We want to classify the gender for each person from the weight to construct this function a training sample is used in which the gender is unknown in the data set most often applied classification rules are based on the each object is assumed to be a member of one (and only one) group for each object, an error is incurred if the object is attached to the wrong group the measurements of all objects of one class are collected the prior probability \( \pi_k \) for each class is equal to minimizing the expected loss of classification.

The purpose of discriminant analysis is to assign objects to one of several (K) groups based on a set of measurements \( X = (X_1, X_2, \ldots, X_p) \) which are obtained from each object:

- each object is assumed to be a member of one (and only one) group \( 1 \leq k \leq K \)
- an error is incurred if the object is attached to the wrong group
- the measurements of all objects of one class are characterized by a probability density \( f_k(X) \)
- we want to find a rule to decide for every object to which class it belongs to

The class conditional densities \( f_k(X) \) are seldom known:

\[ f_k(X) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2} (X - \mu_k)^T \Sigma_k^{-1} (X - \mu_k)} \]

where \( \mu_k \) and \( \Sigma_k \) are the class \( k \) (\( 1 \leq k \leq K \)) population mean vector and covariance matrix.

\( f_k(X) \) are seldom known

The prior and unconditional distribution:

There might be some prior knowledge about the probability of observing a member of class \( k \):

- **prior probability** \( \pi_k \) with \( \pi_1 + \pi_2 + \pi_k = 1 \)
- if the prior probabilities are equal for each \( k \) it leads to Maximum-Likelihood
- to estimate the class conditional densities \( f_k(X) \) and the prior probability \( \pi_k \), a training sample with already correctly classified classes is used.

the unconditional distribution of \( X \) is given by

\[ f(X) = \sum_{k=1}^{K} \pi(k) f(X | k) \]

There might be some prior knowledge about the probability of the class distribution:

\[ f(X) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (X - \mu)^T \Sigma^{-1} (X - \mu)} \]

where \( \mu \) and \( \Sigma \) are the population mean vector and covariance matrix.

The posterior distribution:

Probability for one object with given vector \( X^T = (X_1, \ldots, X_p) \) to belong to class \( k \) is calculated by the Bayes formula:

\[ p(k | X) = \frac{f(X | k) \cdot \pi(k)}{f(X)} \propto f(X | k) \cdot \pi(k) \]

Posterior distribution

an object is assigned to class \( k \), if it has the biggest posterior probability \( p(k | X) \)

\[ \Rightarrow \text{this is equal to minimizing the expected loss} \]

Example

A group of people consist of male and female persons

\( \Rightarrow K = 2 \)

from each person the data of their weight and height is collected

\( \Rightarrow p = 2 \)

the gender is unknown in the data set

we want to classify the gender for each person from the weight and height

\( \Rightarrow \) discriminant analysis

a classification rule is needed (discriminant function) to choose the group for each person

to construct this function a training sample is used in which the gender is already known

\[ \sum_{k=1}^{K} \pi_k f(X | k) \]

Example

A group of people consist of male and female persons

\( \Rightarrow K = 2 \)

from each person the data of their weight and height is collected

\( \Rightarrow p = 2 \)

the gender is unknown in the data set

we want to classify the gender for each person from the weight and height

\( \Rightarrow \) discriminant analysis

a classification rule is needed (discriminant function) to choose the group for each person

to construct this function a training sample is used in which the gender is already known

\[ \sum_{k=1}^{K} \pi_k f(X | k) \]
Log posterior distribution

For easier calculation we take the logarithm of the posterior distribution
\[
\log p(k \mid X) = \log f(X \mid k) + \log \pi(k)
\]

- with the multivariate normal distribution it leads to
\[
\log p(k \mid X) = -\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \frac{1}{2} \log |\Sigma_k| + \log \pi_k
\]

- the constant term \(-\frac{1}{2} \log (2\pi)\) can be omitted as it is the same for each class \(k\)

\[
(1)
\]

Quadratic discriminant analysis

Multiplication with \(-2\) leads to the discriminant function
\[
d_k(X) = (X - \mu_k)^T \Sigma_k^{-1} (X - \mu_k) + \log |\Sigma_k| - 2 \log \pi_k
\]

and to the classification rule
\[
d_k(X) = \min_{1 \leq k \leq K} d_k(X) \iff \max_{1 \leq k \leq K} p(k \mid X)
\]

Using this rule is called the Quadratic Discriminant Analysis (QDA)

For easier calculation we take the logarithm of the posterior distribution
\[
\log p(k \mid X) = \log f(X \mid k) + \log \pi(k)
\]

- with the multivariate normal distribution it leads to
\[
\log p(k \mid X) = -\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \frac{1}{2} \log |\Sigma_k| + \log \pi_k
\]

- the constant term \(-\frac{1}{2} \log (2\pi)\) can be omitted as it is the same for each class \(k\)

Linear discriminant analysis

A special case occurs when all \(k\) class covariance matrices are identical
\[
\Sigma_k = \Sigma
\]

The discriminant function
\[
d_k(x) = (x - \mu_k)^T \Sigma^{-1} (x - \mu_k) - 2 \log \pi_k
\]

simplifies to
\[
d_k(x) = 2\mu_k^T \Sigma^{-1} x - \mu_k^T \Sigma^{-1} \mu_k - 2 \log \pi_k
\]

This is called the Linear Discriminant Analysis (LDA) because the quadratic terms in the discriminant function cancel:

- \(x^T \Sigma^{-1} x\) is the same in every class \(k\) and can be left out
- the decision boundaries are now linear

Estimation

In most applications of linear and quadratic discriminant analysis the parameters \(\mu\) and \(\Sigma\) are estimated by their sample analogs
\[
\hat{\mu}_k = \bar{X}_k = \frac{1}{N_k} \left[ \begin{array}{c} \sum_{i=1}^{N_k} x_{i1} \\ \vdots \\ \sum_{i=1}^{N_k} x_{ip} \end{array} \right] = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix}
\]

and
\[
\hat{\Sigma}_k = \frac{S_k}{N_k} = \frac{1}{N_k} \sum_{i=1}^{N_k} (X_i - \bar{X}_k)(X_i - \bar{X}_k)^T
\]

with \(c(v) = \text{class of } v\text{th observation}\)

Small sample sizes

These estimates are straightforward to compute and represent the corresponding maximum likelihood estimates.

Problem: they are only optimal for \(n \to \infty\) and not for small \(n\)

Small sample sizes

- the \((p \times p)\) covariance matrix estimates become highly variable
- not all of the parameters are even identifiable
- \(\Sigma\) is singular
- the inverse \(\Sigma^{-1}\) does not exist

<table>
<thead>
<tr>
<th>QDA</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>poorly-posed</td>
<td>(N_k \approx p) (N \approx p)</td>
</tr>
<tr>
<td>ill-posed</td>
<td>(N_k &lt; p) (N &lt; p)</td>
</tr>
<tr>
<td>parameters to be estimated</td>
<td>(k \cdot p^2 + p \cdot p^2 + p)</td>
</tr>
</tbody>
</table>

QDA requires generally larger samples size than LDA
Regularized Discriminant Analysis

For ill- or poorly-posed situations:
- parameter estimates can be highly unstable
- high variance

The aim of regularization is to improve the estimates by biasing them away from their sample-based values
- reduction of variance at the expense of potentially increased bias
- the bias variance trade-off is regulated by two parameters
- these parameters control the strength of the biasing

Further regularization

Strategy 1: A less limited approach is represented by
\[ \hat{\Sigma}_k(\lambda) = (1 - \lambda)\hat{\Sigma}_k + \lambda\hat{\Sigma} \]
with \( 0 \leq \lambda \leq 1 \)
- \( \lambda \) controls the degree of shrinkage of the individual class covariance matrix estimates toward the pooled estimate
- \( \lambda = 0 \) gives rise to QDA
- \( \lambda = 1 \) gives rise to LDA

\[ \Rightarrow \] still fairly limited
\[ \Rightarrow \] cannot be used if LDA is ill- or poorly posed

Strategy 2: A less limited approach is represented by
\[ \hat{\Sigma}_k(\lambda, \gamma) = (1 - \gamma)\hat{\Sigma}_k + \gamma \frac{\text{tr} [\hat{\Sigma}_k(\lambda) ]}{p} \hat{\Sigma} \]
with \( 0 \leq \gamma \leq 1 \)
- \( A = \text{sum of eigenvalues} \)
- the additional regularization parameter \( \gamma \) controls shrinkage toward a multiple of the identity matrix for a given value of \( \lambda \)
- decreasing the larger eigenvalues and increasing the smaller ones

\[ \Rightarrow \] shrinkage toward the average eigenvalue of \( \hat{\Sigma}_k(\lambda) \)

Further regularization

Strategy 3: If LDA is ill- or poorly-posed
\[ \hat{\Sigma}_k(\lambda, \gamma) = (1 - \gamma)\hat{\Sigma}_k + \gamma \frac{\text{tr} [\hat{\Sigma}_k(\lambda) ]}{p} \hat{\Sigma} \]
with \( 0 \leq \gamma \leq 1 \)
- \( A = \text{sum of eigenvalues} \)
- the additional regularization parameter \( \gamma \) controls shrinkage toward a multiple of the identity matrix for a given value of \( \lambda \)
- decreasing the larger eigenvalues and increasing the smaller ones

\[ \Rightarrow \] shrinkage toward the average eigenvalue of \( \hat{\Sigma}_k(\lambda) \)

The discriminant function for the Regularized Discriminant Analysis (RDA) is
\[ d_k(X) = (X - \hat{\Theta}_k)^T \hat{\Sigma}_k^{-1}(\lambda, \gamma)(X - \hat{\Theta}_k) + \log [\hat{\Sigma}_k(\lambda, \gamma)] - 2\log \pi(k) \]
- the values for \( \lambda \) and \( \gamma \) are not likely to be known in advance
- we have to estimate them
- the aim is to find values for \( \lambda \) and \( \gamma \) that jointly minimize the future misclassification risk

Methods:
- bootstrapping
- cross-validation

Idea of cross-validation (leave-one-out)
- one particular observation \( X_i \) is removed from the model
- the classification rule is developed on the \( N - 1 \) training observations without \( X_i \)
- then this classification rule is used to classify \( X_i \) and to calculate the loss which occurs if classified to the wrong group
- this is repeated for every observation

\[ \Rightarrow \] the future misclassification risk is estimated by the average of the resulting misclassification loss
- this is done for a number of combinations for \( \lambda \) and \( \gamma \)
Conclusion

The potential for RDA to improve misclassification risk over that of QDA or LDA depend on the situation

- $N_k \gg p$
  - no regularization is needed and QDA can be used
  - model-selection procedure should tend to small values of $\lambda$ and $\gamma$
- $N = p$
  - LDA has been the method of choice in the past
  - regularization can substantially improve the misclassification risk when
    - $\Sigma_k$ are not close to being equal
    - $N$ is even too small for LDA

Part III

Prediction Analysis with Microarrays (PAM)

Class prediction with gene expression data

The aim is to assign people to one of $K (1 \leq k \leq K)$ diagnostic categories based on their gene expression profile

The classification by DNA microarrays is challenging because:

- there is a very large number of genes ($p$) from which to predict classes and only a relatively small number of samples ($N$)
- identify the genes which contribute most to the classification
- reduction of $p$

Example

Small round blue cell tumors (SRBCT) of childhood can be divided in four groups

- Burkitt lymphoma (BL)
- Ewing sarcoma (EWS)
- neuroblastoma (NB)
- rhabdomyosarcoma (RMS)

The DNA microarrays of 88 children with SRBCT were obtained

- 63 of them were already classified right and their data were used as the training sample to estimate the classification rule
- the category for the other 25 children (of which 5 were not SRBCT) was then predicted by this rule
- the aim is to correctly classify the test samples

Example

The data consist of expression measurements on 2,308 genes

- the mean expression value (centroid) was calculated from the training sample for each of the four classes
- then the squared distance from the gene expression profile to each centroid was calculated for each test sample
- the predicted class for a child was the one with the closest centroid
- nearest centroid classification

It would be more attractive if fewer genes were needed
- modification to nearest shrunk centroid classification
- the genes which don't contribute for the class prediction are eliminated

Shrunken centroids

- $\bar{x}_k = \frac{1}{N_k} \sum_{i} x_{ik}$ mean value in class $k$ for gene $i$
- $\bar{x}_i = \frac{1}{\sum_{k=1}^{K}} x_{ik}$ overall centroid for gene $i$

$d_{ik}$ is a $t$ statistic for gene $i$, comparing the mean of class $k$ to the overall centroid

\[
d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k - (s_i + s_0)}
\]

$\Sigma_k$ pooled standard deviation for gene $i$

The same value for every gene, positive constant $\gamma$

Shrunk Centers

The distance for gene $i$ between the mean in class $k$ and the overall mean is shrunk toward zero $\Rightarrow$ soft thresholding

\[
d_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta) +
\]

with

$\Delta$ shrinkage parameter, also called threshold

$\Delta = \left\{ \begin{array}{ll} \gamma & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{array} \right.$

- many of the genes are eliminated as $\Delta$ increases
- $\Delta$ is again chosen by cross-validation ($\Delta = 4.34$ in the example)

The centroids are shrunk towards the overall centroid

$\tilde{x}_{ik} = \bar{x}_i + m_k(s_i + s_0)d_{ik}$

- if $d_{ik} = 0$ for every class $k$ 
  $\Rightarrow \tilde{x}_{ik} = \ldots = \tilde{x}_{ik}$
- the centroid for each class is the same
- if each group has the same mean for one gene, this gene does not help to predict a class and can be left out
- in the example only 43 genes are needed for the class prediction and 2275 are not needed to distinguish between the groups
The discriminant function uses the Mahalanobis metric in computing distances to centroids:

$$\delta_k^{\text{LDA}}(x^*) = (x^* - \bar{x}_k)^T \Sigma^{-1} (x^* - \bar{x}_k) - 2\log \pi_k$$

with
- $\pi_k$ class prior probability
- $W$ pooled within-class covariance matrix

- $\Sigma$ is huge as $p \gg n$ and any sample estimate will be singular
- to cope with this problem a heavily restricted form of LDA is used
- $\Rightarrow \Sigma$ is assumed to be a diagonal matrix

Classification rule

The discriminant function can be rewritten as

$$\delta_k(x^*) = \sum_{i=1}^p \left( \frac{(x_i^* - \bar{x}_k^i)^2}{s_i + s_0} \right) - 2\log \pi_k$$

- again standardized by $s_i + s_0$
- the discriminant function is for one person with $p$ genes
  $x^* = (x_1^*, x_2^*, \ldots, x_p^*)$
- the person is assigned to group $k$ if
  $$\hat{\delta}_k(x^*) = \min_{1 \leq k \leq K} \delta_k(x^*)$$

Literature