On stability of compositional canonical variate vector components

R. A. REYMENT

Palaeozoology Section, Swedish Museum of Natural History, Box 50007, S10405, Stockholm, Sweden (e-mail: richard.reyment@nrm.se)

Abstract: Canonical variate analysis (aka discriminant coordinates) is viewed from the aspect of Aitchisonian compositional data analysis and the concept of stability in canonical vectors examined in relation to their reification (i.e. providing canonical vector components with a practical interpretation). The log-ratio transformation was found to have computational and interpretational advantages over the centred log-ratio transformation. The *ad hoc* application of N. A. Campbell's application of the method of shrinkage estimators to multiple discrimination, and the optimal retention of discrimination power, is exemplified by two cases, one drawn from quantitative sedimentology, the other from biomolecular palaeontology, with the intention of probing the effect of instability on interpreting the relative importance of standardized canonical variate coefficients in relation to the suppression of near-redundant directions of within-groups variation. Pronounced instability in canonical vectors may endanger the validity of an analysis.

Campbell & Reyment (1978) considered the problem of instability in canonical vectors in palaeontology using data on the Nigerian Cretaceous foraminiferal species *Afrobolivina afra* Reyment for exemplifying procedures. It was reported that where instability occurs, stability of the coefficients within the framework of repeated sampling can be achieved by suppression of near-redundant directions of the within-groups variation. The identification of redundant information in multivariate analysis has been slow to enter into the praxis of statistical methods (Campbell 1979, 1980*a*; Seber 1984). So far, interest in investigating reliability in canonical vectorial components has been centred on full-space data, with a preference for biological material. The question of stability for multivariate models in simplex space, such as occur in geostatistics, has been given scant attention, despite its potential significance for good analytical practice. In this note, the problem is illustrated for geochemical data in stratigraphical sedimentology, and for amino acids in biomolecular palaeontology. Both cases are typically compositional in nature.

The presentation followed here is made at two levels: one for the interested geologist with little experience of multivariate statistics; the second, in Appendix A, for anybody desiring to undertake his/her own investigations and with a basic understanding of multidimensional analysis. It should be understood that this article does not profess to have the status of mathematical novelty because it is no more than an exemplification of existing techniques, rather well known in biometrics, to geological materials.

An important issue arising in canonical variate analysis concerning stability in canonical vector coefficients was studied by Campbell (1979, 1980*a*,*b*) who demonstrated that shrunken estimators in discriminant and canonical variate analysis (discriminant coordinates; multiple discrimination) may lead to improved stability of the resulting coefficients when the between-groups sum of squares for a particular principal component (more stringently, latent roots and vectors), defined by the within-groups covariance or correlation matrix, is small and the corresponding latent root is small. Campbell's engagement in the problem arose out of collaboration with marine biologists concerned with practical aspects of the commercial utilization of marine crustaceans and gastropods in Australian and New Zealand waters. The ideas underlying shrinkage constants in multivariate analysis derive from practical problems of stability that occur in multiple regression analysis (Goldstein & Smith 1974; Campbell & Furby 1994; Gui 1999). Granted that the same mathematical structure exists in discriminant functions, the application of ridge-type estimators to that field is a logical step and one that can be tentatively extended to multiple discriminant analysis.

The provenance of the data

In this paper two examples are considered, the one a canonical variate analysis of the geochemistry of Lithuanian Silurian sediments using data kindly made available by Dr Donata Kaminskas, Geology Department, University of Vilnius. The

From: BUCCIANTI, A., MATEU-FIGUERAS, G. & PAWLOWSKY-GLAHN, V. (eds) *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Geological Society, London, Special Publications, **264**, 59–66.
0305-8719/06/$15.00

second example is concerned with a biomolecular study of fossil and living brachiopods (Endo *et al.* 1995). A presentation of the observations used here for the Lithuanian material is given by Kaminskas & Malmgren (2003) in a statistical analysis of three sequences of Silurian sediments dominated by carbonates and mudrocks; in that paper, the compositional nature of the data was taken into account adequately. The brachiopod data concern intra-crystalline molecules isolated from the shells of four species collected from horizons extending over the last 1.47 million years. The material was analysed by immunoassay and amino-acid analysis. The species studied were Japanese occurrences of *Coptothyris grayi* (Davidson), *Terebratalia coreanica* (Adams & Reeve), *Pictathyrispicta* (Dillwyn) and *Laqueus rubellus* (Sowerby). A detailed account of the methods of analysis is provided in Endo *et al.* (1995). The original statistical analysis was made on 14 amino acids determined on 52 specimens. For present purposes the dataset was reduced to eight parts, none of which has zero entries.

Overview of statistical procedures

Aitchison (1986) showed that a canonical variate analysis of compositions can be made using either the log-ratio covariance matrix or the centred log-ratio covariance matrix, in both cases for the within-groups and between-groups sums of squares and cross-products. The advantage of the former formulation is that the within-groups matrix **W** and the between-groups matrix **B** are positive definite and hence possess a normal inverse. The awkward aspect here is that a common divisor is required. The latter has the attractive property that all parts are represented in the formulation, but the centred log-ratio covariance matrix is singular and therefore requires a generalized inverse. Further reservations have been noted by Egozcue *et al.* (2003). For detailed discussions of Aitchison's statistical results in compositional data analysis, the reader is referred to specialist papers appearing in the theoretical section of this issue.

It is recommended that the vector-stability analysis proceeds in the following steps.

1. Make a preliminary (graphical) inspection of the data for redundancy and other singularities.
2. Transform the frequencies for m parts to the desired log-ratio form (Aitchison 1986, 1997). This may be as a simple log-ratio transformation or, *le cas écheant*, a centred log-ratio transformation.
3. Perform a canonical variate analysis of k groups. If the results indicate there could be a likelihood of marked instability in the canonical vectors, apply the method of ridge regression (shrunken estimators) in orientational mode, as described in Campbell (1979, 1980*a*) and Campbell & Reyment (1978), and examine the coefficients of the canonical vectors for significant change.

Canonical variate analysis

The method of canonical variate analysis is usually considered to be a suitable multivariate statistical procedure for treating the problem posed by the simultaneous analysis of several sampling levels in, say, observations made over a stratigraphical sequence. In many applications of canonical variate analysis, the relative magnitudes of the coefficients for the variables (or parts in the case of compositions) standardized to unit variance by the pooled within-groups standard deviations often prove useful indicators of those coefficients that are likely to be influential for discrimination. The success of such an operation presupposes that the coefficients are stable over repeated sampling (Campbell 1980*b*; Campbell & Atchley 1981; Seber 1984).

In accordance with accepted methodology, and adhering to Aitchison's (1986) formulation, the computation of canonical variates may be regarded as a two-stage rotational procedure. The first step rotates to orthogonal variables, which may be referred to as the principal components of the pooled samples. The second rotation corresponds to a principal component analysis of the group means in the space of the orthogonal variables. The first step transforms the within-groups dispersion ellipsoid into a concentration spheroid by scaling each latent vector by the square root of the corresponding latent root.

Consider the variation between groups along each orthogonalized variable (i.e. each principal component). Where there is slight variation between groups along a particular direction, and the corresponding latent root is small, marked instability can be expected in some of the coefficients of the canonical variates, granted that the instability is under the sway of small changes in the properties of the data. An approach that often serves to overcome the problem of instability in canonical coefficients is to add shrinkage, or ridge-type constants (Goldstein & Smith 1974) to the latent roots before they are used to standardize the corresponding principal component. When an 'infinitely large' constant is added, this confines the solution to the subspace orthogonal to the

vector, or vectors affected by the addition. Experience dictates that when the between-groups sum of squares for a particular principal component is small (say, less than 5% of the total between-groups variation) and the corresponding latent root is also small (less than 2–3%) then shrinking of the principal component will often prove useful (Campbell 1980*a*). It is often observed that although some of the coefficients of the canonical vectors corresponding to the canonical variates of interest change magnitude and, moreover, often sign, shrinkage has little effect on the corresponding canonical roots, thus betokening that little discriminatory information has been lost. This indicates that those variables contributing most to the shrunken principal component have little influence for discrimination and may be considered for removal. Furthermore, variables (parts) with small standardized canonical variate coefficients can be vetted for exclusion. A simple account of applications of canonical variate analysis in biological (including palaeontological) studies of Reyment *et al.* (1984, chapter 7).

A brief account of the method of shrinkage, as presented by Campbell (1979, 1980*a*) is given in Appendix A.

The Lithuanian Silurian sediments

The samples derive from three boreholes drilled in connexion with a chemo-stratigraphical study of the Silurian sedimentary sequences of Lithuania, located geographically as outlined below. A detailed account of the geological aspects of the problem are given in Kaminskas & Malmgren (2003).

1. The deepest part of the sedimentary basin was penetrated by the Kurtuvénai-161 borehole in northwestern Lithuania.
2. The intermediate beds were drilled by the Ledai-179 borehole in central Lithuania.
3. The uppermost beds of the sedimentary basin were encountered by the Jocionys-299 borehole in southeastern Lithuania.

The oxides determined were (Paskevicius 1997): SiO_2, Al_2O_3, Fe_2O_3, MnO_2, MgO, CaO, Na_2O, K_2O, TiO_2, P_2O_5. A preliminary appraisal of the data array indicated that two of the oxides were not contributing anything essential to the analysis and they were therefore deleted, after which the constant row-sum constraint was re-established. The reduced set of eight parts then lacked the columns for manganese and phosphate. Recall that the constituent categories of the geochemical array are referred to as 'parts' since they are not variables in the accepted statistical sense. This usage is to underline the fact that deletion of one or more of the proportions necessitates reinstating the constant sum condition which automatically alters the covariance relationships between the parts of each of the constituent rows. Should this step be neglected, then the relationships between parts are rendered spurious. In the case of what may be referred to as true variables, this restriction does not apply. In the ensuing analysis, the common divisor for the log-ratios was taken to be SiO_2. The geochemical analysis of these data is given in Kaminskas & Malmgren (2003), who also report multivariate statistics for comparisons.

The within-groups and between-groups matrices of sums of squares and cross products of the log-ratios in standardized form (correlation mode) are listed in Table 1. There are several very high correlations which, according to the results of Campbell (1979, 1980*a*), can set the stage for instability in canonical vector coefficients. The latent roots and vectors for $\mathbf{W}^*$ are given in Table 1. There are two large latent roots. The two smallest latent roots do not differ greatly from zero. Comparison of these smallest roots with the appropriate values of diag $\mathbf{G}$ shows that the smallest latent root is connected to a larger value than is the sixth latent root. The corresponding latent vector (principal component) represents mainly a bipolar relationship between parts 1 and 2. The third latent vector, which connects to the smallest value of diag $\mathbf{G}$, weighs parts 3 and 5 against part 4. Both of these directions could well be candidates for shrinkage. In order to test this, the two smallest latent vectors were suppressed, in turn (Table 2). The coefficients for $\mathbf{a}_1^U$ highlight the contributions from principal components 2 and 4 to the first canonical variate. The main contributing principal component of $\mathbf{a}_2^U$ to the second canonical variate is the seventh (Table 2). The effect of shrinking the smallest principal component has a notable effect on the sum of the canonical roots and the canonical vectors are perturbed rather strongly. The effect of shrinking the third principal component has virtually no influence on the sum of the canonical roots, which implies that discrimination power is undiminished as a result of the shrinkage exercise. This is even more marked for shrinkage of the sixth principal component. The effect of shrinking on the matching canonical vectors is slight, with the exception of part 1 in the first canonical vector. The conclusion that presents itself here is that shrinkage of the sixth principal component can be expected to improve the statistical quality of the analysis with respect to reification.

Brachiopod biomolecules study

Several species of brachiopods (cf. Introduction) were made the object of this study. Using modern

Table 1. *Log-ratio matrices for the Lithuanian data and spectrum of* $\mathbf{W}^*$ *(n = 221) in correlation mode*

The input within-groups matrix, $\mathbf{W}^*$ (upper triangle) and between-groups matrix $\mathbf{B}^*$ (lower triangle)							
	1	2	3	4	5	6	7
1	0.0190	0.9917	0.1333	−0.0583	0.7437	0.9422	0.9870
2	0.0251	0.0331	0.1365	−0.0446	0.7412	0.9332	0.9905
3	−0.1453	−0.1936	1.2941	0.6040	0.0533	−0.0362	0.1904
4	−0.0939	−0.1233	0.6723	−0.4742	−0.1712	−0.2762	−0.0060
5	0.0778	0.1033	−0.6628	−0.3675	0.3438	0.7276	0.7543
6	0.0183	0.0224	−0.1593	−0.0859	0.0822	0.0197	0.9065
7	0.0180	0.0238	−0.1481	−0.0860	0.0775	0.0185	0.0176
Latent roots of $\mathbf{W}^*$							
	4.5300	1.6563	0.4047	0.3377	0.0573	0.0081	0.0060
Latent vectors of $\mathbf{W}^*$							
Parts	P1	P2	P3	P4	P5	P6	P7
1	−0.4634	0.0437	0.1813	−0.1090	−0.1149	***0.8359***	0.1631
2	−0.4626	0.0519	0.1956	−0.0933	−0.2430	***−0.4779***	0.6700
3	−0.0513	0.6919	***−0.5580***	−0.4522	0.0460	−0.0014	0.0266
4	0.0595	0.6984	***0.4818***	0.4863	0.1984	−0.0129	−0.0231
5	−0.3893	−0.0567	***−0.5947***	0.6984	0.0528	0.0083	0.0315
6	−0.4487	−0.1308	0.1139	−0.2161	0.8186	−0.1487	−0.1721
7	−0.4607	0.0924	0.1372	−0.0632	−0.4619	−0.2248	−0.7018
Diag $\mathbf{G}$ = between groups sums of squares for principal components							
	0.0506	0.9417	***0.1043***	1.0064	0.4938	***0.0287***	0.3715
Trace $\mathbf{G}$ = 2.9970							
Percentage of trace							
	1.687	31.420	***3.480***	33.580	16.477	***0.959***	12.397

Bold italics denote values of interpretational consequence. Key for SiO_2 log-ratios: Al(1), Fe(2), Mn(3), Ca(4), Na(5), K(6), Ti(7).

techniques of molecular biology, it was possible to obtain detailed information on amino acids contained in preserved shell material (Endo *et al.* 1995). The results reported briefly below form part of the imaginative work in molecular palaeontology being carried out by Professor Kazuyoshi Endo and his associates at Tsukuba University, Japan.

The log-ratio covariance matrix $\mathbf{W}$ for the amino acids is listed in Table 3. Inspection of the entries shows that $w_{6,6}$ and $w_{7,7}$ are by far the largest. The palaeobiological and ecostratigraphical

Table 2. *Standardized log-ratio canonical vectors for the Lithuanian data, including shrunken estimates: vectors adjusted to standard deviations*

	Principal component							
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	Canonical roots
$\mathbf{a}_1^U$	0.145	***0.626***	−0.065	***−0.597***	0.413	0.063	−0.223	2.397
$\mathbf{a}_2^U$	−0.026	−0.052	−0.396	−0.505	−0.397	0.178	***0.641***	0.600
	Adjusted canonical variate vectors log-ratios for parts							
	1	2	3	4	5	6	7	Canonical roots
$\mathbf{c}_1^U$	−0.009	−2.600	0.855	0.196	−0.706	1.931	1.179	2.397
$\mathbf{c}_2^U$	3.177	4.946	0.858	−1.281	−0.035	−2.822	−5.561	0.600
$\mathbf{c}_{1\,(0,\ldots,\infty)}^{GI}$	0.565	−0.727	0.981	0.080	−0.639	1.384	−0.894	2.288
$\mathbf{c}_{2\,(0,\ldots,\infty)}^{GI}$	2.301	−0.624	0.622	−1.477	−0.246	−2.278	0.549	0.337
$\mathbf{c}_{1\,(0,\ldots,\infty,\infty)}^{GI}$	−0.602	−2.352	0.854	0.210	−0.713	2.049	1.359	2.388
$\mathbf{c}_{2\,(0,\ldots,\infty,\infty)}^{GI}$	1.534	5.954	0.888	−1.273	−0.063	−2.598	−5.176	0.581

Table 3. *The within-groups matrix* **W** *for log-ratio data, its latent roots and vectors and diag* **G** *for the brachiopod amino acids (n = 52)*

	1	2	3	4	5	6	7
1	3.8060	2.2520	0.2660	0.9430	3.5450	4.9210	0.0120
2	2.2520	1.9600	1.0190	0.7190	1.8670	3.7130	−0.1410
3	0.2660	1.0190	3.7660	1.0980	−1.4410	−1.0780	1.1460
4	0.9430	0.7190	1.0980	4.4320	0.5170	0.2730	5.8970
5	3.5450	1.8670	−1.4410	0.5170	9.8330	12.3510	−1.7170
6	4.9210	3.7130	−1.0780	0.2730	12.3510	***63.8340***	−11.5440
7	0.0120	−0.1410	1.1460	5.8970	−1.7170	−11.5440	***41.0870***
Latent roots of **W**							
	71.7529	37.7762	8.8180	5.4903	2.6884	1.9037	0.2885
Latent vectors of **W** parts							
	1	2	3	4	5	6	7
1	−0.0773	−0.0679	−0.4435	−0.3022	−0.2873	***0.5924***	−0.5177
2	−0.0561	−0.0451	−0.2274	−0.3549	−0.2843	0.2226	0.8287
3	0.0239	−0.0231	0.1003	−0.6952	−0.3017	***−0.6087***	−0.2098
4	0.0249	−0.1701	−0.1044	−0.4734	0.8513	0.1000	0.0262
5	−0.1973	−0.1057	−0.8140	0.2497	0.0780	−0.4675	0.0167
6	−0.9068	−0.3358	0.2533	0.0071	−0.0074	0.0197	−0.0150
7	0.3583	−0.9165	0.0620	0.1138	−0.1214	−0.0113	0.0016
Diag **G** = between groups sums of squares for principal components							
	1	2	3	4	5	6	7
	0.0941	0.2361	0.1007	0.2450	0.3718	***0.0417***	0.3345
Trace **G** = 1.42405							
Percentage of trace							
	6.607	16.580	7.071	17.206	26.111	2.930	23.495

Here and elsewhere in the text, the amino acids designated 1–8 (8 is the divisor) correspond to the designations D, E, G, T, A, P, Y, I for amino acids in Endo *et al.* (1995).

reasons for this are that the amino acids involved are much affected by degeneration over time. The latent roots and vectors of **W** are presented in Table 3 together with the diagonal for the between-groups sums of squares matrix for the principal components, diag **G**. The first latent vector is dominated entirely by amino acid P and the second latent vector by amino acid Y. The smallest value of diag **G**, d_6, is just 2.93% of the trace of **G**, whereas the seventh entry is actually the greatest of all, being 23.5% of the trace of **G**. The principal component corresponding to d_6 is mainly an expression of bipolar covariation in the log-ratio vector component amino acid D (0.6) and vector component amino acid G (−0.6). These several relationships of small between-groups sums of squares

Table 4. *Standardized canonical vectors for the seven log-ratios, and shrunken estimates, for the brachiopod data: vectors adjusted to standard deviations*

	Principal component							
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	Canonical roots
$\mathbf{a}_1^U$	−0.283	−0.524	−0.210	−0.062	0.391	−0.074	0.662	0.723
$\mathbf{a}_2^U$	0.263	0.268	0.352	0.465	0.693	−0.190	0.049	0.436
Adjusted canonical variate vectors log-ratios of parts								
	1	2	3	4	5	6	7	Canonical roots
$\mathbf{c}_1^U$	−0.691	0.974	−0.286	0.264	0.131	0.020	0.032	0.723
$\mathbf{c}_2^U$	−0.368	−0.176	−0.189	0.236	0.041	−0.019	−0.048	0.436
$\mathbf{c}_{1\,(0,\ldots,\infty)}^{GI}$	−0.663	0.991	−0.312	0.261	0.110	0.021	0.034	0.720
$\mathbf{c}_{2\,(0,\ldots,\infty)}^{GI}$	−0.339	−0.103	0.288	0.231	−0.011	−0.012	−0.043	0.424

corresponding to principal component 6 and the small latent root for **W** tend to be amenable to shrinking of the principal component. The coefficients for $\mathbf{a}^{U1}$ highlight the contribution from principal components 2 and 7 to the first canonical variate (Table 4). The main contributing principal component of $\mathbf{a}^{U2}$ to the second canonical variate is the fifth. The effect of shrinking the smallest principal component (which is connected to the largest value in diag **G**) is to perturb the canonical variates and to occasion a serious loss of information in the canonical roots. Shrinkage of the sixth principal component (Table 4) has little effect on the components of all of the canonical vectors and the sums of the canonical roots are likewise little influenced. This indicates that the principal component 6 represents a redundant direction and can be eliminated with very little loss of discriminatory power. The practical effect of such a step is generally to achieve greater stability in the canonical vectors and an improvement with respect to their interpretation.

Comments

The two examples drawn from published investigations presented briefly in this paper serve to exemplify the application of the shrinkage technique for stabilization in the canonical variate analysis of compositional data. One example is for the correlation mode (the Lithuanian sediments), the other uses the theoretically, perhaps more faithful, representation in covariance mode (the brachiopods). In both examples it could be demonstrated that redundant directions of variation have a negative influence on the reification of the canonical vectors and hence the reliability of an analysis. It could also be shown that where there is a marked stratigraphical component in a dataset, such as occurs in long borehole sequences, this can influence the outcome of an analysis because of trend-weighting. The log-ratio transformation was found to be more reliable for the purposes of the shrinkage stabilizing technique. The centred log-ratio covariances are less easy to work with because of the practical difficulty of overcoming and deciphering the effect of the singularity of the input matrices, as well as other factors (cf. Egozcue *et al.* 2003).

Campbell (1979) was careful to point out that shrinkage is not the only way in which stability in discrimination can be approached under the circumstances applying in this paper and he provided alternatives. From the aspect of effective data analysis, selection of parts based on relative magnitudes and stability of coefficients may be preferable to a stepwise procedure. Finally, the successful application of the shrinkage technique often necessitates a considerable amount of work. It may therefore be of value to point out that the author's experience has been that instability in canonical vectors due to redundant directions of variation is not very common in all aspects of applied canonical variate analysis and this seems to be true of geology. This notwithstanding, advice to the analyst is that a trial test of the data is a safety investment that should not be neglected if one of the main goals of a study is the practical problem of the interpretation of the coefficients of the canonical vectors.

Dr D. Kaminskas is thanked for making the Lithuanian data available for analysis. Professor B. A. Malmgren helpfully directed attention to these data. The Trustees of the Swedish Museum of Natural History generously provided working facilities. The author is grateful to Professor K. Endo for continued advice concerning problems in the sphere of molecular palaeontology. A special expression of gratitude is extended to Professor Vera Pawlosky-Glahn for constructive comments, advice and information concerning ongoing research.

Appendix A: Review of the method of shrinkage

It is assumed here that multivariate procedures available for full space carry over, at least approximately, to simplex space (Aitchison 1986, p. 202). The within-groups sums of squares and cross-products matrix **W** on n_w degrees of freedom and the between-groups matrix of sums and squares and cross-products **B** are computed in the usual manner. Here, however, the data matrix is composed preferably of the log-ratio transformed observations (Aitchison 1983, 1986). Campbell (1979, 1980*a*) recommended that the within-groups matrix be expressed in correlation form with similar scaling for **B**. This is not mandatory, however. Aitchison (1997) proved that the correlation coefficient is not defined in simplex space, at least from the aspect of statistical interpretation of the correlation coefficient. In the present connexion, the use of correlations is strictly geometric for achieving spherical distributions with the end in view of bettering analytical stability (Campbell 1980*a*,*b*). The following review is in terms of the correlation mode. Hence:

$$\mathbf{W}^* = \mathbf{S}^{-1}\mathbf{W}\mathbf{S}^{-1} \tag{A1}$$

and

$$\mathbf{B}^* = \mathbf{S}^{-1}\mathbf{B}\mathbf{S}^{-1} \tag{A2}$$

where **S** denotes diag $\mathbf{W}^{1/2}$ and the asterisks indicate the correlation mode to apply. Here, and elsewhere in the text, an apostrophe denotes a transposed matrix. The latent roots e_i and latent vectors $\mathbf{u}_i$ of $\mathbf{W}^*$ are then found. The corresponding orthogonalized variables are the principal components with $\mathbf{E} = \text{diag}(e_1, \ldots, e_v)$ and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_v)$.

Consequently,

$$\mathbf{W}^* = \mathbf{UEU}'. \tag{A3}$$

Usually, the latent vectors are scaled by the square root of their latent roots; this a transformation for achieving within-groups sphericity, which is a manipulation for promoting stability. Shrunken estimators are constructed by adding shrinkage constants k_i to the latent roots diag $\mathbf{E}$ before scaling the latent vectors (Goldstein & Smith 1974; Campbell 1980*a*). Let the shrunken estimators be denoted as

$$\mathbf{K} = \text{diag}(k_1, \ldots, k_v). \tag{A4}$$

In a full-scale study, one would wish to test the effect of using different values of k. In the examples, a very large value has been used (in effect this is functionally infinitely great). Smaller values can be tried to test the consequence of reducing the between-group contribution from a particular component. Define now $\mathbf{U}^*$, the matrix of latent roots inflated by the chosen shrunken estimators.

$$\mathbf{U}^* = \mathbf{U}(\mathbf{E} + \mathbf{K})^{-1/2} = \mathbf{U}^*_{(k_1,\ldots,k_v)} \tag{A5}$$

Next form, in the usual manner, the between-groups matrix of sums and squares and cross-products in the space of the within-groups principal components, that is, form

$$\mathbf{G}_{(k_1,\ldots,k_v)}\mathbf{U}^{*\prime}_{(k_1,\ldots,k_v)}\mathbf{B}^*\mathbf{U}^*_{(k_1,\ldots,k_v)} \tag{A6}$$

Set d_i equal to the ith diagonal element of $\mathbf{G}$. This diagonal element, which is the between-groups sums of squares for the ith principal component, is an important diagnostic tool. The usual canonical vectors $\mathbf{c}^{\mathrm{U}}$ of canonical variate analysis are yielded by

$$\mathbf{c}^{\mathrm{U}} = \mathbf{U}^*_{(0,\ldots,0)}\mathbf{a}^{\mathrm{U}}. \tag{A7}$$

Generalized shrunken (i.e. generalized ridge) estimators are determined directly from the latent vectors $\mathbf{a}^{\mathrm{S}}$ of $\mathbf{G}_{(k_1,\ldots,-k_v)}$ with

$$\mathbf{c}^{\mathrm{S}} = \mathbf{U}^*_{(k_1,\ldots,k_v)}\mathbf{a}^{\mathrm{S}} \tag{A8}$$

The coefficient $a^{\mathrm{U}i}$ involves $d_i/e_i^{1/2}$. This implies that where the latent root is small, the value of $a^{\mathrm{U}i}$ is given by the ratio of two small quantities and hence can be expected to fluctuate widely from sample to sample. A generalized solution results when $k_i = 0$ for $i \leqslant r$ and $k_i = \infty$ for $i > r$ (i.e. the first r columns of $\mathbf{U}^*$). This gives $\mathbf{a}_{i\mathrm{GI}} = \mathbf{a}_{i\mathrm{U}}$ for $i \leqslant r$ and $\mathbf{a}^{\mathrm{GI}} = 0$ for $i > r$.

The generalized inverse solution results from forming

$$\mathbf{G}_{(0,\ldots,0;\,\infty,\ldots,\infty)} = \mathbf{U}^{*\prime}_r\mathbf{B}^*\mathbf{U}^*_r \tag{A9}$$

where $\mathbf{U}^*_r$ corresponds to the first r columns of $\mathbf{U}^*_{(0,\ldots,0)}$. The generalized canonical vectors are given by $\mathbf{c}^{\mathrm{GI}} = \mathbf{U}^*_r\mathbf{a}^{\mathrm{GI}}$, where $\mathbf{a}^{\mathrm{GI}}$, of length r, corresponds to the first r elements of $\mathbf{a}^{\mathrm{u}}$. (Note that $\mathbf{c}^{\mathrm{GI}} = \mathbf{c}^{\mathrm{S}}_{(0,\ldots,0;\,\infty,\ldots,\infty)}$).

Practical considerations

In practice it is frequently found that marked instability in vector coefficients is associated with a small value of a latent root e_v and a correspondingly small diagonal element d_v of $\mathbf{G}$. A useful rule of thumb is to examine the contribution of d_v to the total group separation, to wit, trace $(\mathbf{W}^{-1}\mathbf{B})$. In cases where the ratio of d_i to trace $\mathbf{G}$ is small, and, or, the corresponding ratio of canonical roots is small (<0.05) then little loss of the power of discrimination will result from excluding one or more of the smallest latent vectors or equivalently from suppressing the corresponding principal component. Total suppression is not always a necessity and some smaller value may be chosen such that the effect of the principal component is merely reduced (Campbell & Reyment 1978; Campbell 1979). Campbell (1979) reported that a generalized inverse solution with $r = v - 1$ frequently yields stable estimates. Reyment & Savazzi (1999) provide a compiled computer program for carrying out the computations encompassed by the foregoing algebra. Weihs (1995) has taken up the subject of the graphical representation of canonical variate results. As a complement to his treatise, Campbell constructed a very comprehensive interlocking computer program for the Department of Mathematics and Statistics, *CSIRO*, Wembley, Western Australia for canonical variate analysis which includes shrinkage, robust estimation when the covariances are not homogeneous, M-estimators, graphical procedures and much more.

Recent results by Egozcue *et al.* (2003) on isometric log-ratio transformations are clearly worth extension within the context of stability of latent vector coefficients in compositional data analysis.

References

AITCHISON, J. 1983. Principal component analysis of compositional data. *Biometrika*, **70**, 57–65.

AITCHISON, J. 1986. *The statistical analysis of compositional data*. Chapman & Hall, London.

AITCHISON, J. 1997. The one-hour course in compositional data analysis or compositional data analysis is easy. *In*: PAWLOWSKY-GLAHN, V. (ed.) *Proceedings of the Mathematical Geology third Annual Conference, Barcelona* (September, 1997), 3–35. Vera.

CAMPBELL, N. A. 1979. *Canonical variate analysis: some practical aspects*. PhD thesis, University of London.

CAMPBELL, N. A. 1980*a*. Shrunken estimators in discriminant and canonical variate analysis. *Applied Statistics*, **29**, 5–14.

CAMPBELL, N. A. 1980*b*. Robust procedures in multivariate analysis.1: robust covariance estimation. *Applied Statistics*, **29**, 231–237.

CAMPBELL, N. A. & ATCHLEY, W. R. 1981. The geometry of canonical variate analysis. *Systematic Zoology*, **30**, 268–280.

CAMPBELL, N. A. & FURBY, S. L. 1994. Variable selection along canonical vectors. *Australian Journal of Statistics*, **36**, 177–183.

CAMPBELL, N. A. & REYMENT, R. A. 1978. Discriminant analysis of a Cretaceous foraminifer using shrunken estimators. *Cretaceous Research*, **1**, 207–211.

EGOZCUE, J. J., PAWLOWSKY-GLAHN, V., MATEU-FIGUERAS, G. & BARCELÓ-VIDAL, C. 2003. Isometric log-ratio transformations for compositional data-analysis. *Mathematical Geology*, **35**, 279–300.

ENDO, K., WALTON, D., REYMENT, R. A. & CURRY, G. B. 1995. Fossil intra-crystalline biomolecules of brachiopod shells: diagenesis and preserved geo-biological information. *Organic Geochemistry*, **23**, 661–673.

GOLDSTEIN, M. & SMITH, A. F. M. 1974. Ridge type estimators for regression analysis. *Journal of the Royal Statistical Society*, B**36**, 284–291.

GUI, Q. 1999. Generalized shrunken-type robust estimation. *Journal of Surveying Engineering*, **125**, 177–184.

KAMINSKAS, D. & MALMGREN, B. A. 2003. Comparison of pattern-recognition techniques for classification of Silurian sedimentary rocks from Lithuania based on geochemical data. *Norwegian Journal of Geology*, **84**, 117–124.

PASKEVICIUS, J. 1997. *Geology of the Baltic Republics*. Vilnius University and The Geological Survey of Lithuania.

REYMENT, R. A. 1991. *Multidimensional palaeobiology*. Pergamon Press, Oxford (Appendix by L. F. MARCUS).

REYMENT, R. A. & SAVAZZI, E. 1999. *Aspects of multivariate statistical analysis in geology*. Elsevier, Amsterdam.

REYMENT, R. A., BLACKITH, R. E. & CAMPBELL, N. A. 1984. *Multivariate morphometrics* (2nd edn). Academic Press, London.

SEBER, G. A. F. 1984. *Multivariate observations*. Wiley and Sons, New York.

WEIHS, C. 1995. Canonical discriminant analysis: comparison of resampling methods and convex hull approximations. *In*: KRZANOWSKI, W. J. (ed.) *Recent advances in descriptive multivariate analysis*. Oxford Science Publishing, Oxford, 35–50.