Journal of Classification 7:81-98 (1990)

Between-Group Analysis with Heterogeneous Covariance Matrices: The Common Principal Component Model

W. J. Krzanowski

University of Reading, U.K.

Abstract: Analysis of between-group differences using canonical variates assumes equality of population covariance matrices. Sometimes these matrices are sufficiently different for the null hypothesis of equality to be rejected, but there exist some common features which should be exploited in any analysis. The common principal component model is often suitable in such circumstances, and this model is shown to be appropriate in a practical example. Two methods for between-group analysis are proposed when this model replaces the equal dispersion matrix assumption. One method is by extension of the two-stage approach to canonical variate analysis using sequential principal component analyses as described by Campbell and Atchley (1981). The second method is by definition of a distance function between populations satisfying the common principal component model, followed by metric scaling of the resulting between-populations distance matrix. The two methods are compared with each other and with ordinary canonical variate analysis on the previously introduced data set.

Keywords: Between-group analysis; Canonical variate analysis; Common principal component model; Eigenvalues and eigenvectors; Matusita distance between populations; Metric scaling; Principal component analysis.

Financial support for part of this work was provided by a research grant from Shell Research Limited. All three referees made constructive comments which are gratefully acknowledged.

Author's Address: W. J. Krzanowski, Department of Applied Statistics, P.O. Box 217, University of Reading, Whiteknights, Reading RG6 2AN, U. K.

1. Introduction

We consider situations where individuals in a sample are divided *a priori* into a number of groups and the objective of the statistical analysis is to highlight differences between the groups. Suppose that the p variables $X_1, \ldots, X_p$ are recorded for each of n individuals, and that the *a priori* division of these individuals is into g groups with n_i individuals in the i-th group (so that $\sum_{i=1}^{g} n_i = n$). Write $\mathbf{X}^T = (X_1, \ldots, X_p)$, and let the p observations for the j-th individual of the i-th group be contained in the vector $\mathbf{x}_{ij}^T = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})$ where $i = 1, \ldots, g$ and $j = 1, \ldots, n_i$.

The classical approach to the problem is via canonical variate analysis. The *canonical variates* Y_k associated with the g groups of individuals are linear combinations of the original variates, i.e., $Y_k = \mathbf{a}_k^T \mathbf{X}$ for $k = 1, \ldots, s$ where $s = \min(g-1, p)$. The coefficient vector $\mathbf{a}_k^T = (a_{k1}, a_{k2}, \ldots, a_{kp})$ is the eigenvector corresponding to the k-th largest eigenvalue λ_k of the equation $|\mathbf{B} - \lambda \mathbf{W}| = 0$ in which

$$\mathbf{B} = \frac{1}{g-1} \sum_{i=1}^{g} n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T$$

is the between-groups covariance matrix and

$$\mathbf{W} = \frac{1}{n-g} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T$$

is the (pooled) within-groups covariance matrix. The vectors $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{x}}$ denote the mean vectors in the i-th group and over the whole sample of individuals respectively. Thus the $(\lambda_k, \mathbf{a}_k)$ pairs satisfy the equation $(\mathbf{B} - \lambda_k \mathbf{W})\mathbf{a}_k = \mathbf{0}$, and these pairs are ranked in descending order of the λ_k. The values of the canonical variates for each individual are known as the *canonical variate scores* for that individual. The canonical variate scores for the n individuals can be represented as a set of n points in s-dimensional Euclidean space (the *canonical variate space*), and the first r dimensions of this space provide the optimal r-dimensional configuration of sample points in which to highlight the differences between the groups. The group mean vector $\bar{\mathbf{x}}_i$ transforms to $\bar{\mathbf{y}}_i^T = (\bar{y}_{i1}, \ldots, \bar{y}_{is})$ in the canonical variate space (where $\bar{y}_{ik} = \mathbf{a}_k^T \bar{\mathbf{x}}_i$ for $k = 1, \ldots, s$), and the squared Euclidean distance between $\mathbf{y}_i$ and $\mathbf{y}_j$ in this space is equal to the Mahalanobis D^2 value $(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)^T \mathbf{W}^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)$ between the i-th and j-th groups in the original data space. We will assume throughout that $n - g \geq p$, so that $\mathbf{W}$ is positive-definite and symmetric hence non-singular.

A full development of the above ideas is given in most standard texts on multivariate analysis (e.g., Mardia, Kent and Bibby 1979; Krzanowski 1988), and most statistical package computer implementations of the technique (e.g., BMDP, SAS, GENSTAT, etc.) are based on solution of the generalized eigensystem $(\mathbf{B} - \lambda \mathbf{W})\,\mathbf{a} = \mathbf{0}$. However, there are some alternative formulations which prove useful and which we exploit below. It has long been known that canonical variate analysis can be viewed geometrically as a two-stage rotation. In the first stage the rotation is on the original variables, while in the second stage it is on the group means after transformation to the variables formed at the first stage. Campbell and Atchley (1981) set out the details of this approach. On the other hand, if the intermediate canonical variate coefficients $\mathbf{a}_k$ are not of intrinsic interest but the main objective of the analysis is to obtain the plot of group means in the canonical variate space, then Gower (1966) shows that this objective can be met by a principal coordinate analysis (metric scaling) of the $(g \times g)$ matrix containing the Mahalanobis D^2 values between every pair of groups.

Whichever approach to between-group analysis is adopted, however, a number of assumptions must be satisfied for the analysis to be valid. The individuals in the i-th group are assumed to be a random sample from a population which has mean μ_i and dispersion matrix Ω_i $(i = 1, \ldots, p)$. For some aspects of the analysis these populations must be normal, but this assumption will not concern us here and we will take it as true when necessary. What we focus on is the assumption that all population dispersion matrices are equal, i.e., $\Omega_i = \Omega$ for all i. This assumption must be satisfied if the analysis is to make sense, as W is then an estimate of the common dispersion matrix Ω. A preliminary test of the null hypothesis $H_0 : \Omega_i = \Omega$ for all i can be conducted using a likelihood-ratio test (see e.g., Krzanowski 1988, p.370), and if the result is not significant then a canonical variate analysis can be effected without qualms. If H_0 is rejected, however, then computation of W involves a pooling of heterogeneous covariance matrices and the results of the subsequent canonical variate analysis will not be reliable. Relatively little guidance is available in the literature for dealing with such circumstances. A computationally iterative generalization of canonical variate analysis to the case of heterogeneous covariance matrices has been given by Campbell (1984), while a graphical technique has recently been proposed by Young, Marco and Odell (1987). In both cases no assumptions are made about the structure of the Ω_i.

In many cases where the null hypothesis of equality of dispersion matrices is rejected, there may nevertheless be some additional structure linking these matrices. Assuming them to be completely disparate will thus be unsatisfactory from a statistical point of view, as not only is potentially useful information ignored but also relatively few degrees of freedom are available

for estimation of each separate dispersion matrix. It is better to embody available information in a model, and thereby to improve precision of estimation. Recently, Flury (1988, Chapter 7) has outlined a hierarchical set of models for a collection of dispersion matrices and has studied one of these models, the common principal component model, in some depth. This model appears to be a good intermediate stage between the extremes of equal dispersion matrices or completely disparate dispersion matrices, and seems to be applicable in a range of practical situations. We briefly review the model in Section 2, and show its relevance for a particular set of data. Using this model we then generalize the two-stage canonical variate method in Section 3, and provide a suitable distance measure between populations in Section 4. This distance function can be used as a basis for the metric scaling approach to between-group analysis. The two approaches are then compared in Section 5 with each other and with the standard canonical variate analysis, for the previously introduced data. Apart from questions of parsimony and increased degrees of freedom for estimation, an attractive feature of the common principal component model is that it leads to direct rather than iterative computational techniques.

2. The Common Principal Component Model

Suppose that we reject the null hypothesis $H_0 : \Omega_i = \Omega$ for all $i = 1, \ldots, g$ when it is tested against the general alternative H_a: at least one Ω_i differs from the rest. Flury (1988) argues that there may nevertheless be *some* similarity among all the Ω_i, which should be exploited in any subsequent analysis. He proposes various possible models, the most generally applicable of which is the *common principal component* model. This model states that the Ω_i share the same principal axes, but these axes may be of different sizes and rankings in the different populations. It is thus equivalent to the hypothesis that the Ω_i are all diagonalizable by the same orthogonal matrix, i.e., $H_c : \mathbf{L}^T \Omega_i \mathbf{L} = \mathbf{D}_i$ for $i = 1, \ldots, g$ where $\mathbf{L}$ is a $p \times p$ orthogonal matrix and the $\mathbf{D}_i$ are all diagonal matrices. Note that $H_0 \subset H_c \subset H_a$.

Estimates $\hat{\mathbf{L}}$ and $\hat{\mathbf{D}}_i$ for this model can be obtained using either maximum likelihood (if normality of populations is assumed) or least squares (if no distributional assumptions are made); FORTRAN routines have been provided for the former case by Flury and Constantine (1985), with an amendment by Clarkson (1988a), and for the latter case by Clarkson (1988b). The advantage of the normality assumption, if it is appropriate, is that a likelihood-ratio test of the model is then possible. In fact it is possible to test any of the three hypotheses H_0, H_c and H_a against each other if normality can be assumed. Maximum likelihood estimates of the Ω_i under each of these hypotheses are given by

$$\hat{\Omega}_i = \hat{\mathbf{L}}\,\hat{\mathbf{D}}_i\,\hat{\mathbf{L}}^T \qquad \text{(for } H_c\text{)}$$

$$S_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)^T \qquad \text{(for } H_a\text{)}$$

and

$$\mathbf{W} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)^T \qquad \text{(for } H_0\text{)}$$

Appropriate *L–R* test statistics and asymptotic null distributions are then as follows (Flury 1988, p.70; Krzanowski 1988, p.370):

For H_0 v. H_a

$$T_1 = n\log_e|\mathbf{W}| - \sum_{i=1}^{g} n_i \log_e |S_i|\,,$$

asymptotically χ^2 on $\frac{1}{2}p(p+1)(g-1)$ d.f. under H_0.

For H_c v. H_a

$$T_2 = \sum_{i=1}^{g} n_i \log_e |\hat{\Omega}_i| - \sum_{i=1}^{g} n_i \log_e |S_i|\,,$$

asymptotically χ^2 on $\frac{1}{2}p(p-1)(g-1)$ d.f. under H_c.

For H_0 v. H_c

$$T_3 = n\log_e|\mathbf{W}| - \sum_{i=1}^{g} n_i \log_e |\hat{\Omega}_i|\,,$$

asymptotically χ^2 on $p(g-1)$ d.f. under H_0.

Thus T_2 and T_3, and their associated degrees of freedom, are additive components of the test of H_0 v. H_a. (An alternative decomposition of this test has been given by Manly and Rayner 1987.)

To illustrate the usefulness of the common principal component model, consider the following set of data (kindly provided by Mr. B. Carroll of the British Council). One hundred and sixty Venezuelan students came to Britain

in 1975 to learn English, and were distributed among ten colleges in the North of England. The British Council was interested in determining whether there were differences in performances of these students between the different colleges. The students were given centrally-administered English tests in November 1975 (on arrival), in February 1976 and in June 1976 (on departure). Each test had six constituents: Comprehension, Essay, Cloze (exact), Cloze (acceptable), Structure and Dictation, each of which resulted in a score out of 25. Analysis of the November test results suggested that there were no systematic between-college differences and hence that random allocation of students to colleges had been successful. To investigate whether the colleges exhibited any differential teaching effects, therefore, canonical variate analysis of the June data was considered. Examination of plots and histograms of the data indicated that normality assumptions were reasonable, but in testing homogeneity of dispersions it was found that $T_1 = 255{\cdot}97$ on 189 degrees of freedom. This value is significant at the 0·1% level (Pearson and Hartley 1966, p.137), indicating that the hypothesis of equal dispersion matrices had to be rejected and that standard canonical variate analysis was thus inappropriate.

Consideration of the nature of the data, however, suggests that the common principal component model might be reasonable in the present circumstances. The difficulties encountered by Spanish-speaking students learning English should be fairly universal, irrespective of college attended, and the major sources of variation among students should be similar from college to college. However, each college will have its own teaching strengths, weaknesses and emphases (for example, some colleges may stress formal grammar learning while others may concentrate on aural aspects such as dictation or on creative skills such as essay writing). Variation in performance among students will tend to be reduced in those areas in which a college specializes, at the expense of the other major sources of variation. Thus the dispersion matrices in each college would be expected to have the same principal axes, but the sizes and rankings of these axes might vary from college to college. A decomposition of the likelihood-ratio test statistic T_1 in fact yielded the non-significant value $T_2 = 139{\cdot}39$ on 135 degrees of freedom and the significant ($p < 0{\cdot}001$) value $T_3 = 116{\cdot}58$ on 54 degrees of freedom. It was therefore concluded that the common principal model provides the best summary of this set of data.

3. Generalization of Two-stage Canonical Variate Analysis

Canonical variate analysis requires determination of the eigenvalue/eigenvector pairs $(\lambda_k, \mathbf{a}_k)$ satisfying $(\mathbf{B} - \lambda \mathbf{W})\,\mathbf{a} = \mathbf{0}$. These

eigenvalues and eigenvectors can be obtained equivalently in the following stages (see Campbell and Atchley 1981).

(i) Find the eigenvalues/eigenvectors $(\theta_k, \mathbf{u}_k)$ of $\mathbf{W}$. Write $\Theta = \text{diag}(\theta_1, \ldots, \theta_p)$ and $\mathbf{U} = (\mathbf{u}_1\ \mathbf{u}_2, \ldots, \mathbf{u}_p)$ so that $\mathbf{W} = \mathbf{U}\,\Theta\,\mathbf{U}^T$. Since $\mathbf{W}$ is assumed to be positive definite then $\theta_k > 0$ for all k and hence $\Theta^{-\frac{1}{2}} = \text{diag}(\theta_1^{-\frac{1}{2}}, \ldots, \theta_p^{-\frac{1}{2}})$.

(ii) Transform the original data vectors $\mathbf{x}_{ij}$ to $\mathbf{w}_{ij} = \Theta^{-\frac{1}{2}}\,\mathbf{U}^T\mathbf{x}_{ij}$ for $i = 1, \ldots, g$ and $j = 1, \ldots, n_i$. Hence $\bar{\mathbf{w}}_i = \Theta^{-\frac{1}{2}}\,\mathbf{U}^T\,\bar{\mathbf{x}}_i$ $(i = 1, \ldots, g)$ and $\bar{\mathbf{w}} = \Theta^{-\frac{1}{2}}\,\mathbf{U}^T\,\bar{\mathbf{x}}$. Thus the (weighted) covariance matrix of the $\bar{\mathbf{w}}_i$ is

$$\mathbf{C} = \frac{1}{g-1}\sum_{i=1}^{g} n_i(\bar{\mathbf{w}}_i - \bar{\mathbf{w}})(\bar{\mathbf{w}}_i - \bar{\mathbf{w}})^T = \Theta^{-\frac{1}{2}}\,\mathbf{U}^T\,\mathbf{B}\,\mathbf{U}\,\Theta^{-\frac{1}{2}}\ .$$

(iii) Find the eigenvalues and eigenvectors of $\mathbf{C}$. Then the eigenvalues are the required canonical roots $\lambda_1, \ldots, \lambda_p$ of the original data, and the eigenvectors $\mathbf{c}_1, \ldots, \mathbf{c}_p$ yield the canonical variate coefficients $\mathbf{a}_k$ from $\mathbf{a}_k = \mathbf{U}\,\Theta^{-\frac{1}{2}}\,\mathbf{c}_k$ $(k = 1, \ldots, s)$. Canonical variate group means are just the principal component scores of the $\bar{\mathbf{w}}_i$ in this second-stage principal component analysis.

There is a useful geometrical interpretation of the above algebraic steps. The original data form g groups of points in the p-dimensional space in which the original variates $X_1, \ldots, X_p$ are the reference axes. Stage (i) identifies the principal axes $\mathbf{u}_k$ of the common within-group scatter of the points, and the within-group spread θ_k of points along each axis. Stage (ii) first rotates the axes of the p-dimensional configuration to line up with these principal axes $\mathbf{u}_k$, and then stretches or shrinks the configuration along each axis so that the within-group variance is equalized along all axes (thus converting the within-group-scatter hyperellipsoids into hyperspheres). Group mean vectors are identified as points $\bar{\mathbf{w}}_i$ in this new space. Stage (iii) then effects a principal component analysis of the g points $\bar{\mathbf{w}}_i$ where each point is weighted by the square root of the number of individuals in its corresponding group. These two sequential principal component analyses yield a canonical variate analysis of the original data.

Now suppose that the dispersion matrices are heterogenous, but the data (perhaps after suitable transformation) satisfy the common principal component model. For this model we have $\Omega_i = \mathbf{L}\,\mathbf{D}_i\,\mathbf{L}^T$ $(i = 1, \ldots, g)$. If we write $\mathbf{L} = (\mathbf{l}_1, \ldots, \mathbf{l}_p)$ then the $\mathbf{l}_k$ play the role of the $\mathbf{u}_k$ in the above procedure: we by-pass stage (i), and first rotate the axes of the p-dimensional data configuration to line up with the principal axes $\mathbf{l}_k$. However, since the $\mathbf{D}_i$

differ between groups it is no longer possible to carry out a uniform stretching or shrinking of the configuration along each axis to transform all within-group scatter hyperellipsoids to hyperspheres. Instead, if d_{ij} is the j-th diagonal element of $\mathbf{D}_i$ ($i = 1, \ldots, g$; $j = 1, \ldots, p$), it is necessary to stretch or shrink the j-th axis by the factor $d_{ij}^{-\frac{1}{2}}$ *in the i-th group* to produce hyperspheres. Since the configuration is referred to principal axes, which are statistically independent, in place of explicit stretching or shrinking of the configuration in stage (ii) we can view the d_{ij} as weights in the principal component analysis of the $\bar{\mathbf{w}}_i$ in stage (iii). Where previously each point $\bar{\mathbf{w}}_i$ was weighted by the factor $\sqrt{n_i}$, now each element of $\bar{\mathbf{w}}_i$ must additionally be weighted by the inverse of its standard deviation. Thus the three corresponding stages to the ones above for the common principal component model are as follows.

(i) Find the estimates $\hat{\mathbf{L}}$, $\hat{\mathbf{D}}_i$ in the model $\Omega_i = \mathbf{L}\,\mathbf{D}_i\,\mathbf{L}^T$ ($i = 1, \ldots, g$), using either maximum likelihood or least squares as discussed in Section 2. Use these estimates in place of the parameter values below.

(ii) Transform the original data vectors $\mathbf{x}_{ij}$ to $\mathbf{w}_{ij} = \mathbf{D}_i^{-\frac{1}{2}}\,\mathbf{L}^T\,\mathbf{x}_{ij}$ for $i = 1, \ldots, g$ and $j = 1, \ldots, n_i$. Hence $\bar{\mathbf{w}}_i = \mathbf{D}_i^{-\frac{1}{2}}\,\mathbf{L}^T\,\bar{\mathbf{x}}_i$ ($i = 1, \ldots, g$). Let

$$\bar{\mathbf{w}} = \frac{1}{n} \sum_{i=1}^{g} n_i\,\bar{\mathbf{w}}_i$$

and

$$\mathbf{C} = \frac{1}{g-1} \sum_{i=1}^{g} n_i\,(\bar{\mathbf{w}}_i - \bar{\mathbf{w}})(\bar{\mathbf{w}}_i - \bar{\mathbf{w}})^T .$$

(iii) Find eigenvalues λ_k and eigenvectors $\mathbf{c}_k$ of $\mathbf{C}$ as before. Group means on generalized canonical variates are again the principal component scores of the $\bar{\mathbf{w}}_i$ in this second-stage principal component analysis. From the multiple-transformation problem discussed above, however, it is evident that the co-ordinates in this final space do not have a simple relationship with the original variables across the whole set of data.

Note that this procedure reduces to the earlier version when all dispersion matrices are equal. In this case $\mathbf{D}_i = \Theta$ for all i, $\mathbf{L} = \mathbf{U}$ and $\mathbf{C}$ reduces to $\Theta^{-\frac{1}{2}}\,\mathbf{U}^T\,\mathbf{B}\,\mathbf{U}\,\Theta^{-\frac{1}{2}}$ as before.

4. Distance Between Two Groups

It was pointed out in Section 1 that the squared Euclidean distance between two group mean points in the standard canonical variate space is equal to the (sample) Mahalanobis D^2 between the corresponding groups in terms of the original variates. Consequently a metric scaling of the matrix containing Mahalanobis D^2 values between all pairs of groups will recover the canonical variate configuration of group means (Gower 1966). In this matrix, all D^2 values use the pooled within-group covariance matrix W in their calculation.

Assuming the observed groups to be random samples from given parent populations, the calculated Mahalanobis D^2 values can be viewed as estimates of the squared distances Δ^2 between corresponding populations. For these estimates to be valid ones, the assumptions of normality and common dispersion matrices in all populations (i.e., hypothesis H_0 above) must again be made. These assumptions were made in the original definition of the distance function (Mahalanobis 1931). Subsequently, there have been many general definitions of distance between two populations (e.g., Bhattacharyya 1943; Jeffreys 1948; Kullback and Leibler 1951; Matusita 1956; Ali and Silvey 1966; Rao 1982) and all these definitions yield a monotonic function of $\Delta^2 = (\mu_1 - \mu_2)^T \Omega^{-1} (\mu_1 - \mu_2)$ in the case of two multivariate normal populations with different means μ_1 and μ_2 but the same dispersion matrix Ω. Relaxing the assumption of normality is possible, and the Mahalanobis Δ^2 remains appropriate if the distributions belong to the same family within the class of elliptic distributions (Mitchell and Krzanowski 1985). However, once hypothesis H_0 is violated then the measure Δ^2 is no longer valid. A number of authors have proposed data-based definitions of distance between two groups when their dispersion matrices are heterogeneous (Anderson and Bahadur 1962; Reyment 1962; Chaddha and Marcus 1968; Chernoff 1973). Others have applied one of the formal definitions to derive the theoretical distance between two multivariate normal populations which have different means μ_1, μ_2 and dispersion matrices Ω_1, Ω_2, replacing unknown parameters in the resulting expression by their estimates from the observed data (Kullback 1959; Matusita 1967). If there are similarities among the Ω_i for a set of populations, however, then the same objections can be raised to the application of these distance measures as were raised to existing generalizations of canonical variate analysis: dispersion matrices are only considered in pairwise fashion, so that any similarities linking all g matrices are not exploited, and relatively few degrees of freedom are used for estimation of each distance. In particular, if the common principal component hypothesis is appropriate, then considerable structure among the dispersion matrices is ignored. We therefore consider the calculation of distance between groups

under this model.

First, note that if a generalized canonical variate analysis has been done as indicated in Section 3 above, then the Euclidean distance between a pair of group means in the final derived s-dimensional space is a possible measure of distance between the corresponding groups. However, in general we seek a definition of distance that can be applied to the original data and for this purpose we will use Matusita's (1956) definition. Although in principle we could have used any of the previously suggested definitions, our choice was motivated by the successful application that has been made of this particular definition in a range of statistical problems as well as by the attractive metric properties established under its alternative name of Hellinger distance (Le Cam 1970). A referee has pointed out some conceptual difficulties in defining distance for $g > 2$ groups. However, for the purpose of principal co-ordinate analysis we merely need *pairwise* distances between groups, so in the following we only have to establish a measure of distance between any two of the g groups.

Suppose that $\pi_1, \pi_2, \ldots, \pi_g$ are populations, and $\mathbf{X}$ is a vector-valued random variable defined over a p-dimensional space R such that p_i is the density function of $\mathbf{X}$ in π_i with respect to a suitable measure μ ($i = 1, \ldots, g$). Then the distance between π_i and π_j is defined by

$$\Delta_{ij} = \left\{ \int (\sqrt{p_i} - \sqrt{p_j})^2 d\mu \right\}^{\frac{1}{2}}$$

$$= \sqrt{2(1 - \rho_{ij})}$$

where

$$\rho_{ij} = \int \sqrt{p_i p_j}\, d\mu$$

is the affinity between π_i and π_j.

If π_i denotes the multivariate normal distribution with mean vector μ_i and dispersion matrix Ω_i ($i = 1, \ldots, g$), then Matusita (1967) shows that

$$\rho_{ij} = |\, \Omega_i^{-1} \Omega_j^{-1} \,|^{\frac{1}{4}} \,|\, \frac{1}{2} (\Omega_i^{-1} + \Omega_j^{-1}) \,|^{-\frac{1}{2}}$$

$$\times \exp - \frac{1}{4} \{ \mu_i^T \Omega_i^{-1} \mu_i + \mu_j^T \Omega_j^{-1} \mu_j$$

$$- (\Omega_i^{-1} \mu_i + \Omega_j^{-1} \mu_j)^T (\Omega_i^{-1} + \Omega_j^{-1})^{-1} (\Omega_i^{-1} \mu_i + \Omega_j^{-1} \mu_j) \} \qquad (1)$$

If $\Omega_i = \Omega$ for all i, then this expression simplifies to $\rho_{ij} = \exp(-\frac{1}{8} \Delta^2)$

where $\Delta^2 = (\mu_i - \mu_j)^T \Omega^{-1} (\mu_i - \mu_j)$.

Under the common principal component model we have $\Omega_i = \mathbf{L}\,\mathbf{D}_i\,\mathbf{L}^T$ $(i = 1, \ldots, g)$, where $\mathbf{D}_i = \text{diag}\,(d_{i1}, \ldots, d_{ip})$. It readily follows that:

$$\Omega_i^{-1} = \mathbf{L}\,\mathbf{D}_i^{-1}\,\mathbf{L}^T$$

$$\Omega_i^{-1}\,\Omega_j^{-1} = \mathbf{L}\,\mathbf{D}_i^{-1}\,\mathbf{D}_j^{-1}\,\mathbf{L}^T$$

$$|\,\Omega_i^{-1}\,\Omega_j^{-1}\,| = \prod_{t=1}^{p} \frac{1}{d_{it}\,d_{jt}}$$

$$\Omega_i^{-1} + \Omega_j^{-1} = \mathbf{L}\,(\mathbf{D}_i^{-1} + \mathbf{D}_j^{-1})\,\mathbf{L}^T$$

$$|\,\Omega_i^{-1} + \Omega_j^{-1}\,| = \prod_{t=1}^{p} \left[\frac{d_{it} + d_{jt}}{d_{it}\,d_{jt}}\right]$$

and

$$(\Omega_i^{-1} + \Omega_j^{-1})^{-1} = \mathbf{L}\,\Phi_{ij}\,\mathbf{L}^T$$

where

$$\Phi_{ij} = \text{diag}\left[\frac{d_{i1}\,d_{j1}}{d_{i1} + d_{j1}}, \ldots, \frac{d_{ip}\,d_{jp}}{d_{ip} + d_{jp}}\right].$$

Using these results in (1) and simplifying leads to:

$$\rho_{ij} = 2^{p/2} \left\{\prod_{t=1}^{p} (d_{it}\,d_{jt})^{1/4}\right\} \left\{\prod_{t=1}^{p} (d_{it} + d_{jt})^{-1/2}\right\}$$

$$\times \exp -\frac{1}{4} \Big\{\mu_i^T\,\mathbf{L}\,[\mathbf{D}_i^{-1} - \mathbf{D}_i^{-1}\,\Phi_{ij}^{-1}]\,L^T\,\mu_i$$

$$+ \mu_j^T\,\mathbf{L}[\mathbf{D}_j^{-1} - \mathbf{D}_j^{-1}\,\Phi_{ij}\,\mathbf{D}_j^{-1}]\,\mathbf{L}^T\,\mu_j$$

$$- 2\mu_j^T\,\mathbf{L}\,\mathbf{D}_i^{-1}\,\Phi_{ij}\,\mathbf{D}_j^{-1}\,\mathbf{L}^T\,\mu_j\Big\}\,. \qquad (2)$$

Further algebra establishes that

$$\mathbf{D}_i^{-1} - \mathbf{D}_i^{-1}\,\Phi_{ij}\,\mathbf{D}_i^{-1} = (\mathbf{D}_i + \mathbf{D}_j)^{-1}\,,$$

$$\mathbf{D}_j^{-1} - \mathbf{D}_j^{-1}\,\Phi_{ij}\,\mathbf{D}_j^{-1} = (\mathbf{D}_i + \mathbf{D}_j)^{-1}\,,$$

and

$$\mathbf{D}_i^{-1}\,\Phi_{ij}\,\mathbf{D}_j^{-1} = (\mathbf{D}_i + \mathbf{D}_j)^{-1} \ .$$

Hence if we write $\nu_i = \mathbf{L}^T \mu_i$ then (2) reduces finally to

$$\rho_{ij} = 2^{p/2}\ \frac{|\,\mathbf{D}_i\,\mathbf{D}_j\,|^{1/4}}{|\,\mathbf{D}_i + \mathbf{D}_j\,|^{1/2}}$$
$$\times \exp\left\{-\tfrac{1}{4}\,(\nu_i - \nu_j)^T\,(\mathbf{D}_i + \mathbf{D}_j)^{-1}\,(\nu_i - \nu_j)\right\} \ . \qquad (3)$$

which is easily computed given $\mathbf{L}$, $\mathbf{D}_i$ and μ_i for $i = 1, \ldots, g$.

In a practical application we would use either maximum likelihood or least squares, as in Section 2, to estimate $\mathbf{L}$ and $\mathbf{D}$ $(i = 1, \ldots, g)$ from the observed data, use the estimates in (3) to obtain affinities between all pairs of populations and finally convert these affinities to distances $\Delta_{ij} = \sqrt{2(1-\rho_{ij})}$. The matrix of inter-group distances could then be used in a metric scaling construction of a configuration which highlights the relationships among the g populations. This approach is illustrated in the next section and is compared with both the generalized canonical variate analysis of Section 3 and the standard canonical variate analysis assuming homogeneous dispersion matrices.

5. Example

We return to the Venezuelan students example introduced in Section 2, and now look at the question of differences between the colleges in June. Standard canonical variate analysis and each of the two methods outlined above were applied in turn to the data. Co-ordinates of group means on each of the (generalized) canonical axes are shown for each technique in Table 1, and the canonical roots corresponding to each axis are also given.

Note first that there are six variables and ten colleges in the data set. Thus for both standard canonical variate analysis and the generalization of the two-stage approach the maximum dimensionality is $s = 6$. On the other hand, for metric scaling of the Matusita distance matrix the maximum potential dimensionality is $10 - 1 = 9$ and in this case we actually have a solution in seven dimensions. However, the eigenvalue corresponding to the seventh dimension is so small (0.9% of the total of the eigenvalues) that we incur negligible distortion by ignoring it. Thus to compare the full solutions from the three analyses we can use Procrustes analysis (Gower 1971; Sibson 1978) on three six-dimensional configurations. Each configuration is first normalized so that the sum of squares of all its co-ordinates is unity. The three

Table 1

Co-ordinates of Group Means on (Generalized) Canonical Axes, plus Canonical Roots

Group	Dimension 1	2	3	4	5	6	7
(a) Ordinary Canonical Variates							
1	−1.656	−0.143	0.127	0.359	−0.030	0.012	
2	−0.140	−0.121	−0.131	−0.038	0.034	−0.072	
3	−0.284	0.385	0.259	−0.393	0.261	−0.134	
4	−0.321	0.009	0.266	−0.275	−0.188	−0.038	
5	0.835	−0.892	0.132	−0.117	−0.019	−0.031	
6	0.790	−0.205	−0.033	0.520	0.105	−0.010	
7	0.589	0.520	−0.441	0.114	−0.313	−0.085	
8	0.779	0.564	0.429	0.193	0.098	0.073	
9	−0.016	0.001	−0.003	−0.323	−0.081	0.242	
10	−0.231	0.017	−0.799	−0.131	0.197	0.041	
Canonical Roots	0.639	0.179	0.128	0.086	0.027	0.009	
(b) Two-Stage Generalization							
1	9.716	−0.589	−0.721	−0.472	0.005	0.013	
2	0.731	−0.283	−0.326	0.182	0.379	−0.248	
3	1.180	1.907	1.334	−0.707	0.693	−0.510	
4	1.390	0.345	1.857	0.917	−0.575	−0.085	
5	−2.975	−3.477	1.269	0.166	0.350	−0.427	
6	−4.037	−1.041	−1.620	−1.010	−0.222	0.232	
7	−2.333	1.820	−1.327	1.661	−0.146	−0.271	
8	−3.326	2.088	0.655	−1.176	0.314	0.080	
9	0.255	−0.112	0.881	1.018	1.012	0.712	
10	0.510	−0.135	−1.491	0.506	1.123	−0.402	
Canonical Roots	140.10	25.03	15.25	7.96	2.67	1.23	
(c) Metric Scaling							
1	0.526	0.002	−0.026	0.036	−0.049	−0.006	−0.016
2	0.104	−0.100	−0.115	−0.039	0.016	−0.025	−0.008
3	0.072	0.117	0.034	−0.051	0.019	0.051	0.012
4	0.098	−0.093	0.133	−0.083	0.082	−0.021	−0.007
5	−0.136	−0.114	−0.094	0.099	0.065	0.043	0.008
6	−0.273	−0.091	0.058	0.050	−0.052	−0.051	−0.029
7	−0.126	0.060	−0.068	−0.082	−0.022	−0.058	0.045
8	−0.200	0.019	−0.021	−0.095	−0.052	0.067	−0.029
9	0.010	−0.047	0.097	0.081	−0.047	0.027	0.044
10	−0.074	0.246	0.003	0.084	0.040	−0.027	−0.019
Canonical Roots	0.457	0.121	0.059	0.054	0.024	0.018	0.007

configurations are then compared in pairwise fashion, by rotating the configurations of each pair to positions of best fit (in the least squares sense) with respect to each other. The residual sums of squares among the ten pairs of points in the resulting rotated configurations were as follows: 0.0781 for the standard analysis versus the two-stage generalization; 0.2362 for the standard analysis versus metric scaling; and 0.2109 for two-stage versus metric scaling. These three rotations show that the standard analysis and the two-stage analysis are much more similar to each other than each is to metric scaling.

Next we consider the results of each analysis in terms of the configurations of points representing colleges. The first two canonical variates accounted for 76.5% of the between-college relative to within-college differences in the standard analysis, and the corresponding percentages were 85.8% for the two-stage analysis and 80.1% for metric scaling. Consequently a plot of the group means on the first two (generalized) canonical variates as axes will give an adequate approximation to the full configuration in each analysis; Figure 1 gives the first two dimensions for the standard analysis, Figure 2 that for the two-stage version and Figure 3 that for the metric scaling. Where necessary, directions of axes have been reversed so that the resulting configurations have the same orientations.

It is evident that the configurations from the standard analysis and from the two-stage generalization are very similar, the only minor differences being a greater contraction of the central group (colleges 2, 4, 9, 10) and a slight displacement of college 6 in the latter case. By contrast, the configuration in the metric scaling shows an overall contraction, and additionally college 10 shows considerable displacement from its position in the other two configurations.

A reason was sought for the relative similarity between the configuration of the standard analysis and that of the two-stage generalization. Table 2 shows the common principal components (by columns), and the variances on these components for each of the ten colleges. The patterns of these variances appear to be relatively similar for all colleges except colleges 2, 10 and (perhaps) 8. Omitting these three colleges from the data and retesting the hypothesis of equal dispersion matrices in the remaining seven colleges yielded a χ^2 statistic of 158.8 on 126 degrees of freedom. This represents a rise in significance level from 0.1% to just under 5%, and while there is still some evidence of heterogeneity the practitioner would now be much happier with a standard analysis. Thus the relative similarity between Figures 1 and 2 could be attributed to relatively small differences between the common principal component and the equal dispersion models. Furthermore, the patterns of variances in Table 2 suggest a possible extension to the common component model, in which some of the groups are constrained to have

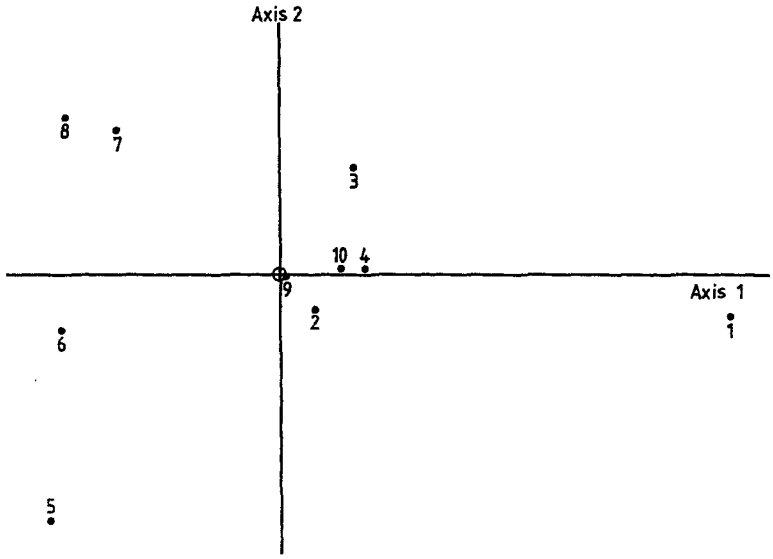

Figure 1. Plot of college means on first two canonical variate axes, standard analysis.

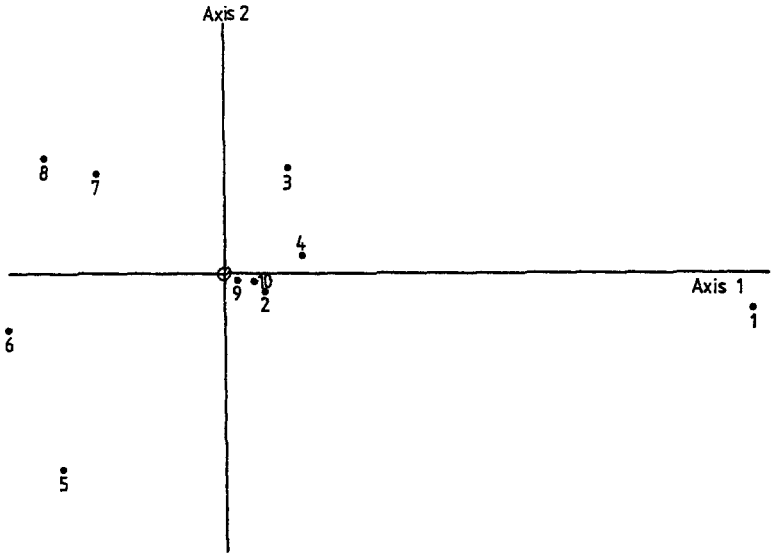

Figure 2. Plot of college means on first two generalized canonical variate axes, two-stage generalization.

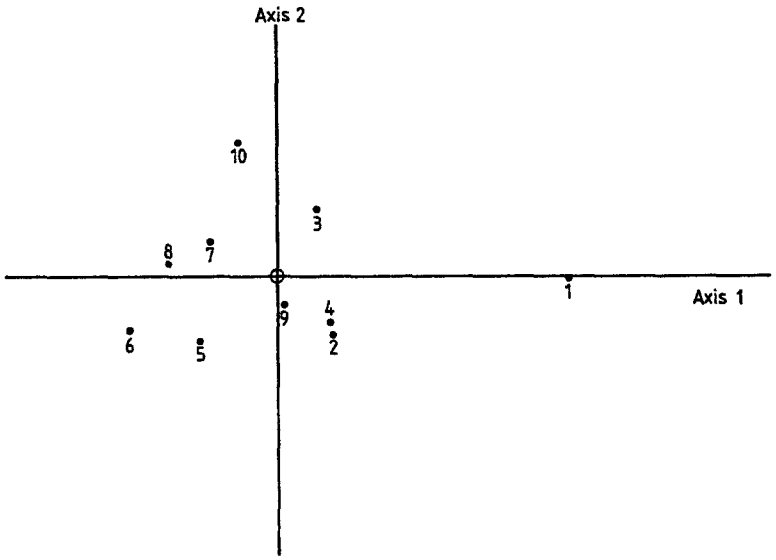

Figure 3. Plot of college means on first two generalized canonical variate axes, metric scaling.

Table 2

Common Principal Components and Variances on these Components for Each College, for June Set of Venezuelan Students' Data

(a) Common Principal Components (by columns)

Test	Component 1	2	3	4	5	6
Comprehension	0.285	0.413	–0.776	–0.080	–0.374	–0.005
Essay	0.432	–0.835	–0.307	–0.109	0.068	0.071
Cloze (exact)	0.439	0.165	0.276	–0.617	0.082	–0.563
Cloze (acceptable)	0.298	0.190	0.234	–0.379	0.022	0.822
Structure	0.416	0.263	–0.115	0.420	0.754	–0.005
Dictation	0.527	0.010	0.399	0.530	–0.529	–0.048

(b) Variances on These Components for Each College

College						
1	16.15	7.28	4.22	3.78	0.99	1.57
2	23.05	12.63	3.08	13.11	2.70	0.92
3	31.11	7.56	8.38	3.83	4.62	0.45
4	33.31	5.32	1.29	3.95	3.39	0.38
5	46.07	6.51	5.08	6.99	4.34	0.80
6	19.16	1.66	4.18	5.07	4.27	0.51
7	30.38	6.72	5.38	4.26	5.85	1.04
8	36.41	3.29	4.57	6.34	9.51	1.10
9	58.59	2.53	6.65	6.00	2.06	0.49
10	40.33	5.71	15.91	4.10	14.83	0.35

equal covariance matrices while the remainder merely have common principal axes. Such a possibility remains to be investigated.

As a final comment, it seems that the differing variance patterns along the common principal components for colleges 2 and 10 versus the rest have been emphasized more in the metric scaling of Figure 3 than in the two-stage generalization of Figure 2, but no further explanation for the discrepancies between these two configurations has yet been found.

6. Conclusion

Two possible methods of between-group analysis have been proposed, for the case where the data follow the common principal component model. Both methods have been applied to some real data and shown to be viable. Also, both methods reduce to standard canonical variate analysis when covariance matrices are homogeneous. However, differences in results between the two methods do exist, and further investigation is necessary before any definite recommendations can be made about their relative merits.

References

ALI, S. M., and SILVEY, S. D. (1966), "A General Class of Coefficients of Divergence of One Distribution from Another," *Journal of the Royal Statistical Society, Series B, 28*, 131-142.

ANDERSON, T. W., and BAHADUR, R. R. (1962), "Classification into Two Multivariate Normal Distributions with Different Covariance Matrices," *Annals of Mathematical Statistics, 33*, 420-431.

BHATTACHARYYA, A. (1943), "On a Measure of Divergence between Two Statistical Populations Defined by Their Probability Distributions," *Bulletin of the Calcutta Mathematical Society, 35*, 99-109.

CAMPBELL, N. A. (1984), "Canonical Variate Analysis with Unequal Covariance Matrices: Generalizations of the Usual Solution," *Mathematical Geology, 16*, 109-124.

CAMPBELL, N. A., and ATCHLEY (1981), "The Geometry of Canonical Variate Analysis," *Systematic Zoology, 30*, 268-280.

CHADDHA, R. L., and MARCUS, L. F. (1968), "An Empirical Comparison of Distance Statistics for Populations with Unequal Covariance Matrices," *Biometrics, 24*, 683-694.

CHERNOFF, H. (1973), "Some Measures for Discriminating Between Normal Multivariate Distributions with Unequal Covariance Matrices," in *Multivariate Analysis III*, Ed. P. R. Krishnaiah, New York: Academic Press, pp. 337-344.

CLARKSON, D. B. (1988a), "A Remark on Algorithm AS211: The F-G Diagonalization Algorithm. Algorithm ASR71," *Applied Statistics, 37*, 147-151.

CLARKSON, D. B. (1988b), "A Least Squares Version of Algorithm AS211: The F-G Diagonalization Algorithm. Algorithm ASR74," *Applied Statistics, 37*, 317-321.

FLURY, B. N. (1988), *Common Principal Components and Related Models*. New York: Wiley.

FLURY, B. N., and CONSTANTINE, G. (1985), "The F-G Diagonalization Algorithm. Algorithm AS211," *Applied Statistics, 34*, 177-183.

GOWER, J. C. (1966), "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis," *Biometrika, 53*, 325-338.

GOWER, J. C. (1971), "Statistical Methods of Comparing Different Multivariate Analyses of the Same Data," in *Mathematics in the Archaeological and Historical Sciences*, Eds. F. R. Hodson, D. G. Kendall and P. Tautu, Edinburgh: University Press, pp. 138-149.

JEFFREYS, H. (1948), *Theory of Probability* (2nd ed.), Oxford: Clarendon Press.

KRZANOWSKI, W. J. (1988), *Principles of Multivariate Analysis: A User's Perspective*, Oxford: Clarendon Press.

KULLBACK, S. (1959), *Information Theory and Statistics*, New York: Wiley.

KULLBACK, S., and LEIBLER, R. (1951), "On Information and Sufficiency," *Annals of Mathematical Statistics, 22*, 79-86.

LE^CAM, L. (1970), "On the Assumptions Used to Prove Asymptotic Normality of Maximum Likelihood Estimates," *Annals of Mathematical Statistics, 41*, 802-828.

MAHALANOBIS, P. C. (1936), "On the Generalized Distance in Statistics," *Proceedings of the National Institute of Science, India, 22*, 49-55.

MANLY, B. F. J., and RAYNER, J. C. W. (1987), "The Comparison of Sample Covariance Matrices Using Likelihood Ratio Tests," *Biometrika, 74*, 841-847.

MARDIA, K. V., KENT, J. T., and BIBBY, J. M. (1979), *Multivariate Analysis*, London: Academic Press.

MATUSITA, K. (1956), "Decision Rule, Based on the Distance, for the Classification Problem," *Annals of the Institute of Statistical Mathematics, 8*, 67-77.

MATUSITA, K. (1967), "Classification Based on Distance in Multivariate Gaussian Cases," *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1*, 299-304.

MITCHELL, A. F. S., and KRZANOWSKI, W. J. (1985), "The Mahalanobis Distance and Elliptic Distributions," *Biometrika, 72*, 464-467.

PEARSON, E. S., and HARTLEY, H. O. (1966), *Biometrika Tables for Statisticians*, Vol. 1, Cambridge: University Press.

RAO, C. R. (1982), "Diversity and Dissimilarity Coefficients: A Unified Approach," *Theoretical Population Biology, 21*, 24-43.

REYMENT, R. A. (1962), "Observations on Homogeneity of Covariance Matrices in Palaeontologic Biometry," *Biometrics, 18*, 1-11.

SIBSON, R. (1978), "Studies in the Robustness of Multidimensional Scaling: Procrustes Statistics," *Journal of the Royal Statistical Society, Series B, 40*, 234-238.

YOUNG, D. M., MARCO, V. R., and ODELL, P. C. (1987), "Quadratic Discrimination: Some Results on Optimal Low-dimensional Representation," *Journal of Statistical Planning and Inference, 17*, 307-319.