Biometrika (2009), **xx**, x, *pp.* 1–12

Printed in Great Britain

Between-Group Metrics

By John C. Gower, Casper J. Albers

Department of Mathematics & Statistics, The Open University, Walton Hall, Milton Keynes MK7 6AA, United Kingdom.

j.c.gower@open.ac.uk, c.j.albers@open.ac.uk

Summary

In canonical analysis with more variables than samples, it is shown that, as well as the usual canonical means in the range-space of the within-groups dispersion matrix, canonical means may be defined in its null space. In the range space we have the usual Mahalanobis metric; in the null space explicit expressions are given and interpreted for a new metric.

Keywords: Between-group distances, Canonical analysis, Mahalanobis distance

1. Introduction

In Canonical Variate Analysis measurements on each of p variables for n samples are distributed among k groups of sizes $n_1 + n_2 + \ldots + n_k = n$. These measurements are available in an $n \times p$ matrix X, assumed column-centered, and therefore of rank at most $\min(n-1, p)$, with group-membership given in an $n \times k$ indicator matrix G. Here, $g_{ij} = 1$ when the ith sample belongs to the jth group but otherwise G is zero. Thus $G1 = 1$ and $1'G = 1'N$, where $N = \mathrm{diag}(n_1, n_2, \ldots, n_k) = G'G$; we also write ${}_nH_n = {}_nG_k N_k^{-1} G'_n$.

With this notation, the usual between and within-group orthogonal decomposition:

$$ {}_nX_p = {}_nG_kN_k^{-1}G'_nX_p + [I - {}_nG_kN_k^{-1}G'_n]_nX_p = HX + (I-H)X $$

has an associated analysis of variance

$$ {}_pX'_nX_p = {}_pX'_nH_nX_p + {}_pX'_n(I-H)_nX_p $$

expressing that the Total sum-of-squares (T) is the sum of the Between-Group sum-of-squares (B) and the Within-Group sum-of-squares (W). Note that the n rows of HX repeat the k different means $n_1, n_2, \ldots, n_k$ times; to get each mean only once, we require $N^{-1}GX$ which we write as $\bar{X}$.

In classical canonical variate analysis, the spectral decomposition $W = U\Sigma^2U'$ underpins the transformation to canonical variables XL where $L = U\Sigma^{-1}$. These define canonical means HXL with inner-products $(HXL)(HXL)' = HXW^{-1}X'H$ that use the metric $LL' = W^{-1}$ to generate Mahalanobis distances between the canonical means; note that $L'WL = I$. The rank of the canonical means is $k-1$ (or less) but they may be approximated in a smaller space, by using a conventional principal components analysis. These two steps (i) define a metric, followed by (ii) a principal components analysis, are usually subsumed into a single two-sided eigenvalue calculation but the two-step process is better for understanding the following.

The above requires that W has full rank p. The case when $p > n$ is increasingly important where much of the interest is in overcoming computational difficulties, perhaps reducing the number of variables by identifying and rejecting those deemed irrelevant or by focussing on some form of functional multivariate analysis (see e.g. Krzanowski, 1995; Mertens, 1998). Here, we explore a novel structural property of canonical analysis that occurs when $p > n$. When $p > n$ then $\text{rank}(W) = n - k$ and $\text{rank}(T) = n - 1$ and

W does not have an ordinary inverse so the Mahalanobis metric is undefined. This need not be a major problem, because we may express the spectral decomposition

$$W = (U_1, U_0) \begin{pmatrix} \Sigma^2 & \\ & 0 \end{pmatrix} \begin{pmatrix} U_1' \\ U_0' \end{pmatrix}$$

where U_1 are the eigenvectors in the range space of W and U_0 those in its null space. Then, we may define canonical means HXL where now $L = U_1\Sigma^{-1}$ in the range space. No longer is $L'WL = I$ but rather $L'WL = I_{n-k}$. Then $(L'WL)(L'WL) = I_{n-k} = L'WL$. We may write this:

$$\begin{pmatrix} L' \\ U_0' \end{pmatrix} WLL'W\,(L, U_0) = \begin{pmatrix} L' \\ U_0' \end{pmatrix} W\,(L, U_0)$$

which, because (L, U_0) is non-singular, gives $W(LL')W = W$ showing that the metric is now a generalised inverse, rather than an inverse, of W. With this minor change, we may proceed as before with a principal components analysis. An interesting thing is that canonical means may also be defined in the null space. This follows from noting that the null vectors satisfy:

$$X'(I - H)XU_0 = 0$$

and so

$$XU_0 = HXU_0. \tag{1}$$

Note that the k different means are repeated $n_1, n_2, \ldots, n_k$ times in the n rows of both XU_0 and, equivalently, HXU_0. Being null vectors of W, the canonical variables XU_0 have zero variability within groups, but the corresponding canonical means HXU_0 have non-zero sums-of-squares. Evidently, the computation of HXU_0 is straightforward, as is any subsequent principal components analysis; an example is given by Gower & Albers

(‘Canonical Analysis: Ranks, Ratios and Fits’, in preparation). For a fuller understanding it is interesting to ask what functional form, analogous to Mahalanobis distance in the range space, is taken by the distance d_{ij} between the ith and jth canonical means in the null space of W. This is our main objective below but first we have to address a minor but troublesome technical matter.

The total dispersion $T = X'X$ has rank $n-1$ so implying an extensive null space of rank $p-n+1$; this null space is also common to the null spaces of B and W. This common null space is uninteresting; we are concerned only with the additional null spaces of W and B that are in the range space of T, especially the intersection of the range space of T and the null space of W which normally has dimension/rank $k-1$. To simplify the following development we assume that the common null space has been eliminated by taking the spectral decomposition $T = V\Lambda V'$ and redefining X as XV. Throughout the following, we assume that X has been so redefined.

This initialisation to give X with $n-1$ columns, eliminates the common null space from the dispersion matrices T, B and W. However, it does not remove null items from X itself. Indeed the vector 1, which eliminates the general mean, is one such null vector and is what gives rise to the rather extensive algebraic manipulations required in the following. Linear combinations among the rows of X will generate additional null vectors in the common null space. The position is complicated, because such linear combinations may be of two, not mutually exclusive, kinds (i) linear combinations within groups and (ii) linear combinations among the group means. Loss of rank within groups merely reduces the number of columns of the redefined X but to handle all variants that include (ii) is not trivial and would greatly extend this short paper. Therefore, apart from some passing concluding remarks, throughout the following, we assume that rank$\bar{X} = k - 1$ and that rank$X \leq n - 1$.

2. Derivation of d_{ij}^2

From here on we shall be working in the null space of W so we drop the suffix from U_0. Starting from $XU = HXU$ for the null-vectors of W, as in (1), we have that

$$
\begin{aligned}
{}_nX_{n-1}U_{k-1} &= G(G'G)^{-1}G'XU \\
&= {}_nG_kA_{k-1}
\end{aligned}
$$

where ${}_kA_{k-1} = \bar{X}U$ are the k group-mean coordinates given in repeated form in XU.

Thus, the calculation $A = (G'G)^{-1}G'XU$ is an expression whose rows give coordinates that generate the distance d_{ij} between each pair of group-mean coordinates. We need an explicit expression for d_{ij}^2 . We do not require A itself, which has the usual rotational indeterminacy, but only AA'. Then, $d_{ij}^2 = (AA')_{ii} + (AA')_{jj} - 2(AA')_{ij}$. Because X is centred, $1'X = 0$ and so $1'XU = 1'GA = 1'NA = 0$. Also, $U'U = I_{k-1}$. From (1) $U'U = ((X'X)^{-1}X'GA)'((X'X)^{-1}X'GA) = A'PA$ where $P = G'X(\Lambda^{-2})X'G$ with $X'X = \Lambda$. We have $1'P = 1'G'X(\Lambda^{-2})X'G = 1'X(\Lambda^{-2})X'G = 0$. So we have to solve $A'PA = I$ for AA'. The difficulty is that both A and P are singular (with rank $k-1$). Hence, consider

$$
\begin{pmatrix} A'N \\ \frac{1}{n}1'N \end{pmatrix} (Q + \lambda 11')(NA, \tfrac{1}{n}N1) \tag{2}
$$

where $Q = N^{-1}PN^{-1} = \bar{X}\Lambda^{-2}\bar{X}'$. The introduction of λ may seem arbitrary but we shall show that it has no substantive effect. On expansion, (2) becomes $\begin{pmatrix} I & 0 \\ 0 & \lambda \end{pmatrix}$ giving:

$$Q + \lambda 11' = \begin{pmatrix} A'N \\ \frac{1}{n}1'N \end{pmatrix}^{-1} \begin{pmatrix} I & \\ & \lambda \end{pmatrix} (NA, \tfrac{1}{n}N1)^{-1}$$

and

$$(Q + \lambda 11')^{-1} = (NA, \tfrac{1}{n}N1) \begin{pmatrix} I & \\ & \frac{1}{\lambda} \end{pmatrix} \begin{pmatrix} A'N \\ \frac{1}{n}1'N \end{pmatrix}.$$

Thus,

$$N^{-1}\left(Q + \lambda 11'\right)^{-1} N^{-1} = \frac{AA' + 11'}{\lambda n^2}. \tag{3}$$

From 3 we may calculate d_{ij}^2. The constant term $11'/\lambda n^2$ has no effect on derived distances and we shall show that $(Q + \lambda 11')^{-1}$ also is invariant to non-zero choices of λ. Thus (3) contains everything needed for finding AA' but the evaluation of $(Q + \lambda 11')^{-1}$ needs some care, because Q is singular and the equivalent of (A4) is unavailable.

For simplicity, we derive d_{12}^2, the other values of d_{ij}^2 following by symmetry. Notation is established via

$$C = \left(\begin{array}{c|c|c} c_{11} & c_{21} & c_1' \\ \hline c_{12} & c_{22} & c_2' \\ \hline & & \\ c_1 & c_2 & C_{12} \\ & & \end{array}\right) \quad \text{where} \quad \begin{cases} C = Q + \lambda 11' \\ C_{12} = Q_{12} + \lambda 11' \\ c_1 = q_1 + \lambda 1 \\ c_2 = q_2 + \lambda 1 \\ (c_{11}, c_{12}, c_{22}) = (q_{11}, q_{12}, q_{22}) + \lambda \end{cases}$$

From (3)

$$\Delta d_{12}^2 = \frac{1}{n_1^2}\det\begin{pmatrix} c_{22} & c_2' \\ c_2 & C_{12} \end{pmatrix} + \frac{1}{n_2^2}\det\begin{pmatrix} c_{11} & c_1' \\ c_1 & C_{12} \end{pmatrix} + \frac{2}{n_1 n_2}\det\begin{pmatrix} c_{12} & c_2' \\ c_1 & C_{12} \end{pmatrix}, \tag{4}$$

where $\Delta = \det C$, and the determinants are the cofactors of c_{11}, c_{22} and c_{12}. Using (A1), (4) becomes

$$\Delta d_{12}^2 = \det C_{12}\left[\frac{1}{n_1^2}\left(c_{22} - c_2' C_{12}^{-1} c_2\right)^{-1} + \frac{1}{n_2^2}\left(c_{11} - c_1' C_{12}^{-1} c_1\right)^{-1} + \frac{2}{n_1 n_2}\left(c_{12} - c_1' C_{12}^{-1} c_2\right)^{-1}\right],$$

which simplifies to

$$n_1^2 n_2^2 \Delta d_{12}^2 = \det C_{12}\left[n_1^2 c_{11} + n_2^2 c_{22} + 2 n_1 n_2 c_{12} - (n_1 c_1 + n_2 c_2)' C_{12}^{-1} (n_1 c_1 + n_2 c_2)\right]. \tag{5}$$

Using (A1) and (A2) we have

$$\begin{aligned}(n_1 c_1 + n_2 c_2)' C_{12}^{-1} (n_1 c_1 + n_2 c_2) &= (C_{12} M 1 - \lambda n 1)' C_{12}^{-1} (C_{12} M 1 - \lambda n 1)\\ &= 1' M C_{12} M 1 - 2\lambda n 1' M 1 + \lambda^2 n^2 1' C_{12}^{-1} 1. \qquad (6)\end{aligned}$$

Bringing everything together using (6), (A4) and (A5), and expanding in terms of q_{ij}, (5) becomes

$$\begin{aligned}n_1^2 n_2^2 \Delta d_{12}^2 = (1 + \lambda 1' Q_{12} 1) \det Q_{12} \Big[& n_1^2 q_{11} + n_2^2 q_{22} + 2 n_1 n_2 q_{12} + \lambda (n_1 + n_2)^2 \\ & - 1' M Q_{12} M 1 - \lambda (1' M 1)^2 + 2 \lambda n 1' M 1 - \lambda^2 n^2 \left(Q_{12}^{-1} - \frac{\lambda Q_{12}^{-1} 1 1' Q_{12}^{-1}}{1 + \lambda 1' Q_{12}^{-1} 1}\right)\Big]\end{aligned}$$

which, on using (A1), simplifies to

$$\begin{aligned}n_1^2 n_2^2 \Delta d_{12}^2 &= \det Q_{12}\left[\lambda (n_1 + n_2)^2 + \lambda (n - n_1 - n_2)(n + n_1 + n_2) - \frac{\lambda n^2 1' Q_{12}^{-1} 1}{1 + \lambda 1' Q_{12}^{-1} 1}\right](1 + \lambda 1' Q_{12}^{-1} 1)\\ &= \det Q_{12}\left[\lambda n^2 \left(1 + \lambda 1' Q_{12}^{-1} 1\right) - \lambda^2 n^2 1' Q_{12}^{-1} 1\right]\\ &= \lambda n^2 \det Q_{12}. \qquad (7)\end{aligned}$$

This simple result shows that d_{ij}^2 is proportional to $\det Q_{ij}$.

To show that d_{ij}^2 is independent of λ requires an analysis of $\Delta = \det(Q + \lambda 1 1')$. Let $R = Q + 1 1'$. Then $I = R^{-1} Q + R^{-1} 1 1'$ and so

$$1' I N 1 = 1' \left(R^{-1} Q + R^{-1} 1 1'\right) N 1$$

$$1'N1 = \left(1'R^{-1}1\right)(1'N1)$$

$$1'R^{-1}1 = 1.$$

Then

$$\begin{aligned}\Delta &= \det(R + (\lambda - 1)11') \\ &= \det R\left(1 + (\lambda - 1)1'R^{-1}1\right) \\ &= \lambda \det R \\ &= \lambda \det(Q + 11').\end{aligned}$$

That $R \neq 0$ is guaranteed by our assumption that $\bar{X} = k - 1$ made at the end of Section 1.

Thus, finally, (7) becomes

$$d_{12}^2 = \frac{1}{\det(Q + 11')} \frac{n^2}{n_1^2 n_2^2} \det Q_{12} \qquad (8)$$

showing that d_{ij}^2 depends only on the group sizes and $\det Q_{12}$. Recall that $Q = \bar{X}\Lambda^{-1}\bar{X}'$ and that Q_{12} is obtained from Q by striking out its first two rows and columns.

3. Alternative expressions

The expression for Δ has manu forms. Of special interest is that derived from writing

$$\Delta = \det \left(\begin{array}{cc|c} c_{11} & c_{12} & c_1' \\ c_{12} & c_{22} & c_2' \\ \hline c_1 & c_2 & C_{12} \end{array}\right).$$

Multiplying the first row by n_1 and then adding q_i $(i = 2, \ldots, k)$ times the other rows, replaces the first row by $n\lambda 1$. This shows that λ may be subtracted from rows $2, \ldots, k$

to give

$$n_1\Delta = \lambda n \det \left(\begin{array}{cc|c} 1 & 1 & 1' \\ q_{12} & q_{22} & q_2' \\ \hline q_1 & q_2 & Q_{12} \end{array} \right)$$

whence similar operations on the columns give

$$\begin{aligned} n_1^2\Delta &= \lambda n^2 \det \left(\begin{array}{c|cc} 1 & 1 & 1' \\ 0 & q_{22} & q_2' \\ \hline 0 & q_2 & Q_{12} \end{array} \right) \\ &= \lambda n^2 \left(q_{22} - q_2' Q_{12}^{-1} q_2 \right) \det Q_{12}. \end{aligned} \tag{9}$$

Similar expressions may be derived by annihilating the second row/column and the first row and second column to give

$$\left. \begin{aligned} n_2^2\Delta &= \lambda n^2 \left(q_{11} - q_1' Q_{12}^{-1} q_1 \right) \det Q_{12} \\ n_1^2\Delta &= \lambda n^2 \left(q_{22} - q_2' Q_{12}^{-1} q_2 \right) \det Q_{12} \\ -n_1 n_2\Delta &= \lambda n^2 \left(q_{12} - q_1' Q_{12}^{-1} q_2 \right) \det Q_{12} \end{aligned} \right\} . \tag{10}$$

Combining, gives the symmetric form

$$(n_1 + n_2)^2\Delta = \lambda n^2 \left[(q_{11} + q_{22} - 2q_{12}) - (q_1 - q_2)' Q_{12}^{-1} (q_1 - q_2) \right] \det Q_{12}$$

which, on substitution into (8) gives

$$d_{12}^2 = \frac{(n_1 + n_2)^2}{n_1^2 n_2^2} \left[(\bar{x}_1 - \bar{x}_2)' \Lambda^{-2} (\bar{x}_1 - \bar{x}_2) - (q_1 - q_2)' Q_{12}^{-1} (q_1 - q_2) \right]^{-1} . \tag{11}$$

Other substitutions for Δ given by (10) give alternative, less symmetric, expressions for d_{ij}^2.

4. Interpretation

From (11), for $k = 2$ we immediately have

$$d_{12}^{-2} = \frac{n_1^2 n_2^2}{n^2} (\bar{x}_1 - \bar{x}_2)' \Lambda^{-2} (\bar{x}_1 - \bar{x}_2) . \tag{12}$$

We next examine the part of expression (11) that is enclosed in square brackets. We have

$$\bar{X} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \hline \bar{X}_{12} \end{pmatrix} \quad \text{and} \quad \begin{matrix} (q_1 - q_2) = (\bar{x}_1 - \bar{x}_2)\Lambda^{-2}\bar{X}_{12}\}' \\ Q_{12} = \bar{X}_{12}\Lambda^{-2}\bar{X}_{12}' \end{matrix} .$$

Hence,

$$d_{12}^{-2} = \frac{n_1^2 n_2^2}{(n_1+n_2)^2} \left[(\bar{x}_1 - \bar{x}_2)'\Lambda^{-2}(\bar{x}_1 - \bar{x}_2) - (\bar{x}_1 - \bar{x}_2)'\Lambda^{-1}R\Lambda^{-1}(\bar{x}_1 - \bar{x}_2) \right]$$

where $R = \Lambda^{-1}\bar{X}_{12}(\bar{X}_{12}\Lambda^{-2}\bar{X}_{12})^{-1}\bar{X}_{12}\Lambda^{-1}$ represents projection in the metric Λ onto the space spanned by the $k-2$ rows of $\bar{X}_{12}\Lambda$. Thus

$$d_{12}^{-2} = \frac{n_1^2 n_2^2}{(n_1+n_2)^2} \left[(\bar{x}_1 - \bar{x}_2)'\Lambda^{-1}(I - R)\Lambda^{-1}(\bar{x}_1 - \bar{x}_2) \right] . \tag{13}$$

Thus, expressions (6), (11) and (13) are our main results.

The interpretation of (13), visualised in Figure 1, is that d_{12}^{-1} is a measure of how far the space of $\bar{x}_1, \bar{x}_2$ is from the space spanned by $\bar{x}_3, \ldots, \bar{x}_k$. These results are expressed in terms of the Λ-metric but this is a function of the initial rescaling and vanishes on transforming back to the scales of the original variables.

If $\bar{x}_1$ and $\bar{x}_2$ lie in R then (13) gives a null projection onto $I - R$ but this case is excluded by our assumption that $\text{rank}\bar{X} = k - 1$. An indication of the corresponding results when $\text{rank}\bar{X} < k - 1$ is given by the collinearity case $k = 3$ and $\text{rank}\bar{X} = 1$. Then, without giving the detailed derivation, it can be shown that d_{12}^2 is proportional to $(\bar{x}_1 - \bar{x}_2)'\Lambda^{-2}(\bar{x}_1 - \bar{x}_2)$. That the projection term vanishes is consistent and plausible but we do not know whether it generalises to other reduced rank cases.

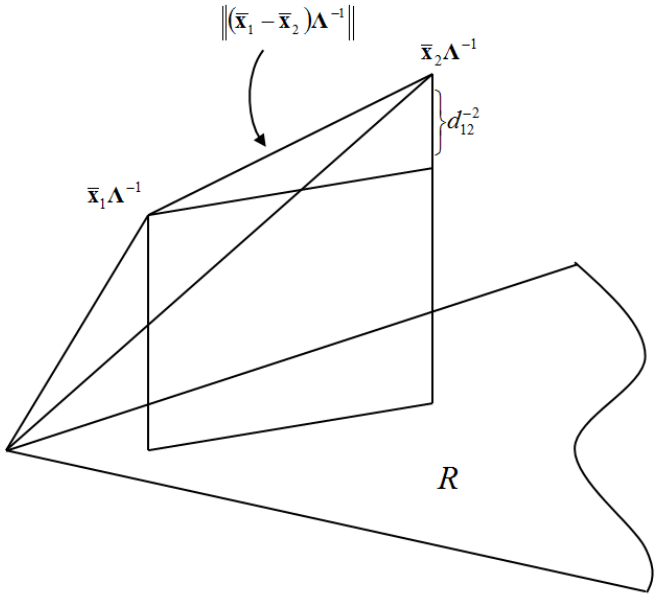

Fig. 1. R is the space spanned by the means $\bar{x}_3, \ldots, \bar{x}_k$. The inverse of the distance between groups 1 and 2 in the intersection space is given by the illustrated projection onto the space orthogonal to R.

5. Basic formulae

This appendix contains some basic results for convenience of reference. Most are well-known and, apart from (A1) and (A2), no derivations are supplied.

Using $1'NQ = 0$ shows that

$$n_1 q_{11} + n_2 q_{12} + 1'Mq_1 = 0$$

$$n_1 q_{12} + n_2 q_{22} + 1'Mq_2 = 0$$

$$n_1 q_1 + n_2 q_2 + Q_{12}M1 = 0$$

where $N = \text{diag}(n_1, n_2, M)$. It follows that

$$n_1^2 q_{11} + n_2^2 q_{22} + 2n_1 n_2 q_{12} = -1'M(n_1 q_1 + n_2 q_2) = 1'MQ_{12}M1. \tag{A1}$$

Similarly,

$$n_1c_1 + n_2c_2 = n_1q_1 + n_2q_2 + \lambda(n_1 + n_2)1$$
$$= -Q_{12}M1 + \lambda(n_1 + n_2)1$$
$$= -C_{12}M1 + \lambda n1. \quad \text{(A2)}$$

Furthermore, we need

$$\det \begin{pmatrix} \alpha & a' \\ a & A \end{pmatrix} = \left(\alpha - a'A^{-1}a\right) \det A, \quad \text{(A3)}$$

$$\det C_{12} = \det(Q_{12} + \lambda 11') = \det Q_{12} \left(1 + \lambda 1' Q_{12}^{-1} 1\right), \quad \text{(A4)}$$

and

$$C_{12}^{-1} = (Q_{12} + \lambda 11')^{-1} = Q_{12}^{-1} - \frac{\lambda Q_{12}^{-1} 11' Q_{12}^{-1}}{1 + \lambda 1' Q_{12}^{-1} 1}. \quad \text{(A5)}$$

References

Krzanowski, W.J., Jonathon, P., McCarthy, W.V. and Thomas, M.R. (1995), Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data. Applied Statistics, 44, 101-115.

Mertens, B.J.A. (1998), Exact principal components influence measures applied to the analysis of spectroscopic data on rice, Journal of the Royal Statistical Society, Series C (Applied Statistics), 47(4), 527 - 542.