2 Canonical Variate Analysis (CVA)

For $k = 1, \ldots, K$ let X_k be an $n_k \times p$ data matrix for a sample of n_k individuals drawn from population k and measured on p variables. Assume that the K populations have a common covariance matrix Σ, but possibly different mean vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$.
Let $\mathbf{x}_{kj}$ be the $p \times 1$ observation vector for the j'th individual in sample k. Then

$$X_k = \begin{bmatrix} \mathbf{x}_{k1}^T \\ \mathbf{x}_{k2}^T \\ \vdots \\ \mathbf{x}_{kn_k}^T \end{bmatrix}.$$

Let $\overline{\mathbf{x}}_k$ be the $p \times 1$ vector of variable means in sample k, so that

$$\overline{\mathbf{x}}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \mathbf{x}_{kj}, \quad k = 1, \ldots, K.$$

Then the overall $p \times 1$ vector of means, $\overline{\mathbf{x}}$, is given by

$$\overline{\mathbf{x}} = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{n_k} \mathbf{x}_{kj} = \frac{1}{n} \sum_{k=1}^{K} n_k \, \overline{\mathbf{x}}_k,$$

where $n = \sum_{k=1}^{K} n_k$.
Define the within-samples sums of squares and products matrix, W, and the weighted between-samples sums of squares and products matrix, B, as follows:

$$W = \sum_{k=1}^{K} \sum_{j=1}^{n_k} (\mathbf{x}_{kj} - \overline{\mathbf{x}}_k)(\mathbf{x}_{kj} - \overline{\mathbf{x}}_k)^{\mathrm{T}}$$

$$\text{and} \quad B = \sum_{k=1}^{K} n_k \, (\overline{\mathbf{x}}_k - \overline{\mathbf{x}})(\overline{\mathbf{x}}_k - \overline{\mathbf{x}})^{\mathrm{T}}.$$

In terms of the above notation, we shall define the canonical variates of the observed $n \times p$ data matrix X, where

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{bmatrix},$$

and show how the canonical variates can be calculated. We shall denote by $\mathbf{x}^T$ the general $1 \times p$ row vector of X.

Definition: The r'th *canonical variate* is that linear compound $\boldsymbol{\ell}_r^T \mathbf{x}$ that maximizes the ratio of between to within group sample variance, subject to having zero within group correlation with the preceding canonical variates 1, ..., $r-1$.

Thus the canonical variates (the CV's) are the linear combinations of the original p variables that *separate* the K sample mean vectors maximally.

Note 1 As described above, canonical variates are essentially optimal discriminant functions that may be defined without reference to any distributional assumptions. They may therefore be used purely as an exploratory, hypothesis generating tool to investigate the differences between the samples. If we are able to assume multivariate normality, then we can carry out a MANOVA test, and if H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_K$, is rejected then we have some evidence that there is something to investigate.

Note 2 From last year we know that finding the first canonical variate gives us the UIT statistic for testing H_0, involving the largest eigenvalue, λ_1, and corresponding eigenvector, $\boldsymbol{\ell}_1$, of $W^{-1}B$. In addition the remaining canonical variates correspond to successive eigenvalues and eigenvectors of $W^{-1}B$.

Note 3 Investigating the eigenstructure of $W^{-1}B$ is not quite as straightforward as investigating the eigenstructure of a covariance or sums-of-squares and products matrix as in Principal Component Analysis. $W^{-1}B$ is not necessarily symmetric. We have no means of knowing immediately whether its eigenvalues are real. Moreover its eigenvectors are not mutually orthogonal. (Real eigenvalues and orthogonal eigenvectors are necessary properties of real, symmetric matrices.) We therefore choose to work with $W^{-\frac{1}{2}}BW^{-\frac{1}{2}}$, which has the same eigenvalues as $W^{-1}B$ and whose eigenvectors $\boldsymbol{\gamma}$, give us the corresponding eigenvectors, $W^{-\frac{1}{2}}\boldsymbol{\gamma}$, of $W^{-1}B$. Since $W^{-\frac{1}{2}}BW^{-\frac{1}{2}}$ is symmetric and positive semi-definite its eigenvalues, λ_i, are nonnegative and the corresponding eigenvectors $\boldsymbol{\gamma}_i$ are mutually orthogonal. From the definitions of W and B, The maximum possible rank of $W^{-1}B$ is clearly $\min(p, K-1)$. (Clearly we must assume that W is of full rank so that both W^{-1} and $W^{-\frac{1}{2}}$ exist.)

Finding the Canonical Variates

To find the first CV, let $y = \boldsymbol{\ell}^T\mathbf{x}$ and define the ratio of the between-group sample variance of y to its within-group sample variance as

$$R(\boldsymbol{\ell}) = \frac{\boldsymbol{\ell}^T B\boldsymbol{\ell}/(K-1)}{\boldsymbol{\ell}^T W\boldsymbol{\ell}/(n-K)}.$$

Clearly this is invariant under any change of scale in $\boldsymbol{\ell}$. We normalize $\boldsymbol{\ell}$ by requiring that

$$\boldsymbol{\ell}^T W\boldsymbol{\ell} = (n-K),$$

ie by requiring that the within-group sample variance of $y = \boldsymbol{\ell}^T\mathbf{x}$ is one.

Thus it is required to maximize

$$R(\boldsymbol{\ell}) \;=\; \boldsymbol{\ell}^T B \boldsymbol{\ell}/(K-1) \tag{1}$$

subject to

$$\boldsymbol{\ell}^T W \boldsymbol{\ell} \;=\; (n-K). \tag{2}$$

We now let $W^{\frac{1}{2}}$ be the symmetric square root of W and also let

$$\mathbf{u} \;=\; \frac{1}{\sqrt{n-K}}\, W^{\frac{1}{2}} \boldsymbol{\ell}, \tag{3}$$

so that

$$\boldsymbol{\ell} \;=\; \sqrt{n-K}\, W^{-\frac{1}{2}} \mathbf{u} \tag{4}$$

and we can re-express the maximization problem of (1) and (2) in terms of $\mathbf{u}$ rather than $\boldsymbol{\ell}$. Substituting (4) into (1) and (2), we obtain the problem:

Maximize

$$\frac{n-K}{K-1}\, \mathbf{u}^T\, W^{-\frac{1}{2}} B W^{-\frac{1}{2}}\, \mathbf{u} \tag{5}$$

subject to

$$\mathbf{u}^T \mathbf{u} \;=\; 1. \tag{6}$$

Now the matrix $W^{-\frac{1}{2}} B W^{-\frac{1}{2}}$ has p nonnegative eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$, where some of these may be zero. Denote the corresponding, mutually orthogonal, eigenvectors by $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_p$. Clearly the maximum in the maximization problem of (5) and (6) is given by

$$\frac{n-K}{K-1}\, \lambda_1,$$

and the maximizing $\mathbf{u}$ is just $\boldsymbol{\gamma}_1$. (cf Principal Components Analysis.) Since λ_1 is also the largest eigenvalue of $W^{-1}B$, this gives us the UIT statistic for 1-way MANOVA as seen last year.

Translating back using equation (4), we see that the maximizing $\boldsymbol{\ell}$ is given by

$$\boldsymbol{\ell}_1 \;=\; \sqrt{n-K}\, W^{-\frac{1}{2}} \boldsymbol{\gamma}_1.$$

Thus the first CV is given by $y_1 = \boldsymbol{\ell}_1^T \mathbf{x}$, where $\boldsymbol{\ell}_1$ is the eigenvector of $W^{-1}B$ corresponding to its largest root λ_1, and scaled so that

$$\boldsymbol{\ell}_1^T [(n-K)^{-1}\, W] \boldsymbol{\ell}_1 \;=\; 1,$$

so that the within-groups variance of y_1 is one.

To find the second CV, it is required to maximize

$$R(\boldsymbol{\ell}) \;=\; \boldsymbol{\ell}^T B\boldsymbol{\ell}/(K-1) \tag{7}$$

subject to

$$\boldsymbol{\ell}^T W\boldsymbol{\ell} \;=\; (n-K) \tag{8}$$

$$\boldsymbol{\ell}^T W\boldsymbol{\ell}_1 \;=\; 0. \tag{9}$$

Once again set

$$\mathbf{u} \;=\; \frac{1}{\sqrt{n-K}}\, W^{\frac{1}{2}}\boldsymbol{\ell} \;\Longrightarrow\; \boldsymbol{\ell} \;=\; \sqrt{n-K}\, W^{-\frac{1}{2}}\mathbf{u}.$$

From this, and recalling that $\boldsymbol{\ell}_1 \;=\; \sqrt{n-K}\, W^{-\frac{1}{2}}\boldsymbol{\gamma}_1$, the problem becomes:

Maximize

$$\frac{n-K}{K-1}\, \mathbf{u}^T\, W^{-\frac{1}{2}} B W^{-\frac{1}{2}}\, \mathbf{u} \tag{10}$$

subject to

$$\mathbf{u}^T\mathbf{u} \;=\; 1 \tag{11}$$

$$\mathbf{u}^T\boldsymbol{\gamma}_1 \;=\; 0. \tag{12}$$

Again by analogy with Principal Components Analysis, it is clear that the maximum is given by

$$\frac{n-K}{K-1}\, \lambda_2,$$

with the maximizing $\mathbf{u}$ given by $\boldsymbol{\gamma}_2$.

Thus the second CV is just $y_2 \;=\; \boldsymbol{\ell}_2^T\mathbf{x}$, where $\boldsymbol{\ell}_2 \;=\; \sqrt{n-K}\, W^{-\frac{1}{2}}\boldsymbol{\gamma}_2$ is the eigenvector of $W^{-1}B$ corresponding to its second largest root λ_2, and scaled so that

$$\boldsymbol{\ell}_2^T[(n-K)^{-1}\, W]\boldsymbol{\ell}_2 \;=\; 1.$$

Thus the within-groups variance of y_2 is one. Clearly

$$\boldsymbol{\ell}_2^T W\boldsymbol{\ell}_1 \;=\; (n-K)\boldsymbol{\gamma}_2^T W^{-\frac{1}{2}} W W^{-\frac{1}{2}}\boldsymbol{\gamma}_1 \;=\; 0$$

and so the first two canonical variates are uncorrelated within groups.

The Canonical Variate Space

In general $\boldsymbol{\ell}_r \;=\; \sqrt{n-K}\, W^{-\frac{1}{2}}\boldsymbol{\gamma}_r, \quad r \;=\; 1,\ldots,p$, where the $\boldsymbol{\ell}_r$ are eigenvectors of $W^{-1}B$ corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0$. We need to specify the scaling for these eigenvectors.

Let $L = (\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_p)$ and $\Gamma = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p)$. Clearly $\Gamma^T\Gamma = \Gamma\Gamma^T = I$, since the $\boldsymbol{\gamma}_r$ are mutually orthogonal. Then

$$L = \sqrt{n-K}\, W^{-\frac{1}{2}}\Gamma \tag{13}$$

Transforming the general vector $\mathbf{x}$ of original variables to the general vector $\mathbf{y}$ of canonical variables, we have

$$\mathbf{y} = L^T\mathbf{x}.$$

(This transforms the $n \times p$ data matrix X to the $n \times p$ matrix Y given by $Y = XL$, the data matrix referred to canonical axes.)

Thus, from (13), see that the within-groups sample covariance matrix for $\mathbf{y}$ is given by

$$(n-K)^{-1}L^TWL = \Gamma^TW^{-\frac{1}{2}}WW^{-\frac{1}{2}}\Gamma = I.$$

That is to say the eigenvectors $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_p$ are scaled so that

$$L^T[(n-K)^{-1}W]L = I. \tag{14}$$

Thus the canonical variates $\mathbf{y}$ are uncorrelated within groups, and the within-group covariance matrix for $\mathbf{y}$ is I. In fact the canonical variates are also uncorrelated between groups. This is easily seen. Clearly

$$BL = WL\Lambda,$$

where $\Lambda = \mathrm{diag}(\lambda_r)$, and pre-multiplying by L^T gives the diagonal matrix

$$L^TBL = L^TWL\Lambda = (n-K)\Lambda \text{ from (14).} \tag{15}$$

From the preceding analysis, we see that

$$(\lambda_1 + \ldots + \lambda_r)/(\lambda_1 + \ldots + \lambda_p)$$

represents the proportion of between-group variation accounted for by the first r canonical variates.

It is easy to show that the canonical variates are scale invariant in the same sense that regression is scale invariant - unlike principal components. (Do as an exercise. Set $\mathbf{z} = D\mathbf{x}$, where D is a diagonal matrix, and consider the first canonical variate.)

In general we use the canonical variates to plot the group centroids with respect to the new canonical variate axes. Therefore let

$$\overline{\mathbf{y}}_k = L^T\overline{\mathbf{x}}_k \quad k = 1, \ldots, K \tag{16}$$

Then $\overline{\mathbf{y}}_k$ is the $p \times 1$ vector of canonical variable means for sample k.

What do distances in canonical variate space represent?

Suppose that we take a pair of samples k_1, k_2. The squared Euclidean distance between $\overline{\mathbf{y}}_{k_1}$ and $\overline{\mathbf{y}}_{k_2}$ is given by

$$(\overline{\mathbf{y}}_{k_1} - \overline{\mathbf{y}}_{k_2})^T(\overline{\mathbf{y}}_{k_1} - \overline{\mathbf{y}}_{k_2}) \;=\; (\overline{\mathbf{x}}_{k_1} - \overline{\mathbf{x}}_{k_2})^T LL^T(\overline{\mathbf{x}}_{k_1} - \overline{\mathbf{x}}_{k_2}) \quad \text{from (16)} \tag{17}$$

Now, from (13) we have

$$LL^T \;=\; (n-K)W^{-\frac{1}{2}}\Gamma\Gamma^T W^{-\frac{1}{2}} = \left(\frac{1}{n-K}\,W\right)^{-1}, \tag{18}$$

the inverse of the pooled within-groups sample covariance matrix for the original variables. Thus, substituting (18) in (17), we have

$$(\overline{\mathbf{y}}_{k_1} - \overline{\mathbf{y}}_{k_2})^T(\overline{\mathbf{y}}_{k_1} - \overline{\mathbf{y}}_{k_2}) \;=\; (\overline{\mathbf{x}}_{k_1} - \overline{\mathbf{x}}_{k_2})^T \left(\frac{1}{n-K}\,W\right)^{-1} (\overline{\mathbf{x}}_{k_1} - \overline{\mathbf{x}}_{k_2}) \;=\; D^2(\overline{\mathbf{x}}_{k_1}, \overline{\mathbf{x}}_{k_2}), \tag{19}$$

the squared Mahalanobis distance between the two samples.

This means that, given the $K \times K$ matrix of between sample Mahalanobis distances, we have identified a set of coordinates reproducing these distances in p dimensions. (In fact in $\text{rank}(W^{-1}B)$ dimensions, where $\text{rank}(W^{-1}B) \leq \min(p, K-1)$.) As usual we now locate the origin at the centroid of the K points $\overline{\mathbf{y}}_k$ $(= L^T\overline{\mathbf{x}}_k)$ representing the sample means in CV space. Then the new coordinates are given by:

$$\overline{Y}_c \;=\; H \begin{bmatrix} \overline{\mathbf{y}}_1^T \\ \vdots \\ \overline{\mathbf{y}}_K^T \end{bmatrix} \;=\; H \begin{bmatrix} \overline{\mathbf{x}}_1^T \\ \vdots \\ \overline{\mathbf{x}}_K^T \end{bmatrix} L \;=\; \overline{X}_c\, L,$$

where, as usual, the $K \times K$ centreing matrix H is given by $H = I - K^{-1}\mathbf{1}\mathbf{1}^T$.

Duality between Canonical Variate Analysis and Classical Scaling

It is natural now to ask whether the canonical axes account successively for the maximum sum of squared between-sample Mahalanobis distances. In other words, is Canonical Variate Analysis (CVA) the exact dual of Classical Scaling (CS) on the matrix of between-sample Mahalanobis distances?

The answer is yes - provided that $n_1 = n_2 = \ldots = n_K$. Otherwise the duality is not exact.

To see this: First consider Classical Scaling. This would analyse $\overline{Y}_c\,\overline{Y}_c^T$, since $\overline{Y}_c$ is a Euclidean representation of the Mahalanobis distances, with its origin at the centroid. Now

$$\overline{Y}_c\,\overline{Y}_c^T \;=\; \overline{X}_c\, LL^T\, \overline{X}_c^T,$$

and analysing the eigenstructure of this matrix is equivalent to analysing the eigenstructure of:

$$LL^T \overline{X}_c^T \overline{X}_c \;=\; (n-K)\, W^{-1}\, B_{\text{uw}} \quad \text{from equation (18)},$$

where B_{uw} is the *unweighted* between-groups sums of squares and products (SSP) matrix. (That is, B_{uw} is not weighted accorded to the sample sizes n_k.)

Thus CS analyses $W^{-1}\, B_{\text{uw}}$, whereas CVA analyses $W^{-1}\, B$.

[Aside: Recall

$$B \;=\; \sum_{k=1}^{K} n_k\, (\overline{\mathbf{x}}_k - \overline{\mathbf{x}})(\overline{\mathbf{x}}_k - \overline{\mathbf{x}})^T, \quad \text{where} \quad \overline{\mathbf{x}} \;=\; \frac{1}{n} \sum_{k=1}^{K} n_k\, \overline{\mathbf{x}}_k,$$

whereas

$$B_{\text{uw}} \;=\; \sum_{k=1}^{K} (\overline{\mathbf{x}}_k - \overline{\mathbf{x}}_{\text{uw}})(\overline{\mathbf{x}}_k - \overline{\mathbf{x}}_{\text{uw}})^T, \quad \text{where} \quad \overline{\mathbf{x}}_{\text{uw}} \;=\; \frac{1}{K} \sum_{k=1}^{K} \overline{\mathbf{x}}_k.\,]$$

<u>Confidence Circles</u>

Since one of the main uses of CVA is to plot sample mean vectors referred to canonical variate axes, it is useful to have some idea of the 'significance' of the distances between samples that we observe. To get this we need to assume multivariate normality. That is we assume that the samples are from K multivariate normal populations, $\text{MVN}(\boldsymbol{\mu}_k, \Sigma)$, $k = 1, \ldots, K$. To check that there are at least some meaningful distances between the samples, we can start by carrying out a 1-way MANOVA to test H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_K$. If this is rejected, then it is worth examining plots of the canonical variate mean vectors with respect to, say, the first two canonical axes. As mentioned last year, for each k we can then superimpose an approximate circular confidence region for $L^T \boldsymbol{\mu}_k = \boldsymbol{\mu}_k^*$ in 2-dimensional CV space, centred on $\overline{\mathbf{y}}_k$. A larger circular confidence region can be calculated for the positions of *individuals* from population k.

Confidence regions for the $\boldsymbol{\mu}_k^*$ in any subspace of canonical variate space are easily derived as follows. We have $\mathbf{x}_{jk} \sim \text{MVN}(\boldsymbol{\mu}_k, \Sigma)$, $k = 1, \ldots, K$; $j = 1, \ldots, n_k$. Assume that we have transformed to the s canonical variates ($s = \text{rank}(W^{-1}B)$ using

$$\mathbf{y}_{kj} \;=\; L^T\, \mathbf{x}_{kj} \;\Longrightarrow\; \mathbf{y}_{kj} \sim \text{MVN}(\boldsymbol{\mu}_k^*, L^T \Sigma L),$$

where L is the $p \times s$ matrix $(\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_s)$.

If the degrees of freedom of W, $(n-K)$, are large, then $\Sigma \;\approx\; (n-K)^{-1}\, W$ and so

$$L^T \Sigma L \;\approx\; (n-K)^{-1}\, L^T W L \;=\; I.$$

Then

$$\mathbf{y}_{kj} \overset{\text{approx}}{\sim} \text{MVN}(\boldsymbol{\mu}_k^*, I),$$

so that

$$\overline{\mathbf{y}}_k \overset{\text{approx}}{\sim} \text{MVN}(\boldsymbol{\mu}_k^*, n_k^{-1}\, I),$$

and

$$\sqrt{n_k}\,(\overline{\mathbf{y}}_k - \boldsymbol{\mu}_k^*) \overset{\text{approx}}{\sim} \text{MVN}(\mathbf{0}, \mathbf{I}).$$

It follows that

$$n_k(\overline{\mathbf{y}}_k - \boldsymbol{\mu}_k^*)^T(\overline{\mathbf{y}}_k - \boldsymbol{\mu}_k^*) \overset{\text{approx}}{\sim} \chi_s^2$$

Hence a $100(1-\alpha)\%$ Confidence Region for the true $s \times 1$ mean vector $\boldsymbol{\mu}_k^*$, relative to the first s canonical axes, is given by the interior of the hypersphere of radius $\sqrt{\chi^2_{s;\alpha}/n_k}$ centred at $\overline{\mathbf{y}}_k$. Thus, if we plot the canonical variate means with respect to the first two canonical axes, we can superimpose on the the plot *approximate* circular confidence regions, centred at the $\overline{\mathbf{y}}_k$ and of radius $\sqrt{\chi^2_{2;\alpha}/n_k}$, for the true means $\boldsymbol{\mu}_k^*$. Confidence circles about $\overline{\mathbf{y}}_k$ for the individuals from each group have *approximate* radius $\sqrt{\chi^2_{2;\alpha}}$. (Note: $\chi^2_{2;0.05} = 5.99$ so that 95% confidence circles for the $\boldsymbol{\mu}_k^*$ have radii $(2.45/\sqrt{n_k})$ in two dimensions.)

Note 1: To assess the importance of the original variables in discriminating between the samples, we can look at the coefficients of these variables in the Canonical Variates. However the coefficients of different variables are not comparable since they depend on the scale on which the variables were measured. For the coefficients to be comparable, we need to scale the original variables to have common with-groups variance. (cf regression coefficients.)

Note 2: The best way to determine the importance of each variable is to try leaving them out of the analysis in turn and assess the change in some suitable criterion (eg % correctly classified). See also Stepwise Discriminant Analysis in the texts.

Note 3: Note that confidence circles are fairly robust to non-normality (Central Limit Theorem).

Note 4: To classify a new unit, $\mathbf{x}$, into one of the K populations, assign $\mathbf{x}$ to the population k for which:

$$D^2(\mathbf{x}, \overline{\mathbf{x}}_k) \;=\; (\mathbf{x} - \overline{\mathbf{x}}_k)^T[(n-K)^{-1}W]^{-1}(\mathbf{x} - \overline{\mathbf{x}}_k)$$

is minimum. A measure of how well the classification works is the percentage of individual units correctly classified. This should be calculated by leaving out each unit in turn and then classifying using the Mahalanobis distances calculated using the remaining units.