2 Mardia, Kent, and Bibby's and Chang's Methods on Selection of Principal Components as Discriminant Variables

2.1 Principal Component Analysis

Let $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ be a p dimensional random vector with covariance matrix $\mathbf{\Sigma}$(or correlation matrix ρ). Suppose $\lambda_1, \lambda_2, \ldots, \lambda_p$ are eigenvalues of $\mathbf{\Sigma}$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, and $\mathbf{e}_i$ is the eigenvector of $\mathbf{\Sigma}$ corresponding to $\lambda_i, i = 1, 2, \ldots, p$. Then the ith principal component is given by

$$\mathbf{Y}_i = \mathbf{e}_i'\mathbf{X} \text{ with } \text{Var}\ (\mathbf{Y}_i) = \lambda_i \quad i = 1, 2, \ldots, p$$

and

$$\text{Cov}\ (\mathbf{Y}_i, \mathbf{Y}_k) = 0 \quad i \neq k.$$

We have

$$\begin{pmatrix} \text{Proportion of total} \\ \text{population variance} \\ \text{due to } kth \text{ Principal} \\ \text{component} \end{pmatrix} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}.$$

Based on the proportion of total population variance, we often take the first k principal components to replace the original p variables.

2.2 Selection of Principal Component Variables

In this section, we give two different methods to select the best principal component variables for further discrimination.

2.2.1 Mardia et al. Method

Suppose $\mathbf{y}$ is a p dimensional random variable with a mixture of two multi-normal distributions with means $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and a common covariance matrix $\boldsymbol{\Sigma}$. That is, a random sample $\mathbf{y}_{11}, \mathbf{y}_{12}, \ldots, \mathbf{y}_{1n_1} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$; independently, a random sample $\mathbf{y}_{21}, \mathbf{y}_{22}, \ldots, \mathbf{y}_{2n_2} \sim N_p\ (\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Suppose the mixing proportions are π and $1-\pi$, and define $\bar{\mathbf{y}}_i = \dfrac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{y}_{ij},\ i = 1, 2;\ \bar{\mathbf{y}} = \dfrac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\mathbf{y}_{ij}$. Then the total sum of squares matrix of the $\mathbf{y}_{ij}$ about the overall mean $\bar{\mathbf{y}}$ is

$$\mathrm{T} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}})(\mathbf{y}_{ij} - \bar{\mathbf{y}})'.$$

We can decompose T into within group and between group sums of squares matrices, $\mathbf{W}$ and $\mathbf{B}$ respectively, that is

$$\mathbf{T} = \mathbf{W} + \mathbf{B},$$

where $\mathbf{W} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)'$ and $\mathbf{B} = \sum_{i=1}^{2} n_i(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})' = c\mathbf{d}\mathbf{d}'$ with $c = \frac{n_1 n_2}{n_1+n_2}$ and $\mathbf{d} = \bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2$. Consider the discriminant problem between the two multi-normal populations. The coefficients of the sample maximum likelihood discriminant function (Mardia et al.1979) are given by $\boldsymbol{\beta} = (\frac{\mathbf{W}}{m})^{-1}\mathbf{d}$, with kth element β_k. Now, partition $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$, and $\mathbf{d}' = (\mathbf{d}_1', \mathbf{d}_2')$, where $\boldsymbol{\beta}_1$ and $\mathbf{d}_1$ have k components. Suppose $\boldsymbol{\beta}_2 = \mathbf{0}$, that is suppose those variables $\mathbf{y}_k, \mathbf{y}_{k+1}, \ldots, \mathbf{y}_p$ have no discriminant power and hence can be safely discarded. And this is equivalent to $D_p^2 = D_k^2$, where :

$$D_p^2 = m\mathbf{d}'\mathbf{W}^{-1}\mathbf{d} \text{ and } D_k^2 = m\mathbf{d}_1'\mathbf{W}^{-1}\mathbf{d}_1.$$

Here, D_p^2 is the Mahalanobis distance computed based on the original p variables and D_k^2 is the Mahalanobis distance computed based on some k variables. This test uses the statistic, called M's statistic,

$$\eta_k = \frac{(m-p+1)}{(p-k)}c(D_p^2 - D_k^2)/(m + cD_k^2). \tag{2.1}$$

And under the null hypothesis, this statistic has the $F_{p-k,m-p+1}$ distribution. If η_k is small, it means that D_k^2 approaches D_p^2. That is, all variables, except these k variables have no discriminant power. When applying this method to select principal components as discriminant variables, we can use the following procedure:

1. Partition $\boldsymbol{\beta}'=(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$, such that$\boldsymbol{\beta}_1 \in \mathbb{R}$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^{p-1}$. Then compute the M's statistic $\eta_1(k) = \frac{(m-p+1)}{(p-1)} c(D_p^2 - D_k^2)/(m + cD_k^2)$ corresponding to the kth principal component, PC_k, $k = 1, 2, \ldots, p$. Let $\eta_1(i)$ be the smallest M's statistic value. We select PC_i in this step.

2. Reset $\boldsymbol{\beta}=(\beta_i, \beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_p)'$. Partition $\boldsymbol{\beta}'=(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$, where $\boldsymbol{\beta}_1$ is a 2-dimensional vector with first component β_i. Then compute the M's statistic $\eta_2(k) = \frac{(m-p+1)}{(p-2)} c(D_p^2 - D_k^2)/(m + cD_k^2)$ of each set $\{\mathrm{PC}_i, \mathrm{PC}_k\}$ $\forall k = 1, 2, \ldots, p, k \neq j$. Let $\eta_2(j)$ be the smallest M's statistic value. We select PC_i, PC_j in this step.

3. Reset $\boldsymbol{\beta}=(\beta_i, \beta_j, \ldots, \beta_p)'$. Partition $\boldsymbol{\beta}'=(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$, where $\boldsymbol{\beta}_1$ is a 3-dimensional vector with first and second components β_i and β_j, respectively. Then compute the M's statistic $\eta_3(k) = \frac{(m-p+1)}{(p-3)} c(D_p^2 - D_k^2)/(m + cD_k^2)$ of each set $\{\mathrm{PC}_i, \mathrm{PC}_j, \mathrm{PC}_k\}$ for $k = 1, 2, \ldots, p, k \neq i, j$. Let $\eta_3(l)$ be the smallest M's statistic value. We select PC_i, PC_j and PC_l in this step.

Continue this process until the smallest η_k is less than the critical value $F_{p-k,m-p+1}(\alpha)$. All principal components selected in the last step will be taken as discriminant variables.

2.2.2 Chang's Method

Let Δ denote the sample Mahalanobis distance between the two sub-populations, and $\mathbf{V}$ be the covariance matrix of $\mathbf{y}$. Then we have

$$\mathbf{V} = \pi(1-\pi)\mathbf{d}\mathbf{d}' + \mathbf{W}$$

$$\Delta^2 = \mathbf{d}'\mathbf{W}^{-1}\mathbf{d}.$$

Consider the Mahalanobis distance based on some (may not be all) principal components. Let $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ be the p eigenvectors of $\mathbf{V}$, and their corresponding eigenvalues are $\lambda_1, \lambda_2, \ldots, \lambda_p$, respectively. In the spectral representation, $\mathbf{V} = \sum_{i=1}^{p} \lambda_i \mathbf{e}_i \mathbf{e}_i'$, where $\mathbf{e}_i'\mathbf{e}_j = 1$ if $i = j$, $\mathbf{e}_i'\mathbf{e}_j = 0$ if $i \neq j$. Moreover, $\mathbf{e}_i'\mathbf{V}\mathbf{e}_i = \lambda_i$, $\mathbf{e}_i'\mathbf{V}\mathbf{e}_j = 0$. Let $B_m = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m)$ and Δ_m be the Mahalanobis distance between the two populations using $B_m'\mathbf{y}$, where $m \leq p$. That is, Δ_m is the Mahalanobis distance based on m principal component values. Chang (1983) gave the following proposition about Δ_m^2.

Proposition

$$\Delta_m^2 = \sum_{i=1}^{m} \frac{(\mathbf{e}_i'\mathbf{d})^2}{\lambda_i} / (1 - \pi(1-\pi) \sum_{i=1}^{m} \frac{(\mathbf{e}_i'\mathbf{d})^2}{\lambda_i}).$$

In particular, for $m = 1$,

$$\Delta_{1k}^2 = \frac{(\mathbf{e}_k'\mathbf{d})^2}{\lambda_k} / (1 - \pi(1-\pi)\frac{(\mathbf{e}_k'\mathbf{d})^2}{\lambda_k}), \text{ where } k = 1, 2, \ldots, p.$$

Chang (1983) called this distance Δ_m as "the information contained in the variables." Moreover, he generated 300 random observations in a 15 dimensional mixture. He found that the discriminant result was the best when taking the first and the last components as the discriminant variables.

When using this method to select principal components as discriminant variables, we can consider the Mahalanobis distance Δ_{1k}^2 computed based on the kth principal component, $k = 1, 2, \ldots, p$. We can reach a selection order of principal components, based on the order (from the largest to the smallest) of Δ_{1k}^2's, where $k = 1, 2, \ldots, p$.