2004 10 08

The Statistical Analysis of Multivariate Serological Frequency Data

Richard A. Reyment

Paleozoologiska avdelningen, Naturhistoriska Riksmuseet, Box 50007, 10405 Stockholm, Sweden

Running Head

MULTIVARIATE ANALYSIS OF SEROLOGICAL DATA

The Statistical Analysis of Multivariate Serological Frequency Data

Richard A. Reyment

Paleozoologiska Avdelningen, Naturhistoriska Riksmuseet, Box 5007, 10405 Stockholm
EMAIL: richard.reyment@nrm.se

Abstract: Data occurring in the form of frequencies are common in genetics, for example, serology. Examples are the AB0 group, the Rhesus group, but also DNA data. The statistical analysis of tables of frequencies is carried out by the available methods of multivariate analysis with usually three principal aims. One of these is to seek meaningful relationships between the components of a data-set, the second is to examine relationships between populations from which the data have been obtained, the third is to bring about a reduction in dimensionality. This latter aim is usually realized by means of bivariate scatter diagrams using scores computed from a multivariate analysis. The multivariate statistical analysis of tables of frequencies cannot safely be made by standard multivariate procedures because they represent compositions and are therefore embedded in simplex space, a subspace of full space. Appropriate procedures for simplex space are compared and contrasted with simple standard methods of multivariate analysis ("raw" Principal Component Analysis). The study shows the differences between a log-ratio model and a simple logarithmic transformation of proportions may not be very great, particularly with respect to graphical ordinations, but important discrepancies do occur. The divergencies between logarithmically based analyses and raw data are, however, great. Published data on rhesus alleles observed for Italian populations are used to exemplify the subject.

Introduction

One of the most common types of observations occurring in applied Genetics concerns compositions, a significant aspect of which is that the data are in the form of frequencies, proportions or percentages, and all of which have the common property that the rows of the data-matrix sum to a constant. This may not seem to be much of an obstacle but there is indeed a geometrical stumbling block involved that may be severe enough as to distort, or even invalidate, an analysis.

The study of compositions is essentially concerned with the relative magnitudes of "ingredients" and not their absolute values such as is the case for, say, measurements on a skull. These ingredients are not variables in the accepted sense of that term in statistics, but *parts*. What justifies this distinction? Consider any vector **x** with non-negative elements

$$x_1 + x_2 + \ldots x_D = 1 \qquad (1)$$

This vector is subject to the "unit-sum constraint", that is, one where a composition **x** is composed of D parts summing to 1. The components of (1) cannot be independent because they are constrained to sum to the same value.

The characteristic features of a compositional data-set are:

a) Each row of the $N \times D$ data-matrix corresponds to a single object (in the present connexion, a biological population or equivalent).

b) Each column of the data-matrix represents the frequency of a single part (for example, an allele).

c) Each row of the data-matrix sums to 1 (for proportions), respectively, 100 (for percentages).

(d) Correlations fluctuate erratically when one or more of the parts is removed from the data matrix (or a new part is added) because of the necessity of re-establishing the constant row-sum.

e) Each entry in the data-matrix is non-negative.

Property d) provides part of the key to understanding the complexity of compositional data. Correlations computed for "normal" data-matrices are invariant to the number of variables included. If you delete one or more variables from the data-set of measurements on some anatomical feature, this has no effect on the correlations between the remaining variables. Deleting a part does, however, change correlations between all remaining parts. Removing the proportion of CaO from a chemical array of values does, for example, influence all other oxides in an unpredictable way, row by row, each of which will after the deletion have a row-

sum differing generally from all other rows of the array and which must be restored to the constant-sum state.

Is all of this a recent discovery? Not at all, and in fact the problem of spurious correlation is perhaps one of the oldest in biometry, but probably the one that is least observed in practice. The main founding father of Biometry, Karl Pearson, wrote in 1897 in his essay on the mathematical foundations to the theory of evolution that a form of spurious correlation may arise when indices are used in the measurement of organisms. The theoretical discussion was backed up using data provided by W. F. R. Weldon for Plymouth shrimps the measurements on which were "standardized" by dividing them by the body-length, thus transforming the original distance-measures into proportions. The modern theory of compositional data analysis is due to Aitchison (1983, 1986, 1997).

Subcompositions: The formation of a subcomposition is not merely a matter of deleting a part from each composition. If *S* is any subset of the parts 1, ..., *D* of a *D*-part composition **x**, and $\mathbf{x}_S$ is the subvector formed from the corresponding components of **x**, then $C(\mathbf{x}_S)$ is called the subcomposition of the parts *S* (Aitchison, 1986, p. 196). The significance of this can be seen from the following example for 5 parts from which parts 1, 4 and 5 are selected to form a subcomposition.

$$(s_1, s_2, s_3) = C(x_1, x_4, x_5)$$

Geometrically, this is a transformation from the original sample space $\mathbf{S}^4$ to a new simplex $\mathbf{S}^2$. An important property of compositional data, and one that overrules the "leave-one-out" manipulation, is that the ratio of any two components must be the same as the ratio of the corresponding two components in the full, original composition. Hence,

$$s_i / s_j = x_i / x_j \qquad (2)$$

which is the attribute of "preserved ratio relationships".

The concepts of covariance and correlation in simplex space.

1. The problem of negative bias. A correlation coefficient computed between two parts is not free to range over the interval (-1, +1). Thus, in the case of two parts, say alleles A and B of the AB0 relationship,

$$\text{Corr}(x_1, x_2) = -1$$

and the product-moment is constrained to taking a specified value.

2. There is no relationship between the product-moment correlations of a subcomposition and those of the full composition. As the dimensionality of a subcomposition is decreased, so do

the crude covariances/correlations fluctuate in sign, which is an outcome of the incoherency of the product-moment correlation coefficient in simplex space (Aitchison, 1997) .

3. The concept of null correlation in reference to simplex space does not have the same meaning with respect to independence as is the case for full-space data. Many futile attempts have been made in the past in geochemistry and analytical chemistry to define a zero correlation in simplex space.

4. The concept of perturbations within the simplex is another fundamental property of compositional data. A perturbation with the original composition **x** is operated upon by the perturbing vector **u** to form a perturbed composition $\mathbf{X} = \mathbf{u}^{o}\mathbf{x}$.. This is familiar to geneticists as the relationships of genotypes before and after selection (Edwards, 2000, Chapter 2).

The logical necessities of scale-invariance, subcompositional coherence and perturbation as fundamental operations in the simplex led Aitchison (1986, 1997) to adopt certain log-ratio forms of defining patterns of compositional variability. These are compatible with the *additive logistic normal class* of distributions on the simplex . One example is the set of final divisor log-ratios

$$y_i = \log(x_i/x_D) \qquad (i = 1, ..., D-1) \qquad (3)$$

As a general rule, in dealing with the statistical analysis of compositions, the appropriate steps are (1) to express the problem in terms of log-ratios of the components; (2) apply the appropriate multivariate methodology for unconstrained vectors to the data, which are now in real space and free of the constant-sum constraint.

Log-ratio covariance-matrices

Thee are three equivalent representations of log-ratio covariances, the variation matrix, the log-ratio covariance matrix and the centred log-ratio covariance matrix; each of them can be derived from either of the others by simple matrix operations (Aitchison,1986, Chap. 4).

The variation matrix **T**

$\mathbf{T}_{D\times D} =$ [var{log(xi/xj)}; $i,j = 1, ..., D$]

This matrix is symmetric with a diagonal of zeros. Although it is not in the form of a covariance matrix it has certain computational advantages in that it treats the parts of a composition symmetrically (i.e. all parts are included on an equal footing). It can be transformed to either of the other covariance representations by means of a simple manipulation (Aitchison, 1986, p. 82).

The log-ratio covariance matrix **S**.

The log-ratio covariance matrix is the covariance matrix of a *d*-dimensional random vector

$\mathbf{y} = \log(\mathbf{x}_{-D} / x_D)$. This vector is located in space R^d. Part D is held fixed, which means that the last component of the vectors of parts in the present representation, x_D, becomes the common divisor of all the log-ratios. Aitchison (1986, p. 92) proved that the order of parts, and the choice of a component divisor, does not influence the outcome of a multivariate analysis. The log-ratio covariance matrix is positive definite and hence has a normal inverse. For discriminant functions and canonical variate analysis, this covariance matrix is often to be preferred.

Centred log-ratio covariance matrix **G**

A symmetric treatment of all D parts of a vector of compositions may be achieved by replacing the single component divisor x_D by the geometric mean of all D components. For a D-part composition, the centred log-ratio covariance matrix of the D-dimensional random vector

$$\mathbf{z} = \log\{\mathbf{x}/g(\mathbf{x})\}$$

where $g(\mathbf{x}) = (x_1, ..., x_D)^{1/D}$ is the geometric mean of the parts, is

$$\mathbf{G} = \quad \text{cov}\left[\log(x_i/g(\mathbf{x})), \log(x_j/g(\mathbf{x}))\right] \qquad (4)$$

This matrix is the one that is interpretationally most useful for many multivariate analogues of full-space statistics. It is easy to explain in that it is symmetric with respect to all parts. The drawback is that is singular and hence does not possess a "normal" inverse and hence where relevant requires a generalized matrix inverse.

Log-contrast principal component analysis

For the purposes of the present exposition, the multivariate method chosen is the widely used one of principal component analysis (Aitchison, 1983; Aitchison, 1986, p. 190) well known from many spheres of quantitative biology.

The covariance matrix used as input is that of equation (4), the centred log-contrast covariance matrix. A log-contrast of a D-part composition $\mathbf{x}$ is defined as any log-linear combination $\mathbf{a}'\log\mathbf{x}$ with

$$a_1 + \ldots + a_D = 0$$

The principal component analysis follows from the reduction of a centred logratio covariance matrix in the usual manner by finding the latent roots and vectors satisfying.

$$(\mathbf{G} - ?_i\mathbf{I})\mathbf{a}_i = \mathbf{0} \qquad (5)$$

Appropriately constructed analogues are available for the multivariate statistical procedures of principal components, principal coordinates, Gabriel's biplot, canonical correlation, discriminant functions and generalized distances, and canonical variates (Aitchison, 1986, Chapters 8 and 9). One particular difficulty that may be mentioned concerns the issue of achieving stability in vector components (Campbell,1979, 1980). A further reference is Reyment and Savazzi (1999, pp 131-133). The rationale for observing the restrictions imposed by simplex space have taken time to achieve recognition in statistical genetics. This situation seems now to be in the process of change and Romano *et al*. (2003) have taken note of this in their work on microsatellite and mtDNA data for Sicily. In that paper, the computations were based on a computer program of Reyment and Savazzi (1999); note, however, that their results seem to have been obtained for the inverse (Q-mode) log-contrast principal component model known as principal coordinate analysis.

Exemplification using Principal Component Analysis. The exemplificationary data were selected from Mourant et al. (1976) from tables of Rhesus blood groups in terms of the CDE standard. Two sets of observations on Italian populations were selected, to wit, Mourant et al. p. 406, Table 4.13 (N=15) and p. 438, Table 4.19 (N=16). There are eight combinations of the three alleles of which six were chosen for the present study, owing to the rarity of two in the populations considered. Subcompositional coherence was maintained in the selection exercise involved in reducing the data set from eight parts to six (Aitchison, 1983, 1986). The set encompassing 15 populations were obtained by tests with anti-C, -D, -E and -c and that for 16 populations by tests for anti-C, -D, -E, -c and -e. The two sets were pooled. These data are typical of what can be expected in serology.

Three sets of principal component computations were carried out, the centred log-ratio analysis, an analysis using only the logarithms of the raw percentages, and analysis on the raw percentages. The $D \times D$ crude log covariance matrix, say $\mathbf{K}$, is defined as:

$$\boldsymbol{K} = [\mathrm{cov}(\log x_i, \log x_j)]. \qquad (5)$$

This is the representation used in the following.

Findings

Comparing latent vectors **.** Table 1 contrasts the three latent vectors connected with the three largest latent roots and the latent vector connected to the smallest valid latent root (i.e

the (*p-1*)th) for the centred log-ratio covariance matrix (CLR), the covariance matrix of log-transformed raw data (LRD), equation (6), and the covariance matrix of raw frequencies (RD). The results for the CLR and LRD vectors differ for the first pairing (i.e,. first latent vectors), less so for the second and third latent vectors. The comparisons for the angles between latent vectors presented in Table 2 expresses this approximately. The finding for the fifth latent vectors is interesting in that the pair are almost identical for both. The angle between these vectors measures 5.56 degrees (0.097 radians). Gower (1967) suggested that the "smallest" latent root of a principal component analysis was worthy of consideration because it represents the direction of least variability. Aitchison (1986, p. 189) noted that for many situations log-transformed data may well lead to a successful principal component analysis in relation to that yielded by the model appropriate to the properties of the additive log-normal distribution. This appears to be the case in the present exemplification in which there is little variation in some of the $\log x_i$.(I am indebted to a referee for bringing this point to my notice). In all cases, the vectors for the raw data differ markedly from the logarithmically based counterparts (Table 1).

Discussion of the ordinations. The plots of the principal component scores for the first two axes for both logarithmically based data-sets yield approximately the same pattern of dispersal points for the CLR and LRD matrices. A notable feature for both samples is that the dispersions express the same fragmentation into two geometrically distinct distributions and indicate that both are heterogeneous in much the same way. The CLR plot shown in Figure 1 contains four outliers. The plot for the LRD set displays two clusters and two outliers (!) with a similar apportionment of points in the clusters (Fig. 2). In both graphs the shapes of the clusters fall into the same two types, the one roughly linear and suggesting close association in the scores and the other a more diffuse bundle of points, the major axis of which lies roughly at right angles to that of the other cluster. The ordination obtained for the scores of the first and second principal components for the raw data is amorphous and it does not appear to contain obviously useful information (Fig. 3). It should be mentioned that the cluster in the lower part of Fig. 3 does not encompass the same set of observations as the similarly shaped constellation of points in Figs. 1 and 2.

As noted above, the graphical analysis is that the data for both logarithmic sets of observations are not homogeneous and that they consist of two well separated clusters. From the statistical point of view, the strongly manifested heterogeneity casts grave doubt on the

validity of any reification made of the latent vectors due to the problem of stability of the components (Campbell, 1979). Inspection of the original data-array discloses that there is a difference between the two major clusters occurring in the ordination for the log-contrast analysis, namely, that one of them (N=13) for contains no non-zero entries, whereas the other contains several zero entries for CdE. The outliers are due to zero observations for Cde and CdE. (The corresponding cluster for the raw logarithms embraces 14 points.) The subset of 13 observational vectors with no non-zero entries was isolated for separate analysis. The resulting ordination is shown in Fig. 4. The points are widely dispersed but with a tendency for grouping in the lower portion of the graph.

Summary and Conclusions

Compositional data possess specific properties such that special adaptations of multivariate statistical analysis are required in order to bring about a geometrically useful analysis. It is shown in the present note that the log-transformed raw proportions, can yield results that are not too far from what is obtained by applying a log-ratio procedure (though not identical as shown by the identification and geometrical locations of atypical observations) particularly where there is little variation in the data as is the case in the present exemplification. This is most clearly manifested for the graphical ordination of principal component scores. It is also demonstrated that abundant zero observations can markedly distort a multivariate compositional analysis. Aitchison (1986, p. 271) concluded that the Box-Cox transformation may in dire cases provide a practical special alternative to the logistic normal model (Box and Cox, 1964).

It is often desired to reify (interpret) the results of a principal component analysis. Caution is called for here, however, since elements of the latent vectors can, and do, often fluctuate wildly under repeated sampling, which implies that any such analysis requires adequate controls on the stability of the vectorial components. This problem has been closely studied by Campbell (1979, 1980).

Acknowledgements

The Trustees of the Swedish Museum of Natural History, Stockholm are thanked for providing working facilities. I wish to thank both referees for valuable comments and constructive criticism. I also wish to acknowledge the instructive discussions I have had with Dr. Peter Forster, Cambridge University, over a period of years.

References

Aitchison, J. (1983). Principal component analysis of compositional data. Biometrika, 70, 57-63.

Aitchison, J.(1986). The statistical analysis of compositional data. Chapman and Hall, London, xv + 416 pp.

Aitchison, J. (1997). One hour course in compositional data-analysis, or compositional data-analysis is easy, pp. 3-35 in Proceedings of the 1997 Annual Conference of the International Association for Mathematical Geology, Universitat Politècnica de Catalunya, edited by Vera Pawlowsky-Glahn, Barcelona.

Box, G. E. P. and Cox, D. R. 1964. The analysis of transformations. Journal of the Royal statistical Society, B 265, 211-252.

Campbell, N. A. 1979. Canonical Variate Analysis. PhD. Thesis. Imperial College, University of London.

Campbell, N. A. 1980. Shrunken estimators in discriminant and canonical variate analysis. Applied Statistics, 29, 5-14.

Edwards, A. W. F.(2000). Foundations of Mathematical Genetics. Cambridge University Press, 2nd Edition, 121 pp.

Gower, J. C. (1967). Multivariate analysis and multidimensional geometry. The Statistician, 17,13-28.

Reyment, R. A., and E. Savazzi, (1999). Aspects of Multivariate Statistical Analysis in Geology. Elsevier Science B. V., Amsterdam, x + 285 pp.

Romano, V., Coli, C., Ragalmuto, A., D'Anna, R. P., Flugy, A., De Leo, G., Giambalvo, O., Lisa, A., Fiorani, O., Di Gaetano, C., Salemo, A., Tamouza, A., Chanon, D., Zei, G., Matullo, G., Piazza, A. (2003). Autosomal microsatellite and mtDNA genetic analysis in Sicily (Italy). Annals of Human Genetics, 67, 42-53.

Figure Captions

Fig. 1. Plot of the first and second principal component scores for the centred log-ratio data . There are two compact clusters oriented roughly at right angles to each other, and three markedly atypical values and one less strongly marked outlier, four in all.

Fig. 2. Plot of the first and second principal component scores for the log-transformed raw data. One cluster is strongly compacted (due to slight variability along the first axis of scores). The second cluster is more inflated than its analogue in Fig. 1. There are two atypical points.

Fig. 3. Plot of the first and second principal component scores for the raw data. The points in the spread-out cluster at the bottom of the figure do not correspond to either of the clusters in Figs 1 and 2, but comprise a mixture of both.

Fig. 4. Plot of the first and second principal component scores for the centred log-ratio data for the reduced array of all entries greater than zero (N=13). The identities of the points are Salerno (1), Bologna (2), Lazio (3), Lazio (4), Lazio (5), Sardinia (6), Palermo (7), Udine (8), Lazio (9), Pavia (10), Pavia (11), Palermo (12), Trentino (13). These points do not correspond to those forming the cluster referred to in Fig. 3.

–

Table 1. Comparison of the first three latent vectors and smallest latent vector for the centred log-ratio covariance matrix (CLR), the covariance matrix of logarithms of raw data (LRD) and the covariance matrix of raw proportions (RP).

	First latent vector			*Second latent vector*		
allele	CLR	LRD	RP	CLR	LRD	RP
CDE	0.197	-0.016	- 0.830	0.026	0.006	-0.395
Cde	0.031	0.180	0.043	-0.811	-0.982	0.184

cDE	0.170	0.010	0.143	0.007	-0.015	0.049
cdE	0.321	-0.078	0.047	0.554	0.046	0.568
cDe	-0.893	0.980	0.059	0.185	0.184	0.010
cde	0.174	0.001	0.532	0.039	0.009	-0.696
?	18 .424	22.457	98.608	4.395	4.507	22.146

	Third latent vector			*Fifth latent vector*		
	CLR	LRD	RP	CLR	LRD	RP
CDE	-0.376	-0.018	0.054	0.165	0.261	-0.020
Cde	0.418	0.060	-0.013	-0.013	-0.012	-0.912
cDE	-0.365	-0.018	-0.853	0.604	0.600	0.131
cdE	0.651	0.996	0.440	0.018	0.019	0.203
cDe	0.027	0.068	0.077	0.006	0.007	0.326
cde	-0.355	0.004	0.266	-0.779	-0.756	-0.048
?	2.651	4.125	8.561	0.356	0.095	0.818

? denotes latent roots.

Table 2: Comparisons between selected pairs of latent vectors

---

Comparison between	*Vector No.*	*Degrees*
CLR_1 and LRD_1	1	26.86
CLR_2 and LRD_2	2	30.80

CLR_3 and LRD_3	3	46.44
CLR_5 and LRD_5	5	5.56
CLR_1 and LRD_1	1	18.00

(N=13; non-zero entries)

CLR = centred log-ratio latent vectors; LRD = latent vectors logarithms of raw proportions