

Multiple group principal component analysis and population differentiation

R. S. THORPE

Department of Zoology, University of Aberdeen, Tillydrome Avenue, Aberdeen, AB9 2TN, Scotland, UK

(Accepted 11 December 1987)

This paper explores the requirements and advantages of multiple group principal component analysis (MGPCA) when it is used to investigate population differentiation. A distinction is drawn between equality of orientation of the within-group axes and equality of variance along these axes. Several examples of the use of MGPCA are discussed and it is shown that MGPCA *per se* does not require equality of variance along the axes although it may be a requirement of some of the techniques subsequently used to analyse the component scores. MGPCA is simple and direct, being based on the mathematically well defined eigenvector analysis of a symmetric positive definite (pooled within-group covariance) matrix and it can be thought of as a step in the computation of canonical variate analysis (CVA). It can be used with CVA (which is the most popular method of biometrically assessing population affinities) to assess the contribution of within-group components to among-group discrimination. It is also one of a range of appropriate techniques that can be used to define (and delete if required) within-group growth effects and is particularly suitable when CVA is being used to assess the population affinities. When used in this way it has the advantage of being more influenced by the groups with the greatest growth range.

Principal component analysis (PCA) of a covariance matrix transforms a set of variables into orthogonal (uncorrelated) components. In order to elucidate the requirements of the pertinent techniques related to PCA, the discussion is based on orthogonal size and shape components derived from linear measurements taken from some aspect of the morphology. This is one of the common biological uses of PCA and related techniques.

When these linear measurements are taken with a view to comparing groups such as geographic populations or sexes, then it is essentially a multiple group problem rather than a problem for simple PCA. The samples should not be pooled irrespective of group because the within-group components (such as growth size) and the between-group differentiation perturb one another (Thorpe, 1976, 1983*a*). Nevertheless, there are a range of ordination and regression techniques (reviewed by Thorpe, 1983*a*) that recognize these problems and are suitable for multiple groups. Common principal component analysis (CPCA), as expounded by Airoldi and Flury in this volume (Airoldi & Flury, 1988), is a welcome addition to this set of techniques.

However, in promoting CPCA, Airoldi & Flury (1988) suggest that another suitable technique, multiple group principal component analysis (MGPCA), has more stringent requirements than CPCA, i.e. they suggest that it requires equality of the within group covariance matrices. Since in MCPCA eigenvectors and eigenvalues are extracted from the pooled within group covariance matrix, one can think of the equality of the covariance matrices as having two facets; 1) the equality of the orientation of the components in character hyperspace as defined by the eigenvectors and; 2) the equality of the spread (variance) of cases along the components as defined

by the eigenvalues. Equality of the eigenvectors (i.e. orientation of the components in hyperspace) is assumed in all uses of MGPCA but equality of the eigenvalues (i.e. variance of the cases along the components) is irrelevant to many applications of MGPCA and in these cases is not a requirement.

Various examples of the use of MGPCA illustrate when equality of the eigenvalues (and hence full equality of the covariance matrices) is, and is not, required:-

Example 1. Two of the advantages of MGPCA are its simplicity and direct relationship to canonical variate/discriminant function analysis (CVA) which is the most frequently used ordination technique in evolutionary studies (Thorpe, 1983a). If one runs MGPCA on a set of characters and then inputs the component scores into CVA this gives identical results to CVA on the original set of characters (Thorpe *et al.*, 1982; Thorpe, 1983a, b; Thorpe & Leamy, 1983). The advantage of the former procedure is that it enables one to assess the contribution of within-group components to the between-group discrimination.

This procedure reveals that there is no correlation between the percentage variance reflected by the within-group components and their contribution to between-group discrimination. Consequently, in studies using intercorrelated linear measurements, the within-group growth/size component may reflect a very large proportion of the within-group variance but have only a minimal influence on the between-group discrimination in CVA (Thorpe *et al.*, 1982; Thorpe 1983a, b; Thorpe & Leamy, 1983; Wiig, 1986; Thorpe & Baez, 1987). A specific example of this is Wiig's (1985) study of male feral American minks where the 'general size' component accounted for almost 60% of the within-group variation but less than a half of a percent of the between-group variation.

If one inputs all the MGPCA components into CVA, then the requirements for MGPCA are the same as for CVA, i.e. equality of the within group covariance matrices (with both the eigenvectors and eigenvalues being equivalent). However, even in these cases, one must bear in mind that minor departures from inequality of the eigenvalues are likely to have only a trivial influence on the relative similarity of the groups.

Example 2. In practice, the most likely cause of inequality of variance along a component when linear measurement of organisms are analysed is inequality in the range of sizes. Inequality of variance along the size component occurs when one population is represented by a sample covering a wide growth range whilst another sample covers only a narrow growth range. In these circumstances (and when a bias in the growth/size between samples occurs as in Corti *et al.*'s (In press) study of fish stocks), then one may wish to forego any between-group discrimination contributed by size and exclude the size/growth vector (Thorpe, 1983a, b). When the source of inequality of the variances is removed by the exclusion of the first component, then a CVA of the $n-1$ components (i.e. a size-out CVA) is appropriate even though the original covariance matrices are not equal.

Example 3. In the previous examples, MGPCA has been linked with CVA but the component scores from MGPCA can be subjected to a range of other techniques which use all, or a selection of, the components (Thorpe *et al.*, 1982; Thorpe, 1983b, Thorpe & Baez, 1987). Moreover, even when the techniques do require equality of the variances, then the selected components may not be the ones with unequal variances. In these cases, MGPCA requires (like CPCA) only the eigenvectors to be equal and not the original covariance matrix. An example of this use of

MGPCA is Thorpe & Baez's (1987) investigations of the lizard *Gallotia galloti* on Tenerife where only the 'size' component and one shape component (no. 7) have the group mean scores mapped on to the localities to reveal the pattern of microgeographic variation.

A further point raised by Airoidi & Flury (1988, and in their reply to these comments), and also previously by Somers (1986), is that in MGPCA the groups with the greatest variance will have more influence on the covariance matrix than those with less variance. In general, this is a trivial problem. As explained above, in the cases where inequality of growth range is the cause of the inequality of variance (as is generally the case with this type of application), then the group with the greatest growth range will have more influence on the direction of the growth vector in MGPCA. However, a narrow growth range may not allow the direction of the growth vector to be adequately defined. Since the growth vector is better defined by a broad range of growth stages than a narrow range, it is an advantage (not a disadvantage) that in MGPCA the direction of the growth vector is more influenced by the groups with a greater growth range.

To sum up. First, for the types of use under consideration, MGPCA, like CPCA, requires equality of the orientation of the components in hyperspace (i.e. equality of eigenvectors). It does not require equality of the variance along the component (equality of the eigenvalues) unless this is a prerequisite of the techniques used to analyse the component scores. Second, MGPCA has the advantage that it allows samples with a greater growth range to have more influence on the direction of the growth vector. Third, if one wishes to go further than defining common components and evaluate the population affinities using the popular technique of CVA, then the related technique of MGPCA is more appropriate than CPCA. Finally, MGPCA is a straightforward technique based on mathematically well-defined eigenvector analysis of a symmetric positive definite matrix. It is a part of the process of CVA (Campbell & Atchley, 1981) which is well established and needs no further mathematical foundation.

In conclusion, neither CPCA nor MGPCA, or any of the other pertinent multi-group techniques (Thorpe, 1983a), has general superiority. Rather, a given technique will be more or less appropriate according to the specific requirements of the application.

REFERENCES

- Airoidi, J.-P. & Flury, B. K. (1988). An application of common principal component analysis to cranial morphometry of *Microtus californicus* and *M. ochrogaster* (Mammalia, Rodentia). *J. Zool., Lond.* **216**: 21–36.
- Campbell, N. A. & Atchley, W. R. (1981). The geometry of canonical variate analysis. *Syst. Zool.* **30**: 268–280.
- Corti, M., Thorpe, R. S., Sola, L., Sbordoni, V. & Cataudella, S. (In press). Multivariate morphometrics in aquaculture: a case study of six stocks of the common carp, *Cyprinus carpio* from Italy.
- Somers, K. M. (1986). Multivariate allometry and the removal of size with principal component analysis. *Syst. Zool.* **35**: 359–368.
- Thorpe, R. S. (1976). Biometric analysis of geographic variation and racial affinities. *Biol. Rev.* **51**: 407–452.
- Thorpe, R. S. (1983a). A review of the numerical methods for recognizing and analysing racial differentiation. In *Numerical taxonomy*: 404–423. Felsenstein, J. (Ed.). Berlin & New York: Springer-Verlag.
- Thorpe, R. S. (1983b). A biometric study of the effects of growth on the analysis of geographic variation: Tooth number in Green geckos (Reptilia: *Phelsuma*). *J. Zool., Lond.* **201**: 13–26.
- Thorpe, R. S. & Baez, M. (1987). Geographic variation within an island: univariate and multivariate contouring of scalation, size and shape of the lizard *Gallotia galloti*. *Evolution, Lawrence, Kans.* **41**: 256–268.
- Thorpe, R. S., Corti, M. & Capanna, E. (1982). Morphometric divergence of Robertsonian population/species of *Mus*: A multivariate analysis of size and shape. *Experientia* **38**: 920–923.

- Thorpe, R. S. & Leamy, L. (1983). Morphometric studies in inbred and hybrid house mice (*Mus* sp.): Multivariate analysis of size and shape. *J. Zool., Lond.* **199**: 421–432.
- Wiig, O. (1985). Multivariate variation in feral American male mink (*Mustela vison*) from Southern Norway. *J. Zool., Lond. (A)* **206**: 441–452.
- Wiig, O. (1986). Sexual dimorphism in the skull of minks *Mustela vison* and otters *Lutra lutra*. *Zool. J. Linn. Soc.* **87**: 163–179.