Z. geol. Wiss., Berlin 40 (2012) 4/5: 213 – 217

On the interpretation of the smallest principal component in geology

Zur Bedeutung der kleinsten Hauptkomponente in den Geowissenschaften

RICHARD A. REYMENT, Stockholm

Key words: Peron-Frobenius theorem, least principal component vector, mineral chemistry, ammonites, ostracods, foraminifers, crabs.

Abstract

The smallest latent root and associated vector of a $k * k$ positive definite square symmetric matrix is shown to have diagnostic value for finding an invariant linear combination if the smallest root is very small, almost zero, and much smaller than the k-1-th root. The matrix must be a covariance matrix and to derive from data that do not house significantly outlying observations. Examples are presented, drawn mostly from the field of palaeobiology of invertebrates. The approach used in identifying the role of a very small latent root and vector cannot be applied to correlation matrices nor to a covariance-equivalent matrix of compositional data for mathematical reasons.

Zusammenfassung

Der kleinste Eigenwert einer $k * k$ positiv definiten quadratischen Matrix – obgleich beinahe gleich Null – und dessen Eigenvektor, können von diagnostischem Wert sein, um eine invariante Relation zu identifizieren. Die Matrix muss unbedingt die Kovarianzmatrix sein. Die angeführten Fälle stellen sowohl „echte" Beispiele von invarianten Verhältnissen bei kleinstem Eigenvektor als auch fragliche Modelle dar.

1. Mineralchemie: Die geothermische Eigenanalyse SAXENAs (1969) wird als Beispiel eines fehlerhaften Vorbildes angeführt.
2. Die Analyse der Morphologie der Ammonitengattung *Schloenbachia* (Kreide) stellt das Vorkommen eines unveränderlichen Verhältnisses bezüglich der Gehäuseform dar, und zwar hinsichtlich Windungsbreite und Breite der Ventralseite.
3. Die Ammonitengattung *Discoscaphites* (Kreide) und das Problem der Biometrie des Gehäuses ist ein Beispiel für die Bedeutung der Invariantentheorie in der Biologie der Mollusken. LANGMAN & WAAGE (1993) versuchten, eine Hauptkomponentenanalyse ihres Materials durchzuführen, ohne verstanden zu haben, dass man es mit den Eigenschaften der logarithmischen Spirale zu tun hatte – nämlich, dass die logarithmische Spirale ihre eigene reziproke Polare in Bezug auf jede gleichseitige Hyperbel hat, welche ihren Asymptotenpunkt als Mittelpunkt hat und auf der sie beruht. Infolgedessen ist eine Hauptkomponentenanalyse der Lateralseite des Ammonitengehäuses wertlos, bestenfalls nur wenig zielstrebig.
4. Die Ostracodenart *Veenia fawwarensis* (Kreide). Diese Analyse stellt ein mögliches Beispiel der Unveränderlichkeit bei gewissen Variablen der Skulptur des Gehäuses dar.
5. Die Foraminiferengattung *Afrobolivina* (Kreide und Paläozän). Zwei gute Beispiele von Unveränderlichkeit bei der Gestaltung des oberen Schalenteiles (Foramen).
6. Die lebende Krabbe *Carcinas maenas*: Dieses Material (Bassin d'Arcachon, Frankreich) ist ein gutes Beispiel von Unveränderlichkeit bezüglich der Schalenbreite sowohl bei Männchen als auch bei Weibchen.

Introduction

Ever since the introduction of latent roots and vectors of a positive definite square symmetric matrix were introduced into multivariate statistical analysis interest has been centered on interpreting the first few latent vectors with the end in view of learning as much as possible about those linear combinations of the variables involved that are providing most of the variability in the material.

However, there is another line of enquiry that ought to be of interest, but which has remained largely unasked, notwithstanding that several mathematical statisticians have wondered over this lack of interest. GNANADESIKAN & WILK (1969), GOWER (1967), GNANADESIKAN (1977), MARDIA et al. (1979) have pointed out that the information resident in the smallest (zero, or almost zero) latent root should be of interest for finding a linear combination of variables which is invariant in the material under study. That is, that combination which is constant, or almost constant, for variables measured in the same metric. The justification for this may not be immediately obvious. GNANADESIKAN and WILK (1969), in a geometrically constructed example, showed the manner in which the smallest latent root and its associated vector, can be used for probing a structural relationship.

It is not always appreciated that the interpretation of principal components is based on an artifact, which is the case in multivariate statistical analysis. The Perron-Frobenius theorem states that among the latent roots of a positive matrix A there will be a real positive value, $\lambda = x$, the maximum root, the value of which is not surpassed by any other latent root of the matrix and which has a positive latent vector $x > 0$ (cf. ZURMÜHL, 1964, p. 219). Moreover, MARDIA et al. (1979, pp. 235, 241) noted that there is an indeterminacy involved in reifying principal components, for example, in the case where (p Ï k) latent vectors are equal or almost equal.

GOWER (1967) made several observations of importance in a critique of Principal Components as a statistically relevant tool. Some are obvious, for example all the variables in the data matrix must be measured in the same units. For many applications, the extraction of the latent roots and vectors is made on the correlation matrix in the belief that this will "stabilize" the data. Another ploy is to work on the logarithms of the observations. In the situations studied in the present note, neither of the foregoing procedures is permissible granted that we are looking for intrinsic structural information. Hence, the reification of the smallest principal component is only valid for data that have not been 'tampered with'. Geometrically, the components of the smallest principal component vector constitutes the best ($n - 1$) flat (i.e. the multidimensional analogue of a plane – GOWER, 1967) which fits the points the coordinates of which are the direction cosines of the normal to the flat.

DEMPSTER (1969, p. 139) commented on the flavor of vagueness and arbitrariness encapsulated in the method of principal components. He made the point that it is mathematically feasible for the last latent vector corresponding to the smallest latent root should be the only one of use in predicting some scientifically important feature.

Empirical studies (REYMENT, 1978, 1979) show that for the elements of the last latent vector to be useful, the sample size must normally be large, around 100 objects for 10 variables. This is not a hard and fast rule, however, as the variances and covariances should be stable (i.e., not be influenced by atypical observations). The examples briefly presented in the next section have been chosen in order to illustrate situations where the invariant role of the smallest latent vector is reasonably clear and other situations where this is not so transparent and where this would be incorrect.

Examples

1. A problem in mineral chemistry

SAXENA (1969) was concerned with studying solid solutions of silicates and geo-thermometry from the aspect of the distribution of iron and magnesium in co-existing garnet and biotite using principal component analysis. Ninety three samples of rocks formed at various pressures and temperatures were analyzed with respect to the following variables (at the outset we note that the distribution coefficient of SAXENA is a ratio constituted by two of the variables involved in the component analysis. This is never a sound construction):

x_1 – the distribution coefficient on a one-cation exchange basis; $K = x_3(1\text{-}x_2)/(1 - x_3)x_2$ and x_2 – Fe in garnet; x_3 – Fe in biotite; x_4 – Mn in garnet; x_5 – Ca in garnet; x_6 – Al^{IV} in biotite; x_7 – Al^{VI} in biotite; and x_8 – Ti in biotite.

SAXENA extracted the principal components of the correlation matrix of his data and interpreted the results as distinguishing between rocks of low and high metamorphic grade. The smallest latent root accounts for 0.32 % of the total variance with the associated latent vector

(– 0.58, 0.66, 0.50, 0.09, 0.00, – 0.01, – 0.04, – 0.15).

This vector was interpreted by SAXENA as representing an invariant relationship between the first three variables. On closer inspection, it becomes apparent that the relationship is in fact expressing x_1 in terms of x_2 and x_3. In other words, the smallest principal component is no more than an artifact, produced by the derived variable x_1.

2. *Schloenbachia*, a morphologically complicated genus of ammonites

The Albian ammonite genus *Schloenbachia* is remarkable for its great variability in apertural shape characteristics. A sample of 18 well preserved shells was subjected to a principal component extraction of the covariance matrix based on six variables observed on the apertural surface, to wit:

1 – maximum diameter of the conch,
2 – maximum breadth of the last whorl,
3 – minimum breadth of the last whorl,
4 – diameter at the beginning of the last whorl,
5 – breadth of the venter at the beginning of the last whorl,
6 – distance across the last whorl to the point of intersection with the second last whorl.

The smallest latent root corresponds to 0.57% of the total variability. The associated latent vector is

(– 0.12, 0.33, 0.32, 0.03, – 0.87, 0.07).

This vector indicates an invariant relationship to exist between variables 2, 3, and 5. That is between two whorl-breadth measures and the width of the venter.

3. *Discoscaphites*, a heteromorphic ammonite genus, and a constraint

Ammonites are coiled in close approximation to the logarithmic spiral. However, it has not been recognized by workers interested in studying the biometrics of ammonite shells that the logarithmic spiral imposes a constraint on the quantitatively appraisable variability of conchs in lateral aspect if the suite of selected variables are part of the spiral growth pattern (KLEIN, 1926, pp. 171–173). As an example of this fact, a simple analysis of four lateral measurements of the conchs in Maastrichtian macroconchs of *Discoscaphites conradi* (Morton), to wit, maximum diameter, two whorl dimensions and umbilical width, yielded the latent roots

Latent roots	(1)	(2)	(3)	(4)
	1.0586	0.0125	0.0005	0.0002
Percentages	98.76	1.17	0.05	0.02

There are three very small latent roots of which two are almost zero, whereas the largest latent root accounts for almost all of the variability. In approximate terms, the rank of the covariance matrix from observations in the plane of coiling is almost of unit rank. What does this signify? It is expressing the fact that the logarithmic growth spiral locks the variability in a firm grip and hence the small latent roots cannot be interpreted as natural expressions of invariance in a morphometric sense as was done for example, by LANDMAN & WAAGE (1993, p. 248).

4. The ostracod species *Veenia fawwarensis*

Veenia fawwarensis Honigstein is a Santonian (Cretaceous) species occurring in the sequence at Shiloah, Jerusalem (REYMENT & SAVAZZI, 1999, p. 248). The measurements made on the carapace are (1) the length of the carapace, (2) the height of the carapace, (3) the distance from the eye tubercle to the adductorial boss, (4) the distance from the adductorial boss to the posteroventral angle, (5) the distance from the posterodorsal angle to the posteroventral angle, (6) the distance from the eye tubercle to the posterodorsal angle, (7) the maximum width of the carapace. The material was collected from nine well defined limestone levels in the Shiloah quarry.

The latent roots of the covariance matrix for 79 specimens are

(1)	(2)	(3)	(4)	(5)	(6)	(7)
223.417	66.401	31.513	14.440	7.629	4.707	4.029

The last two latent roots are almost equal in magnitude, which implies that a clear-cut indication of invariance can hardly be expected. The smallest root is connected to the latent vector

(– 0.320, – 0.084, 0.149, 0.874, – 0.281, – 0.094, – 0.126).

There could possibly indicate that the distance between the adductorial boss and the posteroventral angle could be invariant over time.

5. The benthic genus of foraminifers *Afrobolivina*

5.1. *Afrobolivina afra* – a Campano-Maastrichtian foraminiferal species

Nine characters were measured on 234 individuals of the benthic foraminiferal species *Afrobolivina afra* Reyment from the Lower Maastrichtian of eastern Nigeria (REYMENT, 1991). These data yielded the latent *root*

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
499.069	47.296	34.545	8.300	6.493	6.089	4.877	3.712	0.351

The latent vector of the smallest latent root, which accounts for 0.057% of the total variance, is

(0.001, – 0.046, 0.258, 0.016, 0.001, – 0.119, 0.990, – 0.010, – 0.490)

There is only one element of high dignity, to wit, the width of the penultimate chamber. This indicates that the exposed part of the second last chamber does not vary over time and hence is invariant. The explanation of this seeming anomaly is not difficult to find. *Afrobolivina* is strongly polymorphic with respect to the number of chambers. This polymorphism does not influence the structure of the ultimate chamber to any marked extent. Here, we see that the expression of invariance is bound to the development of the penultimate chamber.

5.2. *Afrobolivina africana* (Graham, de Klasz, Rérat) – a Paleocene foraminifer

It is of interest to investigate the properties of the Paleocene derivative of *A. afra*. The study material is composed of 401 complete tests from a borehole at Akisinde, Nigeria. The six characters selected for principal component analysis are illustrated in (REYMENT, 1966, p. 328).

The percentages of each root extracted from the covariance matrix are:

(1)	(2)	(3)	(4)	(5)	(6)
47.83	26.70	12.13	7.18	4.82	1.34

The sixth latent root is appreciably smaller than the first five. The corresponding latent vector is

(– 0.040, 0.823, – 0.056, – 0.010, – 0.027, – 0.078).

The component 2, breadth of the test across the ultimate and penultimate chambers, dominates the picture and thus suggests an invariant relationship. This agrees with what was found for *Afrobolivina afra*, discussed in the foregoing example, but in the present case the invariance extends to the last two chambers and hence the foramen. The reason for this may lie with the polymorphic status of the Cretaceous species *A. afra*, whereas *A. africana* is not markedly polymorphic.

6. The recent crab species *Carcinus maenas* L.

We shall now consider the crab species *C. maenas* not least because of its historical significance in the original development of multivariate biometrics.

The idea of using the reduction of a covariance matrix seems to have been first invoked by TEISSIER (1938), who called his approach 'analyse factorielle'. What TEISSIER actually did was not factor analysis *sensu stricto* and, moreover, he only computed the first latent vector of the covariance matrix for males and females of the species *Maia squinada*. In other words, he computed the first latent vector for his species of crabs.

The material of *Carcinus maenas* used here derives from the Bassin d'Arcachon, France where I made a marine biological survey of the Bassin in 1966 with emphasis on the crustaceans and, in particular, the ravages made by the introduced muricid drill *Urosalpinx cunifera* (in the bilge-water of North American ships of war). The individuals for biometric study were all collected from the wading zone in the vicinity of the village of Le Canon in the northwestern sector of the bay. The variables selected for analysis were (1) maximum breadth of carapace, maximum height of carapace (2) distance between the posterior of the carapace to line of maximum breadth (3) width of posterior margin of the carapace (4). Additionally, measures on the left and right chelipeds were made, four measures in all. The eighth latent root for males constitutes 0.076% of the total variability with the associated latent vector

(– 0.877, 0.383, 0.122, 0.112, 0.219, – 0.065, 0.070, 0.007).

The eight latent roots for females constitutes 0.028% of the total variability and is associated with the latent vector

(0.662, – 0.745, – 0.011, 0.081, – 0.015, 0.000, – 0.011, 0.011).

For both males (n=55) and females (n=21), the first variable indicates invariance, to wit, mainly in carapace breadth. There is for females also an important loading for the second variable, maximum height of the carapace. This difference has attributed to the influence of sexual dimorphism (REYMENT, 1969). Be

it noted that all crabs selected for statistical appraisal were, when collected, locked in copulation and hence adult individuals.

Concluding comments

The examples presented here were selected for their value in illustrating situations where the smallest principal component is of significance for expressing invariance in a set of variables. Additionally, examples are presented where the smallest component does not permit a reasonable basis for concluding an invariant relationship to exist. A case of an imposed constraint, whereby the covariance matrix is virtually of unit rank due to the effect of the logarithmic spiral that controls the shape of the shells of coiled organisms is discussed. This constraint makes the multivariate analysis of, for example ammonite conchs, questionable when observed in lateral aspect and all variables are under the influence of the logarithmic growth factor.

References

DEMPSTER, A. P. (1969): Continuous Multivariate Analysis.- Addison-Wesley Publishing Company: 388 p., Boston MA (US)

GNANADESIKAN, R. (1977): Methods for Statistical Data analysis of Multivariate Observations. - Wiley and Sons: 311 p., New York.

GNANADESIKAN, R., WILK, M. B. (1969): Data analytic methods in multivariate statistical analysis. - In: KRISHNAIAH, P. R. [ed.]: Multivariate Analysis. - Academic Press: 593–638, New York.

GOWER, J. C. (1967): Multivariate analysis and multidimensional geometry. - The Statistician, **17**: 13–28, Oxford etc.

JÖRESKOG, K. G., KLOVAN, J. E., REYMENT, R. A. (1976): Geological Factor Analysis. - Elsevier: 178 p., Amsterdam (NL).

KLEIN, F. (1926; reprint 1968): Vorlesungen über höhere Geometrie. - Otto Springer Verlag: 405 p., Berlin.

LANDMAN, N. H., WAAGE, K. M. (1993): Scaphitid ammonites of the Upper Cretaceous (Maastrichtian) Fox Hills formation in South Dakota and Wyoming. - Bull. American Museum Natural History, no. 215: 257 p., New York

MARDIA, K. V., KENT, J. T., BIBBY, J. M. (1979): Multivariate Analysis. - Academic Press: 521 p., New York.

REYMENT, R. A. (1966): *Afrobolivina africana* (Graham, de Klasz, Rérat): Quantitative Untersuchung der Variabilität einer paleozänen Foraminifere. - Eclogae Geologicae Helvetiae, Birkhäuser, **59**: 319–337, Basel etc.

REYMENT, R. A. (1969): Some case studies of the statistical analysis of sexual dimorphism. - Bull. Geological Institutions University of Uppsala, N. S., **1**: 97–119, Uppsala.

Реймент, Р. А. (1978): Интерпретация наименьшей главной компоненты [REYMENT, R. A.: Interpretation of the smallest principal component]. - Юбилейный сборник „А. Б. Вистелиус“ Академии Наук СССР [Spec. Publ. Acad. Sci. USSR “A. B. Vistelius”], **254**: 163–167, Ленинград [Translated into English language in: Bull. Geol. Instn. Univ. Uppsala, N. S., **1979** (8): 1–4, Uppsala].

REYMENT, R. A. (1991): Multidimensional Palaeobiology. - Pergamon Press: 377 p., Oxford.

REYMENT, R. A., SAVAZZI, E. (1999): Aspects of Multivariate Statistical Analysis in Geology. - Elsevier: 285 p., Amsterdam.

SAXENA, S. K. (1969): Silicate solid solutions and geothermometry. - Contributions to Mineralogy and Petrology, Springer, **22**: 259–267, Berlin/ Heidelberg

TEISSIER, G. (1938): Un essai d'analyse factorielle. Les variants sexuels de *Maia squinada*. - Biotypologie, **7**: 73–96, Paris.

ZURMÜHL, R. (1964): Matrizen und ihre technischen Anwendungen. - Springer Verlag, 4th ed.: 452 p., Berlin/ Göttingen/Heidelberg.

Manuscript received: August 17, 2011
Manuscript accepted: January 30, 2012

Address of the author:
Prof. Dr. Richard A. Reyment, Department of Palaeozoology, Naturhistoriska Riksmuseet, Stockholm (Sweden)
richard.reyment@nrm.se