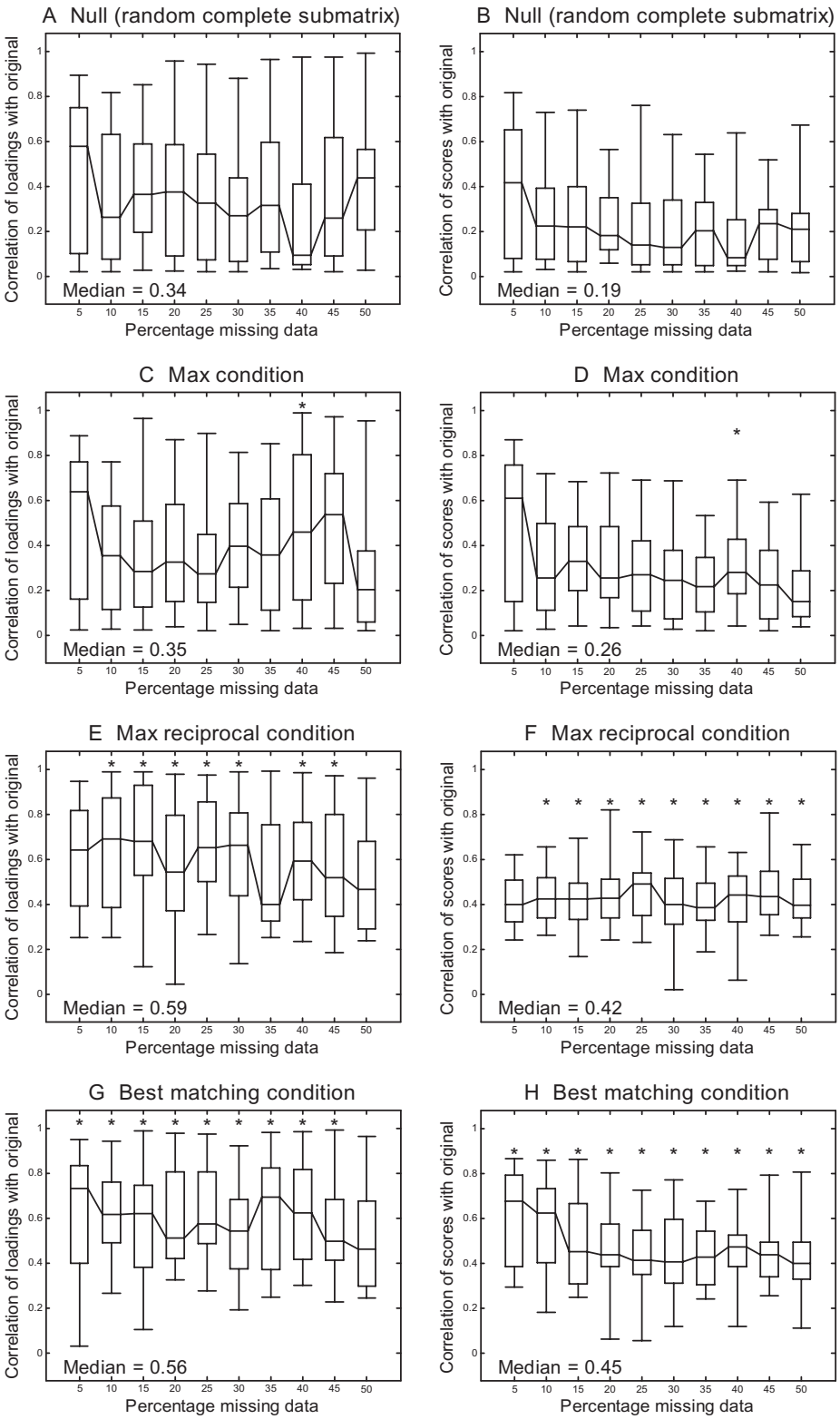

Figure 9. Results from simulations of missing completely at random data, based on discriminant analyses of random subsets of the *Cottus* data. Shown are distributions of correlations of character loadings (first column) and specimen scores (second column) from discriminant function analyses of random subsets with the original loadings and scores from the analysis of complete data (Fig. 5D). Panels are as indicated in Figure 6.

a set using one of a number of different methods. However, the question of how much missing data is too much is not easily answered. Despite the ample literature on missing-value estimation, there is still little empirical guidance for researchers. Each method is based on particular assumptions, often involving homogeneity of observations and characters, multivariate normality, and particular random distributions of missing values within the data set or particular processes hypothesized to have produced the missing values. Although some comparative Monte Carlo studies have been carried out to compare methods, these studies generally: (1) concern types of data (Roth, Switzer & Switzer, 1999) or methods (Bello, 1993a, 1995; Liu *et al*., 1997) that are not relevant to morphometric studies; (2) have been aimed at evaluating the effects of missing data on particular kinds of analyses (Basilevsky *et al*., 1985; Lee, 1986; van der Heijden, de Vries & van Hooff, 1990; Gornbein, Lazaro & Little, 1992; Lien & Rearden, 1992; Little, 1992; Twedt & Gill, 1992; Bello, 1993a; Scheiner, 1993; Brown, 1994; Carrano, Janis & Sepkoski, 1999); or (3) are based on unrealistically stringent assumptions.

We are currently carrying out sets of Monte Carlo simulations of missing-data prediction for biologically realistic kinds of data structure (Strauss *et al*., 2003). The results from these studies suggest that, for moderate numbers of variables, the EM method (Dempster *et al*., 1977) is both accurate and precise for relatively high proportions of missing data (up to almost 50% for data sets with few characters), even in the presence of group structure or character suites. However, for cases in which the proportion of missing data is sufficiently high that the investigator does not wish to include all characters and specimens, we suggest here a protocol for objectively identifying the 'best' specimens or characters to add to a subset of complete data for estimation purposes. The procedure begins with the submatrix having the best value for one of the optimization criteria (e.g. best matching condition) and sequentially adds specimens or characters until some specified maximum threshold of missing data (e.g. 10%) have been estimated. There are two possible choices of how to augment the initial submatrix: (1) because the statistical power of a multivariate analysis depends primarily on the ratio of numbers of specimens to characters (Tabachnick & Fidell, 2006), specimens can be added to the initial submatrix, holding the number of characters constant or (2) if the number of characters in the initial submatrix is too few to be meaningful for the purpose of the study, some combination of characters and specimens can be added. A Matlab function *addmiss*, which performs these operations, is available at the website cited above.

Adding specimens for a constant number of characters

It is generally advisable to maximize the number of observations relative to the number of variables for a multivariate analysis; a minimum of two or three observations per variable is often cited as a rule of thumb. Thus, an obvious procedure would be to select an initial submatrix of complete data having both adequate condition and a sufficient number of variables, and to add to it the specimens having small numbers of missing values. If the addition of all specimens does not exceed the specified maximum threshold of missing data, then all can be added. If the threshold is exceeded, however, then some subset of additional specimens must be selected. Although, in general, this might be the subset of specimens having the fewest missing values, in practice, this might not be the best solution because the addition of specimens can alter the covariance structure among the characters, which in turn affects the condition of the covariance matrix.

The addition of specimens can be done quasi-optimally in a stepwise fashion (analogous to the introduction of variables into a stepwise regression) by (1) adding a single specimen; (2) predicting the missing values for that specimen; and (3) calculating and recording the condition of the matrix with the addition of that specimen. (4) This is repeated for all specimens, and the particular specimen responsible for the largest condition index is identified. This specimen is then added to the initial submatrix and the procedure is repeated, adding specimens one at a time until the threshold of missing data has been reached. An alternate criterion would be to add specimens so long as the condition of the augmented matrix is 'better' than that of the initial complete submatrix. Finally, all of the missing values are then re-estimated as a set so as to be maximally consistent with one another, thus minimizing their prediction errors.

The effect of adding specimens in this way is illustrated by an analysis of the *Rhamphorhynchus* data (Figs 1A, 10). The original data matrix had 16 characters and 96 specimens, with 35.3% missing values. The complete submatrix having the maximum condition index, max C', has four characters and 61 specimens (although the best complete submatrix for some larger number of characters could be used instead). A ceiling of 15% missing data allowed in this case the addition of 23 specimens (Fig. 10A), for a total of 84 specimens, for which values were estimated via the EM algorithm. Although condition of the covariance matrix initially increases as the first few specimens are added, it then decreases with increasing number of specimens, to the point that, after 19 specimens have been added, the condition falls to below the estimated condition index of the complete submatrix. This occurs

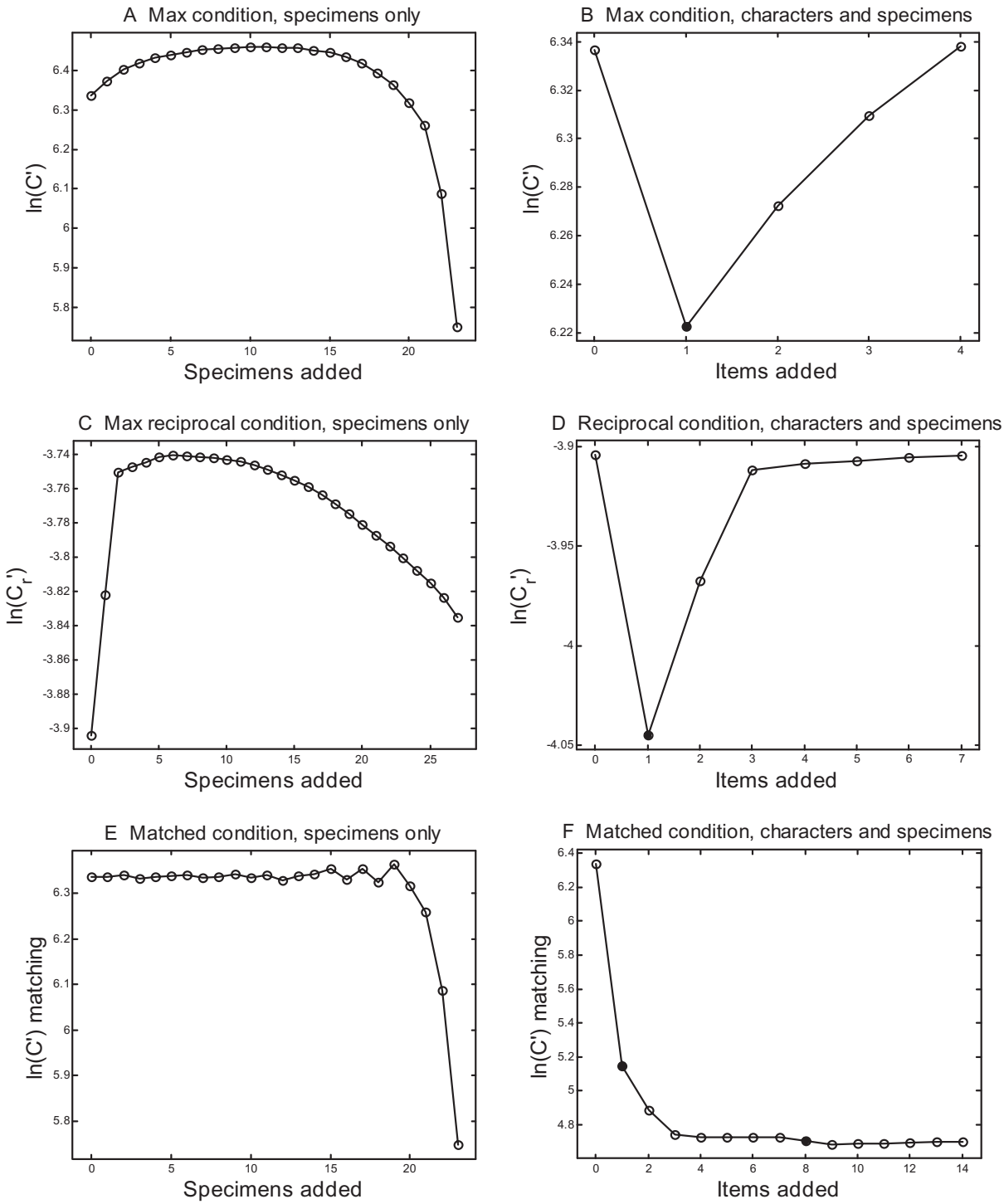

Figure 10. A, change in the maximum condition index as a function of adding, to the initial submatrix, specimens providing up to 15% missing data, for which values have been estimated via the EM algorithm. Data on *Rhamphorhynchus* are from Wellnhofer (1975). B, change in the maximum condition index as a function of adding both characters and specimens providing up to 15% missing data, for which values have been estimated via the expectation-maximization (EM) algorithm. Solid circles represent addition of characters, open circles the addition of specimens. C, D, as in (A) and (B), except showing change in the reciprocal condition index as a function of specimens or characters plus specimens. E, F, As in (A) and (B), except showing change in the best matching condition index as a function of specimens or characters plus specimens.

ostensibly because the additional information provided by the specimens does not offset the internal redundancy introduced by the estimated missing values. Note, however, that the change in condition due to adding specimens having missing data is small in absolute measure.

The results for adding specimens to the complete submatrix having the best reciprocal condition, max C'_r, are qualitatively similar (Fig. 10C), although the identities of the particular characters and specimens chosen are very different. The size of the best complete submatrix is three characters and 51 specimens, substantially smaller than in the preceding case. Addition of 'best' specimens produces an initial increase in C'_r, peaking after addition of six specimens, followed by a smooth decline in condition. In this case, 27 specimens have been added before the ceiling of 15% missing data is attained due to the smaller average number of missing values per specimen, but because of the smaller initial set of observations the final sample size is smaller (78 rather than 84). The absolute change in reciprocal condition is, again, very small.

When the complete submatrix is chosen by the best matching criterion, min $\Delta C'$, the results are qualitatively different from the preceding cases (Fig. 10E). The complete submatrix having the best matching condition index is quite different in size: 12 characters, but only 14 specimens. For the initial specimens added, the condition of the augmented matrix for which missing data are estimated continues to match that of the original complete submatrix. Only after 20 specimens have been added does the condition of the augmented matrix begin to degrade in relation to the original. This is a much more satisfying performance than with the max C' and max C'_r criteria, but the size of the final augmented matrix is quite different: 12 characters and 32 specimens. The disadvantage to this is that the ratio of specimens/characters is much lower; the advantage is that, even though the number of specimens is half as large as for the other optimization criteria, three- to four-fold as many characters have been included in the final matrix.

Adding characters and specimens concurrently

The procedure above can be modified slightly to add both characters and specimens in a stepwise fashion. First, a character is added to the initial submatrix by: (1) adding a single character; (2) predicting the missing values for that character; (3) calculating and recording the condition of the matrix with the addition of that character; and (4) repeating this procedure for all characters not already in the submatrix. The character responsible for the best condition index (according to one of the three optimization criteria) is then identified and added to the initial submatrix. Because the addition of a character for the same number of specimens will generally decrease the condition of the covariance matrix, one or more specimens can then be added, using the method described above, until the matrix condition has been restored or improved (or, alternately, can be added in proportion to the specimen-to-character ratio of the original data matrix, whichever is smaller). Another character and set of specimens can then be added in the same way, and the procedure is repeated until the threshold of missing data has been exceeded. Finally, all of the missing values are re-estimated as a set to minimize their prediction errors.

The second column of Figure 10 illustrates this procedure for the *Rhamphorhynchus* data. For the max C' criterion (Fig. 10B), the initial condition for the complete submatrix is as in Figure 10(A). Addition of character 4 (that character providing the greatest C' among all remaining characters) reduces the condition of the covariance matrix somewhat; addition of several 'best' specimens then restores the loss. Note that, since the best complete submatrix by this criterion has four characters but 61 specimens, the addition of a single character introduces a much larger proportion of missing data than does the addition of a single specimen. Thus, only four items (one character and three specimens) are added before the missing-data threshold is met, in contrast to the 23 specimens when only specimens are added. When the max C'_r criterion is applied to select the characters and specimens to be added (Fig. 10D), the results are qualitatively similar, although the identities of the particular characters and specimens selected are very different. The final matrix size under the 15% missing-data ceiling is four characters $\times$ 57 specimens, as compared to five characters $\times$ 64 specimens for the max C' criterion.

The best-matching condition criterion (Fig. 10F) again produces substantially different results. To the initial complete submatrix of 12 characters and 14 specimens are added two additional characters and 12 specimens. An interesting aspect of the pattern of augmentation is that, in contrast to the other optimization criteria, the addition of specimens apparently does not compensate for the decrease in condition due to the addition of characters. Although the explanation for this pattern is not apparent, analyses of several other data sets produce results qualitatively similar to this. Further investigation of this behaviour is warranted.

DISCUSSION

Although it is well known that multivariate analyses of morphometric data can be much more powerful and informative than univariate or bivariate analyses (Willig & Owen, 1987; Freeman & Jackson, 1990), such methods require complete data matrices and

therefore are often impractical with fossils or delicate extant material due to the presence of incomplete or missing structures. In the face of limited data, investigators commonly delete the characters or specimens from the analysis that sacrifice the fewest complete measurements. Although obvious (but wasteful) in principle, this procedure is complicated in practice by the fact that deletion of different subsets of characters and specimens may produce nearly the same amount of complete data but may have widely varying effects on subsequent analyses. There has been no previous systematic attempt to assess the ramifications of deleting varying subsets of data before multivariate analyses are carried out, other than trying a number of combinations and attempting to evaluate the 'robustness' of the results from different analyses. We have proposed one objective method for doing so, based on several possible measures of the statistical properties of the resulting data matrix. Although there is no guarantee that the reduced data matrix selected by such a criterion will produce the most meaningful biological interpretations, a well-conditioned matrix will at least provide the most stable statistical results and, if matched to the original matrix, will have the greatest likelihood of revealing the same underlying structure.

The examples that we have used involve taxonomic comparisons at the species and genus level. Regardless of the optimization criterion used, predictions of missing values are interpolations based upon correlations among the characters, and thus depend upon those correlations being biologically relevant. The correlations are biologically relevant if the characters measured among taxa are homologous in some sense (Rae, 1998; MacLeod, 2001; Guerrero, De Luna, & Sánchez-Hernández, 2003). Thus, although correlations are likely to be relevant at the population and species levels, they might or might not be relevant at higher taxonomic levels. If the data are appropriate for a morphometric analysis, they are also likely to be appropriate for predicting missing data.

The results from our simulations suggest that the best overall strategy in the face of substantial amounts of missing data is to choose the complete subset that best matches the estimated condition index of the full matrix. If the signal-to-noise ratio of the data is high (as in the *Canis* analyses), the best-matching criterion performs as well as, and often better than, a randomly selected subset of data. But if the structure is more subtle (as in the *Cottus* analyses), then the performance of the best-matching criterion is much better than random for both kinds of multivariate analysis considered in the present study. When using an initial complete submatrix as a basis to augment with a subset of imputed missing values, the best-matching criterion usually provides submatrices that are well-balanced in their inclusion of characters and specimens.

Two caveats are warranted. First, these simulations were carried out under a missing completely at random (MCAR) model, which might not be warranted for all data sets. However, Strauss *et al*. (2003) demonstrated that the presence of multiple groups of individuals or multiple character suites apparently does not markedly degrade the performance of several multivariate missing-data estimation methods, and thus might have at most minor effects on estimation of matrix condition in the presence of structure. Second, PCA and DFA are fairly straight-forward applications of eigenanalysis, and more complicated methods (e.g. canonical correlation, canonical correspondence analysis, or partial least squares regression) or noneigenanalysis methods (Bello, 1993a; Bello, 1993b; Kramer & Konigsberg, 1999) might be more sensitive to the criteria for selecting complete subsets of data.

The ability to predict or impute missing values based on the covariances among characters would seem to provide a potentially useful way to summarize relationships among taxa and characters in the presence of incomplete data. Every missing value that is estimated typically allows the introduction of many more 'good' values into the analysis (and thus increases the degrees of freedom for the actual values introduced, but not for those estimated from the existing data). Although imputation methods are well established in psychometrics and other scientific disciplines and in the clinical medical literature, their use in morphometric studies is uncommon and has remained largely untested (Gunz *et al*., 2002; Gauthier *et al*., 2003; Strauss *et al*., 2003). The stepwise methods that we have described here provide an initial, conservative approach to the estimation of missing data, based again on the principle of providing the most stable statistical results while attempting to minimize both the loss of actual data and the number of missing values estimated. For specimens and characters to be added to the reduced data matrix with this method not only must they have, on average, fewer missing values than others in the full data set, but also they should contribute to the robustness and internal consistency of the covariance matrix. This means that they will reinforce the patterns already present in the reduced data matrix and are unlikely to be multivariate outliers.

ACKNOWLEDGEMENTS

We thank S. C. Bennett for sending us his *Pteranodon* data set in electronic form. F. R. O'Keefe, a reviewer of a previous version of this manuscript, suggested the criterion of matching the covariance structure of complete submatrices to that of the original matrix.

REFERENCES

Allison PD. 2001. Missing data. *Sage University papers series on quantitative applications in the social sciences,* 07–136. Thousand Oaks, CA: Sage Publications.

Arbuckle JL. 1996. Full information estimation in the presence of incomplete data. In: Marcouldies GA, Schumacker RE, eds. *Advanced structural equation modeling techniques*. Mahwah, NJ: Lawrence Erlbaum Associates, 243–277.

Atanassov MN. 1996. Origin of dog *Canis familiaris* L. and early dog breeds of Bulgaria [in Bulgarian]. MSc Thesis, Sofia University.

Atanassov MN, Strauss RE. 1999. Morphometric analysis of *Rhamphorhynchus* from the Late Jurassic of southern Germany. *Journal of Vertebrate Paleontology* **19:** 30A–30A.

Atanassov MN, Strauss RE. 2000. Morphometric analysis of Late Jurassic pterodactyloids from Germany and France. *Journal of Vertebrate Paleontology* **20:** 27A–27A.

Basilevsky A, Sabourin D, Hum D, Anderson A. 1985. Missing data estimators in the general linear model: an evaluation of simulated data as an experimental design. *Communications in Statistics Series B (Simulation and Computation)* **14:** 371–394.

Beale EML, Little RJA. 1975. Missing values in multivariate analysis. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* **37:** 129–145.

Bello AL. 1993a. A simulation study of imputation techniques in linear, quadratic and kernel discriminant analyses. *Journal of Statistical Computation and Simulation* **48:** 167–180.

Bello AL. 1993b. Choosing among imputation techniques for incomplete multivariate data: a simulation study. *Communications in Statistics Series A (Theory and Method)* **22:** 853–877.

Bello AL. 1995. Imputation techniques in regression analysis: looking closely at their implementation. *Computational Statistics and Data Analysis* **20:** 45–57.

Bennett SC. 1991. Morphology of the Late Cretaceous pterosaur Pteranodon and systematics of the Pterodactyloidea. PhD Dissertation, University of Kansas.

Bennett SC. 1995. A statistical study of *Rhamphorhynchus* from the Solnhofen Limestone of Germany: year-classes of a single large species. *Journal of Paleontology* **69:** 569–579.

Bennett SC. 1996. Year-classes of pterosaurs from the Solnhofen Limestone of Germany: taxonomic and systematic implications. *Journal of Vertebrate Paleontology* **16:** 432–444.

Brown RL. 1994. Efficacy of the indirect approach for estimating structural equation models with missing data: a comparison of five methods. *Structural Equation Modeling* **1:** 287–316.

Carrano MT, Janis CM, Sepkoski JJ Jr. 1999. Hadrosaurs as ungulate parallels: lost lifestyles and deficient data. *Acta Palaeontologica Polonica* **44:** 237–261.

Chen Y, McInroy JE. 2002. Estimation of symmetric, positive definite matrices from imperfect measurements. *IEEE Transactions on Automatic Control* **47:** 1721–1725.

Cole DA, Maxwell SE, Arvey R, Salas E. 1994. How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin* **115:** 465–474.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B (Statistical Methodology)* **39:** 1–38.

Essl A. 1991. Choice of an appropriate bending factor using prior knowledge of the parameters. *Journal of Animal Breeding and Genetics* **108:** 89–101.

Flury BK. 1988. *Common principal components and related multivariate models*. New York, NY: Wiley.

Fomby T. 1998. *Messy data: missing observations, outliers, and mixed-frequency data*. New York, NY: Elsevier.

Frane JW. 1976. Some simple procedures for handling missing data in multivariate analysis. *Psychometrika* **41:** 409–415.

Freeman S, Jackson WM. 1990. Univariate metrics are not adequate to measure avian body size. *Auk* **107:** 69–74.

Gauthier O, Landry PA, Lapointe FJ. 2003. Missing data in craniometrics: a simulation study. *Acta Theriologica* **48:** 25–34.

Gornbein JA, Lazaro CG, Little RJA. 1992. Incomplete data in repeated measures analysis. *Statistical Methods in Medical Research* **1:** 275–295.

Groß J. 2003. *Linear regression. Lecture notes in statistics 175*. Berlin: Springer-Verlag.

Guerrero JA, De Luna E, Sánchez-Hernández C. 2003. Morphometrics in the quantification of character state identity for the assessment of primary homology: an analysis of character variation of the genus *Artibeus* (Chiroptera: Phyllostomidae). *Biological Journal of the Linnean Society* **80:** 45–55.

Gunz P, Mitteroecker P, Bookstein FL, Weber GW. 2002. Approaches to missing data in anthropology. *Collegium Antropologicum* **26:** 78–79.

Hayes JF, Hill WG. 1980. A reparameterisation of a genetic selection index to locate its sampling properties. *Biometrics* **36:** 237–248.

Hayes JF, Hill WG. 1981. Modification of estimates of parameters in the construction of genetic selection indices ('bending'). *Biometrics* **37:** 483–493.

van der Heijden PGM, de Vries H, van Hooff JA. 1990. Correspondence analysis of transition matrices, with special attention to missing entries and asymmetry. *Animal Behaviour* **40:** 49–64.

Higham NJ. 1989. Matrix nearness problems and applications. In: Gover MJC, Barnett S, eds. *Applications of matrix theory*. Oxford: Oxford University Press, 1–27.

Higham NJ. 2002. Computing the nearest correlation matrix – a problem from finance. *IMA Journal of Numerical Analysis* **22:** 329–343.

Hill WG, Thompson R. 1978. Probabilities of non-positive definite between-group or genetic covariance matrices. *Biometrics* **34:** 429–439.

Houck MA, Gauthier JA, Strauss RE. 1990. Allometric scaling in the earliest fossil bird, *Archaeopteryx lithographica*. *Science* **247:** 195–198.

Hu H. 1995. Positive definite constrained least-squares esti-

mation of matrices. *Linear Algebra and its Applications* **229:** 167–174.

Jolliffe IT. 1982. A note on the use of principal components in regression. *Journal of the Royal Statistical Society Series C (Applied Statistics)* **31:** 300–303.

Jorjani H, Klei L, Emanuelson U. 2002. Combining disparate estimates of genetic correlations. *Interbull Bulletin* **29:** 1–3.

Jorjani H, Klei L, Emanuelson U. 2003. A simple method for weighted bending of genetic (co)variance matrices. *Journal of Dairy Science* **86:** 677–679.

Kaiser HF, Dickman K. 1962. Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika* **27:** 179–182.

Kinziger AP, Raesly RL, Neely DA. 2000. New species of *Cottus* (Teleostei: Cottidae) from the middle Atlantic eastern United States. *Copeia* **2000:** 1007–1018.

Knol DL, Ten Berge JMF. 1989. Least-squares approximation of an improper correlation matrix by a proper one. *Psychometrika* **54:** 53–61.

Kramer A, Konigsberg LW. 1999. Recognizing species diversity among large-bodied hominoids: a simulation test using missing data finite mixture analysis. *Journal of Human Evolution* **36:** 409–421.

Kupiec PH. 1998. Stress testing in a value at risk framework. *Journal of Derivatives* **1998** (**3**)**:** 7–25.

Lee WT. 1986. Estimation for structural equation models with missing data. *Psychometrika* **51:** 93–99.

Lien D, Rearden D. 1992. A note on estimating regression coefficients with missing data. *Econometric Reviews* **11:** 119–122.

Little RJA. 1992. Regression with missing X's: a review. *Journal of the American Statistical Association* **87:** 1227–1237.

Little RJA, Rubin DB. 1987. *Statistical Analysis with Missing Data*. New York, NY: Wiley.

Liu WZ, White AP, Thompson SG, Bramer MA. 1997. Techniques for dealing with missing values in classification. *Lecture Notes in Computer Science* **1280:** 527–536.

Lucas C. 2001. Computing nearest covariance and correlation matrices. MS Thesis, University of Manchester.

MacLeod N. 2001. Landmarks, localization, and the use of morphometrics in phylogenetic analysis. In: Edgecombe G, Adrain J, Lieberman B, eds. *Fossils, phylogeny, and form: an analytical approach*. New York, NY: Kluwer Academic/Plenum, 197–233.

Mathworks. 1997. *Matlab reference guide*. Natick, MA: Mathworks.

Millis SR. 2003. Statistical practices: the seven deadly sins. *Child Neuropsychology* **9:** 221–233.

Proschan MA, McMahon RP, Shih JH, Hunsberger SA, Geller NL, Knatterud G, Wittes J. 2001. Sensitivity analysis using an imputation method for missing binary data in clinical trials. *Journal of Statistical Planning and Inference* **96:** 155–165.

Rae TC. 1998. The logical basis for the use of continuous characters in phylogenetic systematics. *Cladistics* **14:** 221–228.

Rebonato R. 1999. *Volatility and correlation*. New York, NY: John Wiley and Sons.

Roth PL, Switzer FS, Switzer DM. 1999. Missing data in multiple item scales: a Monte Carlo analysis of missing data techniques. *Organizational Research Methods* **2:** 211–232.

Rubin DB. 1976. Inference with missing data. *Biometrika* **63:** 581–592.

Rubin DB. 1996. Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91:** 473–489.

Schafer JL, Olsen MK. 1998. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research* **33:** 545–571.

Scheiner SM. 1993. MANOVA: multiple response variables and multispecies interactions. In: Scheiner SM, Gurevitch J, eds. *Design and analysis of ecological experiments*. New York, NY: Chapman & Hall, 94–112.

Strauss RE. 1989. Associations between genic heterozygosity and morphological variability in freshwater sculpins, genus *Cottus* (Teleostei: Cottidae). *Biochemical Systematics and Ecology* **17:** 333–340.

Strauss RE. 1991. Correlations between heterozygosity and phenotypic variability in *Cottus* (Teleostei: Cottidae): character components. *Evolution* **45:** 1950–1956.

Strauss RE, Atanassov MN, Oliveira JA. 2003. Evaluation of the principal-component and expectation-maximization methods for estimating missing data in morphometric studies. *Journal of Vertebrate Paleontology* **23:** 284–296.

Tabachnick BG, Fidell LS. 2006. *Using Multivariate Statistics*, 5th edn. Boston, MA: Allyn & Bacon.

Twedt DJ, Gill DS. 1992. Comparison of algorithms for replacing missing data in discriminant analysis. *Communications in Statistics Series A (Theory and Method)* **21:** 1567–1578.

Wellnhofer P. 1970. Die Pterodactyloidea (Pterosauria) der Oberjura-Plattenkalke Süddeutschlands. *Bayerische Akademie der Wissenschaften, Abhandlungen* **141:** 1–133.

Wellnhofer P. 1974. Das fünfte Skelettexemplar von *Archaeopteryx*. *Palaeontographica, Abteilung A (Paläozoologie, Stratigraphie)* **147:** 169–216.

Wellnhofer P. 1975. Die Rhamphorhynchoidea (Pterosauria) der Oberjura-Plattenkalke Suddeutschlands. *Palaeontographica, Abteilung A (Paläozoologie, Stratigraphie)* **148:** 1–33.

Wellnhofer P. 1988. Ein neues Exemplar von *Archaeopteryx*. *Archaeopteryx* **6:** 1–30.

Wellnhofer P. 1993. Das siebte Exemplar von *Archaeopteryx* aus den Solnhofener Schichten. *Archaeopteryx* **11:** 1–47.

Wilks SS. 1932. Moments and distributions of estimates of population parameters for fragmentary samples. *Annals of Mathematical Statistics* **3:** 165–195.

Willig MR, Owen RD. 1987. Univariate analyses of morphometric variation do not emulate the results of multivariate analyses. *Systematic Zoology* **36:** 398–400.