5.4 Canonical Variates Analysis

Suppose that we have available an $n \times m$-dimensional data matrix X, in which the n individuals are divided into g groups with n_k individuals in the k-th group, so that $n = \sum_{k=1}^{g} n_k$. We want to investigate possible differences between the groups.

The matrix of observations is following

$$\underline{\mathsf{X}} = \begin{pmatrix} \mathsf{x}_{11}^{(1)} & \dots & \mathsf{x}_{1m}^{(1)} \\ \mathsf{x}_{21}^{(1)} & \dots & \mathsf{x}_{2m}^{(1)} \\ \vdots & & \vdots \\ \mathsf{x}_{n_1 1}^{(1)} & \dots & \mathsf{x}_{n_1 m}^{(1)} \\ \vdots & & \vdots \\ \mathsf{x}_{11}^{(g)} & \dots & \mathsf{x}_{1m}^{(g)} \\ \mathsf{x}_{21}^{(g)} & \dots & \mathsf{x}_{2m}^{(g)} \\ \vdots & & \vdots \\ \mathsf{x}_{n_g 1}^{(g)} & \dots & \mathsf{x}_{n_g m}^{(g)} \end{pmatrix}$$

Let

$$\bar{\mathsf{x}}_j^{(k)} = \frac{1}{n_k} \sum_{l=1}^{n_k} \mathsf{x}_{lj}^{(k)}, \quad k = 1, \dots, g, \ j = 1, \dots, m$$

be the mean value of the j-th variable in the k-th group, and let

$$\bar{\mathsf{x}}_j = \frac{1}{n} \sum_{k=1}^{g} \sum_{l=1}^{n_k} \mathsf{x}_{lj}^{(k)}, \quad j = 1, \dots m$$

be the mean of the j-th variable, which may also be written as

$$\bar{\mathsf{x}}_j = \frac{1}{n} \sum_{k=1}^{g} n_k \frac{1}{n_k} \sum_{l=1}^{n_k} \mathsf{x}_{lj}^{(k)} = \frac{1}{n} \sum_{k=1}^{g} n_k \bar{\mathsf{x}}_j^{(k)}.$$

Now, let $\underline{\bar{X}}$ be a $(g \times m)$-dimensional matrix of centered means,

$$\underline{\bar{X}} = \begin{pmatrix} \bar{\mathsf{x}}_1^{(1)} - \bar{\mathsf{x}}_1 & \dots & \bar{\mathsf{x}}_m^{(1)} - \bar{\mathsf{x}}_m \\ \bar{\mathsf{x}}_1^{(2)} - \bar{\mathsf{x}}_1 & \dots & \bar{\mathsf{x}}_m^{(2)} - \bar{\mathsf{x}}_m \\ \vdots & & \vdots \\ \bar{\mathsf{x}}_1^{(g)} - \bar{\mathsf{x}}_1 & \dots & \bar{\mathsf{x}}_m^{(g)} - \bar{\mathsf{x}}_m \end{pmatrix}$$

Then matrix $\underline{B} = \underline{\bar{X}}'\underline{\bar{X}}$ represents **between group variation**, that is

$$\underline{B} = \underline{\bar{X}}'\underline{\bar{X}} = \begin{pmatrix} \sum_{k=1}^{g}(\bar{\mathsf{x}}_1^{(k)} - \bar{\mathsf{x}}_1)^2 & \dots & \sum_{k=1}^{g}(\bar{\mathsf{x}}_1^{(k)} - \bar{\mathsf{x}}_1)(\bar{\mathsf{x}}_m^{(k)} - \bar{\mathsf{x}}_m) \\ \vdots & \ddots & \vdots \\ & \dots & \sum_{k=1}^{g}(\bar{\mathsf{x}}_m^{(k)} - \bar{\mathsf{x}}_m)^2 \end{pmatrix}$$

Also, denote by $\underline{X}_k$, $k = 1, \dots, g$, matrix of centered data for group k, that is

$$\underline{X}_k = \begin{pmatrix} \mathsf{x}_{11}^{(k)} - \bar{\mathsf{x}}_1^{(k)} & \dots & \mathsf{x}_{1m}^{(k)} - \bar{\mathsf{x}}_m^{(k)} \\ \mathsf{x}_{21}^{(k)} - \bar{\mathsf{x}}_1^{(k)} & \dots & \mathsf{x}_{2m}^{(k)} - \bar{\mathsf{x}}_m^{(k)} \\ \vdots & & \vdots \\ \mathsf{x}_{n_k 1}^{(k)} - \bar{\mathsf{x}}_1^{(k)} & \dots & \mathsf{x}_{n_k m}^{(k)} - \bar{\mathsf{x}}_m^{(k)} \end{pmatrix}$$

Then

$$\underline{W}_k = \underline{X}_k'\underline{X}_k$$

is sum of squares and products matrix for the k-th group, after centering within the k-th group. Let

$$\underline{W} = \sum_{k=1}^{g} \underline{W}_k.$$

$\underline{W}$ represents **the within-groups variability**. Furthermore,

$$\underline{S} = \underline{B} + \underline{W}$$

is the sum of squares and products matrix for centered observations $\underline{X}$. $\underline{B}$, $\underline{W}$ and $\underline{S}$ are symmetric non-negative definite matrices.

The function

$$F = \frac{\frac{1}{g-1}\,\underline{w}'\underline{B}\underline{w}}{\frac{1}{n-g}\,\underline{w}'\underline{W}\underline{w}} \underset{H_0}{\sim} \mathcal{F}_{g-1,n-g},$$

where H_0 is the hypothesis that there is no difference between groups, has $\mathcal{F}$ distribution with $g-1$ and $n-g$ degrees of freedom and is a test function for this hypothesis. The hypothesis will be rejected if the ratio is large, what means that the within-groups variability is smaller than the between-groups variability. Hence, we determine vector w so as to maximize the function $\mathcal{F}$.

For fixed g and n it is equivalent to maximize

$$f(\underline{w}) = \frac{\underline{w}'\underline{B}\underline{w}}{\underline{w}'\underline{W}\underline{w}}.$$

Differentiating $f(\underline{w})$ with respect to w and setting the result to zero yields

$$\frac{2\underline{Bw}(\underline{w}'\underline{Ww}) - 2\underline{Ww}(\underline{w}'\underline{Bw})}{(\underline{w}'\underline{Ww})^2} = 0$$

Multiplying both sides by $\frac{1}{2}\underline{w}'\underline{Ww}$ we get

$$\underline{Bw} - \underline{Ww}\left(\frac{\underline{w}'\underline{Bw}}{\underline{w}'\underline{Ww}}\right) = 0.$$

Denote

$$l = \frac{\underline{w}'\underline{Bw}}{\underline{w}'\underline{Ww}},$$

which is a constant value (maximum of $f(w)$. Then, we may write

$$\underline{Bw} - \underline{Ww}l = 0$$

and

$$\left(\underline{W}^{-1}\underline{B} - l\underline{I}\right)\underline{w} = 0.$$

So, l must be an eigenvalue of $\underline{W}^{-1}\underline{B}$ and $\underline{w}$ the corresponding eigenvector of $\underline{W}^{-1}\underline{B}$. Furthermore, l is the maximum of $f(\underline{w})$ and so $\underline{w}$ is the eigenvector corresponding to the largest eigenvalue of $\underline{W}^{-1}\underline{B}$.

The vector $\underline{w}$ determines the linear combination of the centered data, i.e., $\underline{w}'\underline{x}$.

Definition
The first canonical variate is the vector $\underline{w}_1$ in $\mathbb{R}^m$ which maximizes $\underline{w}'\underline{Bw}$ subject to $\underline{w}'\underline{Ww} = 1$.

Note
$\underline{w}_1$ maximizes the ratio of between-groups spread in direction $\underline{w}_1$ to within-groups spread in this direction.

Definition
The i-th canonical variate is the vector $\underline{w}_i \in \mathbb{R}^m$ such that

1. $\underline{w}_i'\underline{Ww}_j = \begin{cases} 1, & \textit{if } i = j; \\ 0, & \textit{for } j = 1, \ldots, i-1. \end{cases}$

2. *$\underline{w}_i$ maximizes $\underline{w}'\underline{Bw}$ subject to (1).*

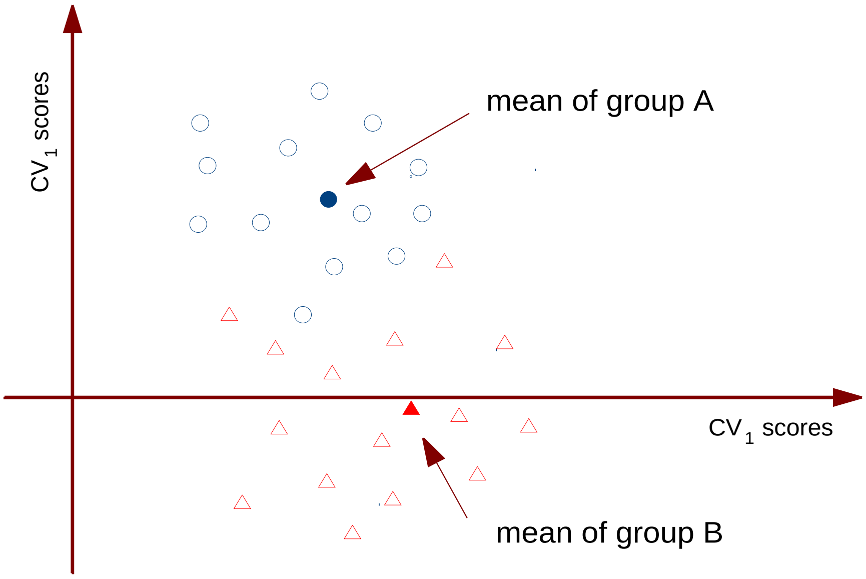

Figure 5.12: First and Second CV scores and their means

Since the eigenvalues l_i, $i = 1, \ldots, s$, $s = min\{m, g - 1\}$, measure how much between-groups variability is taken up by each canonical variate, we need

$$\frac{l_1 + l_2}{\sum_{i=1}^{s} l_i} \tag{5.3}$$

to be large to represent the groups well in a two-dimensional space spanned by $\underline{w}_1$ and $\underline{w}_2$.
When there are many groups and large amount of data it may be more reasonable to plot means of the group CVs only.

If the ratio (5.3) is not large, then more dimensions may be needed. In such case, other techniques, like Andrews curves, will generally provide a good adjunct to CVA.

Notes

- $rank(B) \leq g - 1$, so we have at most $g - 1$ canonical variates;
- $\underline{w}_1$ gives most information about the differences between groups provided that each of the $\underline{W}_i$ has approximately the same structure (similarly as in one-way ANOVA);
- plotting points in CV_1 and CV_2 space gives best two-dimensional picture showing the separation of the groups.