Mathematical Geology, Vol. 36, No. 5, July 2004 (© 2004)

Instability in Principal Component Analysis and the Quantification of Polyphenism in Palaeontological Data[1]

Richard A. Reyment[2]

The occurrence of cryptic polyphenism (variation in morphological properties within a single species) in ammonites is used to exemplify the application of the multivariate set of techniques known in analytical chemistry as cross-validation to quantify and isolate deviating specimens (ecomorphs) in a genetically homogeneous sample. A byproduct of the analysis bears on a method of identifying redundant variables. A species of Nigerian Turonian (Cretaceous) ammonites of the genus Thomasites *is used in the exemplification.*

KEY WORDS: cross-validation, principal component analysis, ecomorphs, ammonites.

INTRODUCTION

The application of methods of multivariate statistical analysis in the Earth Sciences tends at times to follow an uncritical approach. The ready availability of programmed computational procedures can easily lull the unwary user into accepting that the processing of a data-set will be correct and that the results that issue therefrom are a true picture of the statistical relationships. All too often, however, deviations from multivariate Gaussian are not recognized by the practitioner. In the normal course of events, analytical instability due to heterogeneity in a data-set is not a desirable property. The reasons for this are that any attempt at reifying the latent vectors of a principal component extraction cannot lead to a useful result if the chance composition of the sample has a decisive influence on the magnitudes of these vectors. In this note I show that in biology (including palaeontology), at least, heterogeneity in data can actually be used to enhance the biological value of a study with respect to disclosing phenotypic variation in morphological characteristics. Many groups of animals display a wide range of

[1]Received 30 July 2003; accepted 4 March 2004.
[2]Paleozoologiska avdelningen, Naturhistoriska Riksmuseet, Box 50007, 10405 Stockholm, Sweden; e-mail: richard.reyment@nrm.se

0882-8121/04/0700-0629/1 © 2004 International Association for Mathematical Geology

variability such that it is not always obvious that divergent morphological manifestations are genuine members of a taxon. A well known case among living organisms concerns the shell-shape of the acorn barnacle (Lively, 1986), which is controlled by the intensity of predation and the nature of the predator (resulting from preferential predation on a particular shell-morph). Ostracod shells often display polymorphism in shape and size (Reyment, 1991). If this condition in an ostracod species is not recognized then a statistical analysis of a sample containing two or more morphs cannot be invested with much credibility.

The data used for illustrating the present note are taken from a study of Nigerian Turonian (Upper Cretaceous) ammonites (Reyment, 2003) of the trans-Saharan tectono-eustatic epicontinental transgression (the prototype of classical Suessian eustacy). During late Cenomanian time, probably one ammonite species entered the Saharan realm from the Tethys. In reaction to the special ecological conditions pertaining in the long, shallow transcontinental sea, a speciation event took place the effects of which, however, were greatly complicated by ecophenotypic interaction (Reyment, 2003). In the past this has led to a proliferation in the number of species and generic names applied to the morphological variations (Barber, 1957; Pervinquière, 1907). More recent work has recognized this artificial taxonomy (e.g. Meister, 1989). Two schools of thought have crystallized over the last few years concerning the evolutionary relationships in the northern Nigerian forms. One claims that all of the morphs belong to a single highly variable species, to wit, *Thomasites gongilensis* (Woods). The second interpretation is that there are two closely related species, *Thomasites* (*Thomasites*) *gongilensis* (Woods) and *Thomasites* (*Bauchioceras*) *nigeriensis* (Woods). Under the hypothesis that all specimens in the sample derive from the same multivariate statistical universe, it can be reasonably expected that no clearly marked "outliers" will show up in the analyses. If, however, the data-set is statistically heterogeneous, and hence not taxonomically unique, appropriately constructed multivariate procedures will identify the divergent individuals.

THE DATA

The data used to illustrate the procedure consist of seven distance-measures made on the shells of specimens of *Thomasites* and its subgenus *Bauchioceras* from the lowermost Turonian (Cretaceous) of northeastern Nigeria and equivalent strata in Tunisia. These are (1) maximum diameter of the conch, (2) maximum breadth of the shell, (3) maximum breadth across the venter, (4) maximum ventral breadth of the whorl at two right-angles from the distal ventral extremity, (5) maximum ventral breadth of the penultimate whorl, (6) maximum height of the final whorl, (7) maximum umbilical breadth. The sample comprises 16 specimens. The locations of the measures on *Thomasites* (*Thomasites*) are shown in Figure 1. Sketches of typical ventral shapes are shown in Figure 2.

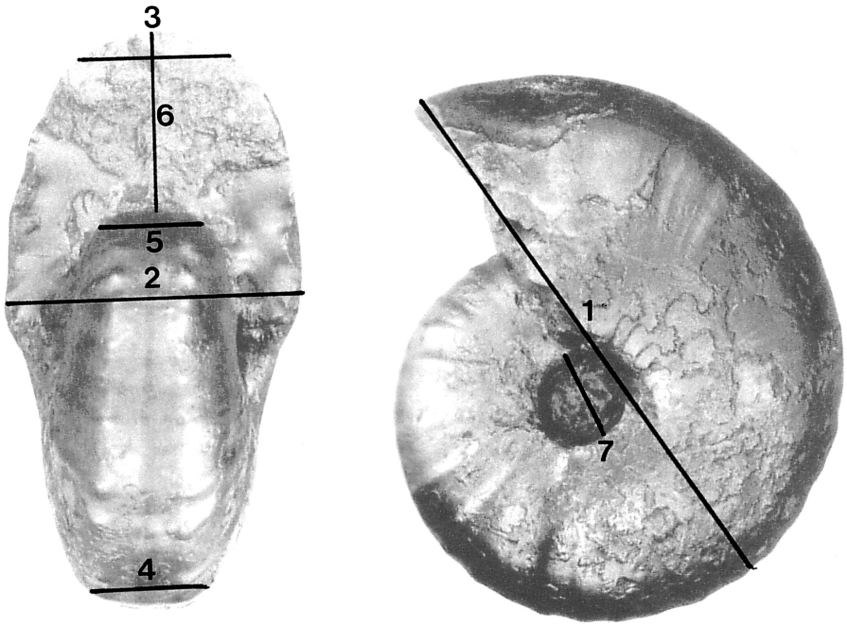


Figure 1. The locations of the variables measured on the ammonite shells.

The biological question of interest is whether any of the specimens of the sample deviate essentially from the main body of the data (as atypical and, or, influential observations). The presence of markedly deviant specimens may disclose polymorphism or polyphenism in the sample. In many cases, a simple bivariate scatter-plot can reveal aberrant observations, but this is not always true. An appropriate multivariate analysis will usually do better, but even then some atypical individuals may escape detection. This is not always a serious matter except in the study of the population dynamics of fossil and living populations.

PRINCIPAL COMPONENT ANALYSIS AND CROSS-VALIDATION

Analytical procedures used here are on the basis of Krzanowski (1982, 1987, 1988), including the data-analytical concept of cross-validation (Wold, 1978): Computational details are provided in Reyment and Savazzi (1999, pp. 120–129). The main questions posed by this synthesis are the following:

1. Can any variables be excluded on the grounds of redundancy?
2. Which observations deviate in some multivariate manner from the main body of the data?

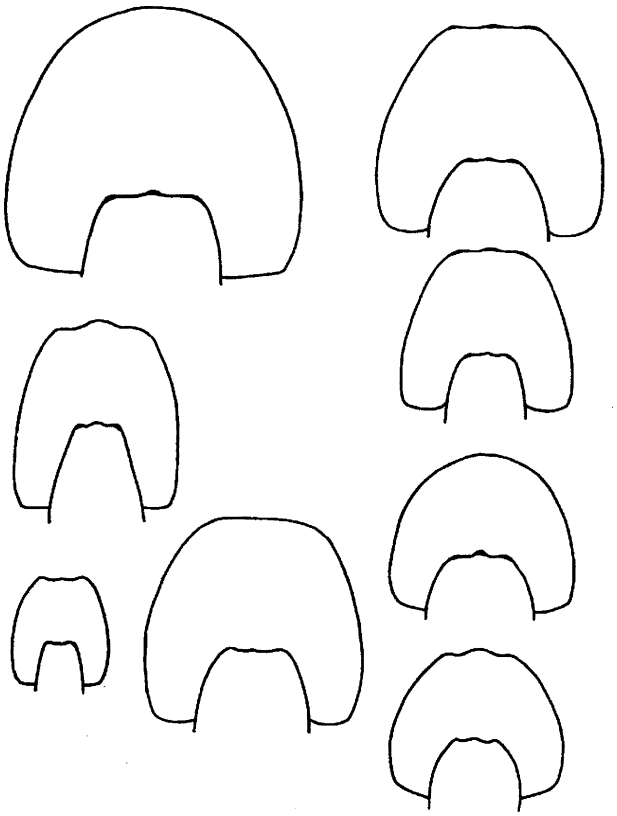


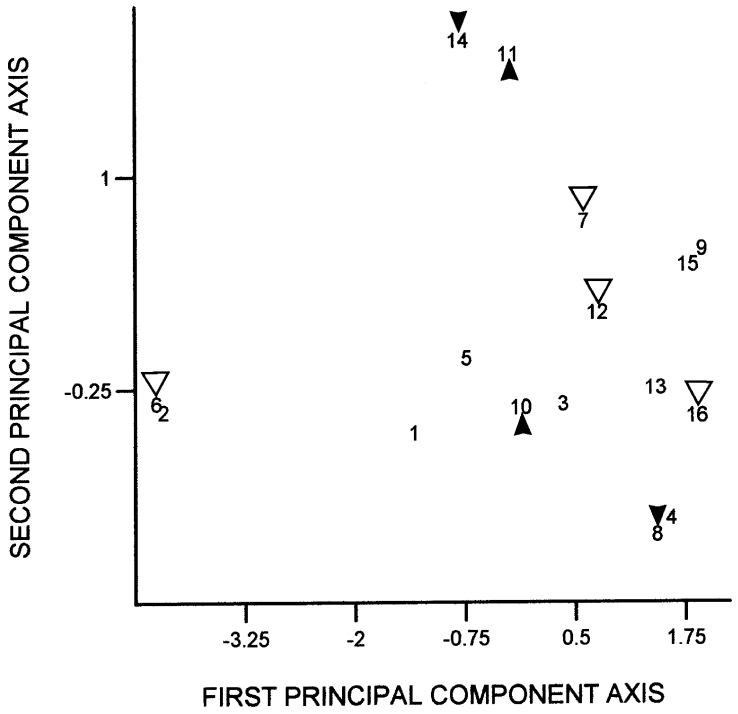
Figure 2. Examples of typical ventral configurations of the *Thomasites* morphological complex.

Two types of divergent values are here of interest. An *atypical observation* is one that differs markedly from the rest of the sample with respect to the measures on the set of traits. It can often be readily picked out from inspection of a multivariate ordination diagram. An *influential observation* is one that causes a marked change in the analysis when it is excluded, but the measures observed on it show no visible divergencies.

Cross-validation is basically an exploratory technique that looks for interesting patterns and scans a data matrix for redundant components. The steps in the calculations are:

1. Compute the principal components of the covariance matrix **S** or the correlation matrix **R** of the $(n \times p)$ data matrix **X**:

$$\mathbf{S} = \mathbf{VLV}' \tag{1}$$



A4 **Figure 3.** Plot of the first and second principal component scores for all variables. *Key*: The filled upward pointing triangles denote the specimens (10 and 11) deviating with respect to both variance and correlation. Inverted filled triangles denote specimens deviating with respect to variance. Open inverted triangles denote specimens deviating with respect to correlation.

with the usual scaling $\mathbf{V}'\mathbf{V} = \mathbf{1}$. Alternatively, the singular value decomposition of $\mathbf{X}$ can be used

$$\mathbf{X} = \mathbf{UDV}' \tag{2}$$

noting that $d_i^2 = (n-1)l_i$ $(i = 1, \ldots, p)$.

2. Compute the scores of the principal components:

$$\mathbf{Z} = \mathbf{XV}'$$

If the data-structure is essentially m-dimensional, then the variation in the remaining $(p - m)$ dimensions can often be disregarded.

3. Determine a criterion denoted W_m by Krzanowski (1987). This criterion is computed from the average squared discrepancy between the actual

Table 1. Latent Roots of Correlation Matrix for the Septivariate Ammonite Data

	Latent roots						
Latent vectors	5.5314	0.7159	0.4247	0.2047	0.0726	0.0368	0.0139
Component 1	0.3742	0.3778	0.4034	0.3657	0.4044	0.3570	0.3602
Component 2	0.5068	0.0626	−0.1391	−0.5033	−0.0756	0.5552	−0.3907
Component 3	−0.1443	0.6247	0.3129	−0.1774	0.1987	−0.3690	−0.5328
Component 4	0.2620	−0.2577	0.3938	0.5375	−0.4429	−0.0179	−0.4737
Component 5	−0.0986	0.6221	−0.2794	0.1731	−0.6628	0.1470	0.1854
Component 6	0.6930	0.0471	−0.0858	−0.1536	−0.1312	−0.6344	0.2587
Component 7	−0.1540	−0.0864	0.6925	−0.4889	−0.3723	0.0611	0.3286

A3

and predicted values of the data matrix. This is done by the method of cross-validation as follows.

Divide **X** into several groups. Then delete each group in turn from the data matrix **X** and compute the values of the predictor from the remaining data, and then predict the deleted values. In practice, the deleted group can be conveniently made to be just one row of **X** (i.e., a single individual). The manner in which the method is usually applied involves deletion of variables as well as individuals (i.e., columns as well as rows of the data-matrix).

4. Informative variables: The comparison of two m-dimensional configurations of the same n points may be conveniently done by applying Procrustean analysis (Gower, 1971; Schönemann and Carroll, 1970), in which the sum of distances is found between corresponding points of the two configurations, after matching under translation, rotation, and reflection. A large sum of squares on the omission of variable x_i indicates a discrepancy between the two configurations and hence suggests that x_i is an important variable. A small sum of squares obtained on the omission of a variable, on the other hand, indicates a close match between the two configurations thus suggesting that this variable can be ignored without undue loss of efficiency. Table 2 lists the residual sums of squares when the 16 points on the first principal component using all variables are matched successively by means of "Procrustean Analysis" applied to the 16 points on the first principal component obtained from the four (16×6) matrices formed by deleting each variable in turn. The second column is the same as above but now matches points on the first two principal components with the first two principal component projections on deleting each variable in turn. Likewise for the third column.

Deleting variables 3 and 5 yields the smallest residual sums of squares in all three columns. Both of these variables are measures of ventral width. There may be a simple reason for this because the standard errors

Table 2. Identification of Redundant Variables by Deleting Each Variable in Turn. Residual Sums of Squares for the Procrustean Fit of New Scores to Old Scores

Variable removed from the analysis	Principal component-spaces examined		
	$P1$	$P1 + P2$	$P1 + P2 + P3$
1	1.0879	2.5482	1.6797
2	1.0117	1.0381	2.6641
3	**0.9244**	**1.0006**	**1.1960**
4	1.0814	2.5938	2.1174
5	**0.8999**	**0.9137**	**1.0028**
6	1.0949	4.2699	2.0639
7	1.0570	2.2256	3.2349

Note. Bold entries denote deletions that do not perturb the principal component residuals thus making them possible candidates for omission from the analysis. *P* denotes "principal component."

for the variables computed by jacknifing are very large and of about equal magnitude (0.4276 for variable 3 and 0.4877 for variable 5). This is probably a reflection of the practical difficulty of obtaining consistent measures of the distance across the venter of the shell-types with a rounded venter (but not the morphs with a flattened or truncated venter—cf. Fig. 2).

5. Determine the influence of each specimen: Proceed by computing *critical angles*. The critical angle is a measure of influence of each individual in the sample, with $t = \arccos(\tau)$ (where τ denotes the smallest element of the diagonal matrix $\mathbf{L}$ of Eq. (1)). Large values of the critical angle denote highly *influential* observations in the sample. The influence of each of the specimens in the sample may be gauged as follows. Let $\hat{\mathbf{U}}$ contain the principal component coefficients when the ith row of $\mathbf{X}$ is deleted. A comparison of this matrix with the original $\mathbf{U}$ will indicate the influence that the ith individual has on the outcome of the analysis. The comparison is made by computing the *critical angles* between principal component subspaces of a common data space, **A2**

$$\psi = \arccos(d)$$

where d is the smallest element of $\mathbf{D}$ (diagonal) in the singular value decomposition of $\mathbf{X}$ (Eq. (2)) with respect to the optimal dimensionality (Krzanowski, 1988, pp. 167–176).

The maximum critical angle ψ is a suitable measure of influence of each individual in the sample. The larger the value of this angle, the greater is the perturbation to the principal component caused by the deletion of the corresponding specimen and hence an indicator of the importance of that specimen in respect of its

departure from the dominant morphological expression prevailing in the sample. There is no established test for assessing the criticality of an angle. Krzanowski (1988) suggests as a guideline, based on experience, values of 10 and in excess thereof as indicative of significance. My personal experience suggests that the question of whether or not a critical angle expresses a genuine deviation must be interpreted in relation to the average of all the angles of the sample. This notwithstanding, in the present study a value of 10 turns out to be quite reasonable. In Table 3, the results for the maximum critical angles are given together with a comment on the shell-shape, that is, the morphological variant typifying each specimen.

Table 3 contains six columns of computational results. The columns 1, 2, and 3 denote the largest critical angle for each specimen, deleted in turn, for one, two, and three principal components respectively in decreasing order of magnitude of the associated latent root. The figures printed in bold type denote those units which, when omitted from the analysis, cause the greatest perturbations. The three columns bearing negative signs are for the three smallest principal components. The motivation for the bipartite form of the table of critical angles is that the largest principal components identify dispersion outliers (roughly in line with the major ellipsoidal axis), whereas the smallest principal components often identify correlation outliers (roughly in line with the minor ellipsoidal axis).

Repeating the computations with variables 3 and 5 omitted, caused an increase in homogeneity of the morphologies and more closely approximated by a multivariate Gaussian model. There are now only two "correlation atypicalities," specimens 10 and 11 and two "variance atypicalities," specimens 11 and 14. Specimens 10 and 11 produced both atypicalities in the septivariate analysis whereas in the quinquevariate set, only specimen 11 displays the two properties. It is significant that specimen 11 is the most markedly deviant shell-type in both appraisals (seven variables and five variables). This result is again a reflection of the role of ventral breadth as a marker of apertural shape polymorphism in these ammonites. Although, as already noted, variables 3 and 5 could not be measured at the same level of accuracy as the others on all specimens, it would be wrong to remove them from the analysis owing to their diagnostic morphological and taxonomic significance.

The plot of the scores for the first and second components shows that there is a lack of cohesion in shape-discrimination. The specimens identified as being atypical do not show up in the standard principal components plot at all. The conclusion suggested by the analysis is that there are several morphometrically characterized morphs in the sample, some of which are qualified by deviations occurring in the largest principal components, variation-sensitive ones, and others in the smallest principal components, correlation-sensitive ones. In only two cases, specimens 10 and11, an evolute morph, are both categories represented.

Table 3. Maximum Critical Angles Obtained From the Successive Deletion of Specimens

	Largest principal component (variance sensitive)			Smallest principal component (correlation sensitive)			Shape characterization
Deleted sp.	1	2	3	−3	−2	−1	"Ecomorphs"
1	0.9627	2.1867	2.1940	3.3645	3.3972	9.4564	Semiglobose
2	4.1401	4.9415	2.7651	2.5130	3.0913	2.3679	Semiglobose
3	0.1564	0.6207	0.5871	0.6991	1.4268	0.2040	Semiglobose
4	2.0181	9.3592	7.9128	4.5258	4.5033	6.3299	"Tectiform" & semiglobose
5	0.2929	0.4160	0.9139	1.7699	1.7474	0.5452	"Tectiform" & semiglobose
6	3.5802	3.5961	3.7162	4.0806	**17.8207**	**23.2104**	Globose
7	0.7229	5.2821	7.4871	**12.0428**	**10.9215**	8.3897	Globose
8	1.4566	**18.1667**	5.9262	5.7950	6.0606	4.8740	Globose
9	1.8659	4.0718	4.8630	6.7362	9.4318	9.4970	Compressed involute
10	0.5883	7.8695	**17.4664**	**10.3253**	**12.6431**	**16.9753**	Evolute
11	1.2133	**22.3242**	**40.5280**	**52.3826**	**35.7912**	**21.9959**	Evolute
12	0.4245	0.8753	3.3729	5.7134	**23.3753**	**17.1583**	Tricarinate compressed
13	1.2743	1.9906	5.4211	4.5679	4.6352	1.1200	Compressed involute
14	0.7022	**34.9400**	**10.4666**	5.4032	5.7945	0.3039	Compressed involute
15	1.8783	4.1978	5.4054	5.2375	5.0180	3.0962	Compressed involute
16	1.4724	2.0782	2.0942	3.8427	**13.3583**	**11.1234**	Evolute

Note. Bold entries denote specimens which when removed produced a large critical angle whereby indicating influential morphometrically deviating specimens.

FINAL REMARKS

Cross-validation is a synthesis of ad hoc techniques that were originally proposed by analytical chemists. The application to biological problems has passed smoothly and, more recently, the fields of geochemistry, morphometrics, psychometry, agronomy, and genetics have been added. So far, the technique has been of a rather informal nature in that exact tests have yet to be devised for establishing significance criteria. In the present analysis, it could be demonstrated that a major factor controlling the morphological variability lies with the development of the ventral configuration of the shell and the degree of evolution of the whorls.

ACKNOWLEDGMENT

I thank Professor John C. Davis for valuable comments which greatly improved the scope of this paper.

REFERENCES

Barber, W. M., 1957, The lower Turonian ammonites of northeastern Nigeria: Bull. Geol. Surv. Niger., v. 26, 86 p.

Gower, J. C., 1971, Statistical methods of comparing different multivariate analyses of the same data, *in* Hodson, F. R., Kendall, D. G., and Tautu, P., eds., Mathematics in the archaeological and historical sciences: University Press, Edinburgh, p. 138–149.

Krzanowski, W. J., 1982, Between-group comparison of principal components—some sampling results: J. Stat. Comput. Sim., v. 15, p. 141–154.

Krzanowski, W. J., 1987, Selection of variables to preserve multivariate data structure, using principal components: Appl. Stat., v. 36, p. 22–33.

Krzanowski, W. J., 1988, Principles of multivariate analysis. A User's Perspective: Oxf. Stat. Sci. Ser., v. 3, 563 p.

Lively, C., 1986, Predator-induced shell dimorphism in the Acorn Barnacle *Chthamalus anisopoma*: Evolution, v. 40, p. 232–242.

Meister, C., 1989, Les ammonites du Crétacé supérieur d'Ashaka, Nigéria: Bull. Cent. Rech. Explor.-Prod. Elf-Aquitaine, v. 13, no. Suppl., p. 1–84.

Pervinquière, L., 1907, Etudes de paléontologie tunisienne. 1. Céphalopodes des terrains secondaires: Mém. carte géol. Tunisie, Paris, 483 p.

Reyment, R. A., 1991, Multidimensional palaeobiology: Pergamon Press, Oxford, 377 p.

Reyment, R. A., 2003, Morphometric analysis of variability in the shell of some Nigerian (Cretaceous) ammonites: Cretaceous Res., v. 24, p. 789–803.

Reyment, R. A., and Savazzi, E., 1999, Aspects of multivariate statistical analysis in geology: Elsevier, Amsterdam, 285 p. (Compact disc of programs included)

Schönemann, P., and Carroll, R. M., 1970, Fitting one matrix to another under choice of a central dilation and rigid motion: Psychometrika, v. 35, p. 245–255.

Wold, S., 1978, Cross-validatory estimation of the number of components in factor and principal component models: Technometrics, v. 20, p. 397–405.

Queries to Author:

A1: Au: Kindly check the RRH. Is it OK?
A2: Au: Kindly note that “arcos” has been changed to “arccos” throughout. Is it OK?
A3: Au: please cite the table in the text.
A4: Au: kindly cite the figure in the text.