Discriminant Analysis

Discriminination
and
Classification

2-group Discrimination

Consider 2 populations, labeled π_1 and π_2.

Assume that p-variate observations exist on each of the members of these 2 populations.

Let $\mu_1 = E(\boldsymbol{x} \mid \pi_1)$

$\mu_2 = E(\boldsymbol{x} \mid \pi_2)$

$\Sigma = Var(\boldsymbol{x})$

which is the same for π_1 and π_2

2-group Discrimination

Consider the linear combination of the x's

$$y = \mathbf{l}'\mathbf{x}$$

Then $\mu_{1y} = E(\mathbf{y} \mid \pi_1) = \mathbf{l}'\boldsymbol{\mu}_1$

$$\mu_{2y} = E(\mathbf{y} \mid \pi_2) = \mathbf{l}'\boldsymbol{\mu}_2$$

$$\sigma_y^2 = \mathbf{l}'\boldsymbol{\Sigma}\mathbf{l} = \Delta^2$$

2-group Discrimination

We wish to find an $\mathbf{l}$ such that

$$\frac{\left(\mu_{1y} - \mu_{2y}\right)^2}{\sigma_y^2} = \frac{\left(\mathbf{l}'\boldsymbol{\mu}_1 - \mathbf{l}'\boldsymbol{\mu}_2\right)^2}{\mathbf{l}'\boldsymbol{\Sigma}\mathbf{l}}$$

$$= \frac{\mathbf{l}'\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)'\mathbf{l}}{\mathbf{l}'\boldsymbol{\Sigma}\mathbf{l}}$$

$$= \frac{\left(\mathbf{l}'\boldsymbol{\delta}\right)^2}{\mathbf{l}'\boldsymbol{\Sigma}\mathbf{l}} \quad \text{is maximized.}$$

2-group Discrimination

where $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$

This is given by

$$\mathbf{l} = c\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = c\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

for any $c \neq 0$.

Fisher's Linear Discriminant Function

Choosing c = 1 gives
Fisher's Linear Discriminant Function:

$$y = \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)' \boldsymbol{\Sigma}^{-1}\mathbf{x} = \mathbf{b}'\mathbf{x}$$

Furthermore, the maximum is given by

$$\max_{\mathbf{l}} \frac{\left(\mathbf{l}'\boldsymbol{\delta}\right)^2}{\mathbf{l}'\boldsymbol{\Sigma}\mathbf{l}} = \boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta} = \Delta^2 = \sigma_y^2$$

Classification of a New Observation

Let $m = \frac{1}{2}\left(\mu_{1y} + \mu_{2y}\right)$

$$= \frac{1}{2}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2\right)$$

be the midpoint between the means of the two univariate populations

Classification of a New Observation

Consider an observation $\mathbf{x}_0$ taken from one of these two populations, with the precise membership of $\mathbf{x}_0$ unknown.

Let $y_0 = \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\boldsymbol{\Sigma}^{-1}\mathbf{x}_0$

Note that $\mathrm{E}\left(y_0 \mid \pi_1\right) - m \geq 0$

$$\mathrm{E}\left(y_0 \mid \pi_2\right) - m < 0$$

Classification of a New Observation

Thus, for classifying $\mathbf{x}_0$,

Allocate x_0 to π_1 if $y_0 \geq m$

Allocate x_0 to π_2 if $y_0 < m$

Classification using Sample Statistics

Consider N_i p-variate observations from population π_i, arranged in the data matrix

$$\mathbf{X}_i' = \begin{bmatrix} \mathbf{x}_{i1} & \mathbf{x}_{i2} & \cdots & \mathbf{x}_{iN_i} \end{bmatrix} \qquad \text{for } i = 1, 2$$

Let $\bar{\mathbf{x}}_i$ denote the sample mean vector,

$\mathbf{S}_i$ the (unbiased) sample covariance matrix.

Classification using Sample Statistics

Assume, for now, that the population covariance matrices are equal ($\Sigma_1 = \Sigma_2$), and let $\mathbf{S}_p$ denote the pooled unbiased estimate of the common Σ.

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S}_p = \frac{(N_1 - 1)\mathbf{S}_1 + (N_2 - 1)\mathbf{S}_2}{N_1 + N_2 - 2}$$

Classification using Sample Statistics

Fisher's sample linear discriminant function is given by:

$$y = \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)' \mathbf{S}_p^{-1} \mathbf{x}$$

Classification using Sample Statistics

Let $\hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$

$$= \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

denote the midpoint between the two univariate sample means

Classification using Sample Statistics

For classifying an unknown observation $\mathbf{x}_0$,

Allocate x_0 to π_1 if $y_0 \geq \hat{m}$ (or $y_0 - \hat{m} \geq 0$)

Allocate x_0 to π_2 if $y_0 < \hat{m}$ (or $y_0 - \hat{m} < 0$)

Classification using Sample Statistics

Note that in Fisher's linear discriminant function

$$y = \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)' \mathbf{S}_p^{-1} \mathbf{x}$$
$$= \hat{\mathbf{l}}' \mathbf{x}$$

$\hat{\mathbf{l}}$ maximizes the ratio

$$\frac{\left(\bar{y}_1 - \bar{y}_2\right)^2}{s_y^2} = \frac{\left(\hat{\mathbf{l}}'\mathbf{d}\right)^2}{\hat{\mathbf{l}}'\mathbf{S}_p\hat{\mathbf{l}}} \qquad \text{where } \mathbf{d} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$$

Classification using Sample Statistics

Furthermore, the maximum is given by

$$\max_{\hat{\mathbf{l}}} \frac{\left(\mathbf{l}'\mathbf{d}\right)^2}{\mathbf{l}'\mathbf{S}_p\mathbf{l}} = \mathbf{d}'\mathbf{S}_p^{-1}\mathbf{d} = D^2$$

where

$D =$ Mahalanobis distance between $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$

Classification Rule

A classification rule breaks the p-dimensional space into 2 mutually exclusive and exhaustive regions R_1 and R_2, where we classify $\mathbf{x}$ in π_i if $\mathbf{x} \in \mathrm{R}_i$

Probability of Classification Table

		Classify as π_1	Classify as π_2
True Population	π_1	$\mathrm{P}(\mathbf{x} \in \mathrm{R}_1 \mid \pi_1)$	$\mathrm{P}(\mathbf{x} \in \mathrm{R}_2 \mid \pi_1)$
	π_2	$\mathrm{P}(\mathbf{x} \in \mathrm{R}_1 \mid \pi_2)$	$\mathrm{P}(\mathbf{x} \in \mathrm{R}_2 \mid \pi_2)$

Total Probability of Misclassification

$\text{TPM} = P(\mathbf{x} \in R_1|\pi_2) + P(\mathbf{x} \in R_2|\pi_1)$

(assuming $P(\pi_1) = P(\pi_2) = ½$)

The optimum error rate (OER) is obtained by choosing R_1 and R_2 such that the TPM is minimized.

Computing the Probability of misclassifying an observation

Assume $\pi_i : N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$

Computing the Probability of misclassifying an observation

P[misclassifying a π_1 observation into π_2]

$$= \mathrm{P}[2|1]$$

$$= \mathrm{P}[y < m]$$

$$= \mathrm{P}\left[y < \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right]$$

$$= P\left[\frac{y - \mu_{1y}}{\sigma_y} < \frac{\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1}{\Delta} \right]$$

Computing the Probability of misclassifying an observation

P[misclassifying a π_1 observation into π_2]

$$= \mathrm{P}\left[z < \frac{-\frac{1}{2}\Delta^2}{\Delta} \right]$$

$$= \Phi\left(-\frac{1}{2}\Delta \right)$$

Computing the Probability of misclassifying an observation

Similarly, $P(1|2) = \Phi\left(-\frac{1}{2}\Delta\right)$

So, $OER = \frac{1}{2}P(2|1) + \frac{1}{2}P(1|2)$

$= \Phi\left(-\frac{1}{2}\Delta\right)$

Estimating the Probability of misclassifying an observation

Whenever the population parameters are unknown, the error rate must be estimated using other procedures

Estimating the Probability of misclassifying an observation

Method 1:

Use $\Phi\left(-\frac{1}{2}\mathrm{D}\right)$

where $\mathrm{D}^2 = \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)' \mathbf{S}_p^{-1} \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)$

Estimating the Probability of misclassifying an observation

Method 2: Resubstitution Estimate
or Apparent Error Rate (APER)

The Apparent Error Rate can be estimated from the following membership table:

Membership Table

		Predicted Membership π_1	Predicted Membership π_2	
Actual Membership	π_1	n_{1C}	n_{1M}	n_1
	π_2	n_{2M}	n_{2C}	n_2

$$\text{APER} = \frac{n_{1M} + n_{2M}}{n_1 + n_2}$$

It can be shown that APER underestimates the error rate.

Estimating the Probability of misclassifying an observation

Method 3: Cross-validation
or estimates from Hold-out Sample

Split sample into a training sample and a validation sample. Construct the classification function using the training sample. Evaluate the error rate using this classification function on the validation sample.

Estimating the Probability of misclassifying an observation

Method 3: Cross-validation
or estimates from Hold-out Sample

Problems:

1. Requires large samples.
2. The function evaluated is not the function of interest. The function of interest uses <u>all</u> the sample values

Estimating the Probability of misclassifying an observation

Method 4: Jack-knife procedure

a) Start with the n_1 observations in π_1. Omit one observation from this group and construct a classification function based on the remaining $(n_1\text{-}1)+n_2$ observations.
b) Classify the observation left out.

Estimating the Probability of misclassifying an observation

Method 4: Jack-knife procedure

c) Repeat steps (a) and (b) until all n_1 observations in π_1 are classified. Let $n_{1M}^{(H)}$ be the number of hold-out observations misclassified in π_1.

Estimating the Probability of misclassifying an observation

Method 4: Jack-knife procedure

d) Repeat steps (a), (b), and (c) for the n_2 observations in π_2. Let $n_{2M}^{(H)}$ denote the number of hold-out observations misclassified in π_2.

Estimating the Probability of misclassifying an observation

Method 4: Jack-knife procedure

Then the estimates of misclassification Probabilities are:

$$\hat{\mathrm{P}}(2\,|\,1) = \frac{n_{1M}^{(H)}}{n_1}$$

$$\hat{\mathrm{P}}(1\,|\,2) = \frac{n_{2M}^{(H)}}{n_2}$$

Estimating the Probability of misclassifying an observation

Method 4: Jack-knife procedure

Also, an estimate of the expected actual error rate is:

$$\frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2}$$

The General Classification Problem

Includes

- Prior probabilities associated with each population
- Costs of misclassification

The General Classification Problem

Let p_i be the prior probability of occurrence of π_i.

For example, π_1 might be much larger than π_2, and based on the size of the populations, obtaining an observation from π_1 is more likely a priori.

Probability of Classification Table with Priors

		Classify as π_1	Classify as π_2
True Pop^{n}	π_1	$P(1\|1)p_1$ $=P(\mathbf{x} \in R_1\|\pi_1)\ P(\pi_1)$	$P(2\|1)p_1$ $=P(\mathbf{x} \in R_2\|\pi_1)\ P(\pi_1)$
	π_2	$P(1\|2)p_2$ $=P(\mathbf{x} \in R_1\|\pi_2)\ P(\pi_2)$	$P(2\|2)p_2$ $=P(\mathbf{x} \in R_2\|\pi_2)\ P(\pi_2)$

Minimize Total Probability of Misclassification (TPM)

$$TPM = P(\mathbf{x} \in R_1|\pi_2)p_2 + P(\mathbf{x} \in R_2|\pi_1)p_1$$
$$= P(1|2)p_2 + P(2|1)p_1$$

Cost of Misclassification Table

		Classify as π_1	Classify as π_2
True Pop	π_1	0	$c(2\|1)$
	π_2	$c(1\|2)$	0

Minimize Expected Cost of Misclassification (ECM)

$$ECM = c(1|2)\, P(1|2)\, p_2 + c(2|1)\, P(2|1)\, p_1$$

Note: If $c(1|2) = c(2|1) = 1$ (equal costs) then ECM = TPM.
This would be the case if no costs were specified.

Regions that minimize ECM

$$R_1 : \frac{L(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma})}{L(\mathbf{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma})} \geq \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1}$$

$$R_2 : \frac{L(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma})}{L(\mathbf{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma})} < \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1}$$

Estimated minimum ECM rule

For classifying an unknown observation $\mathbf{x}_0$,

Allocate x_0 to π_1 if $\mathbf{d}'\mathbf{S}_p^{-1}\mathbf{x}_0 - \hat{m} \geq \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right]$

Allocate x_0 to π_2 if $\mathbf{d}'\mathbf{S}_p^{-1}\mathbf{x}_0 - \hat{m} < \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right]$

Expressing Prior Probabilities in such a way to include costs

- Most statistics computing packages allow the specification of prior probabilities, but do not have a way to enter costs.
- If this is the case, costs can be incorporated into the prior probabilities in the following way:

$$p_1^* = \frac{p_1 \, c(2|1)}{p_1 \, c(2|1) + p_2 \, c(1|2)} \qquad p_2^* = \frac{p_2 \, c(1|2)}{p_1 \, c(2|1) + p_2 \, c(1|2)}$$

Anderson's Classification Function

$$w = \mathbf{d}' \mathbf{S}_p^{-1} \mathbf{x}_0 - \hat{m}$$

Classification when $\Sigma_1 \neq \Sigma_2$ Quadratic Classification Rule

Assume $\pi_i : \mathrm{N}_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)$

The classification function d_i is:

$$d_i = -\frac{1}{2}\left(\mathbf{x} - \overline{\mathbf{x}}_i\right)' \mathbf{S}_i^{-1}\left(\mathbf{x} - \overline{\mathbf{x}}_i\right) - \frac{1}{2}\log\left|\mathbf{S}_i\right| + \log\left[p_i \, \mathrm{c}(j \mid i)\right]$$

Classification when $\Sigma_1 \neq \Sigma_2$ Quadratic Classification Rule

The (Bayes) allocation rule that minimizes ECM is:

Allocate x_0 to π_1 if $d_1 \geq d_2$

Allocate x_0 to π_2 if $d_1 < d_2$

Classification with 2 or more Populations

Let there be g populations:

$$\pi_1, \cdots, \pi_g \qquad g \geq 2$$

Let

p_i = prior probability of π_i

$$\mathrm{P}(j \mid i) = \mathrm{P}\left(\text{classifying } \mathbf{x} \text{ in } \pi_j \mid \mathbf{x} \text{ belongs in } \pi_i\right)$$

$$\mathrm{c}(j \mid i) = \mathrm{Cost}\left(\text{classifying } \mathbf{x} \text{ in } \pi_j \mid \mathbf{x} \text{ belongs in } \pi_i\right)$$

Classification with 2 or more Populations

The Total Probability of Misclassification is:

$$\mathrm{TPM} = \sum_{i=1}^{g} \sum_{\substack{j=1 \\ j \neq i}}^{g} \mathrm{P}(j \mid i)\ p_i$$

Classification with 2 or more Populations

Assume $\pi_i : \mathrm{N}_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right)$

(equal variance covariance matrices)

Let the classification function be:

$$d_i = \frac{1}{2}\left(\mathbf{x} - \overline{\mathbf{x}}_i\right)' \mathbf{S}_p^{-1}\left(\mathbf{x} - \overline{\mathbf{x}}_i\right) - \log\left(p_i\right)$$

Classification with 2 or more Populations

Then the allocation rule that minimizes TPM is:

Allocate $\mathbf{x}$ to π_j if $d_j = \max_i\left(d_i\right)$

Classification with 2 or more Populations

Incorporating costs of misclassification:

The Expected Cost of Misclassification is:

$$\mathrm{ECM} = \sum_{i=1}^{g} p_i \left(\sum_{\substack{j=1 \\ j \neq i}}^{g} \mathrm{P}(j \mid i)\, \mathrm{c}(j \mid i) \right)$$

Classification with 2 or more Populations

Incorporating costs of misclassification:

Let $\mathrm{L}\left(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right)$

be the multivariate normal density function associated with the i^{th} population.

Classification with 2 or more Populations

Incorporating costs of misclassification:

Then the allocation rule that minimizes ECM is:

Allocate $\mathbf{x}$ to π_j for which $\sum_{\substack{i=1 \\ i \neq j}}^{g} p_i \, \mathrm{L}\left(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right) \mathrm{c}(j \mid i)$

is the smallest.

Classification with 2 or more Populations

Now assume $\pi_i : \mathrm{N}_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)$

(<u>unequal</u> variance covariance matrices)

Let the (quadratic) classification function be:

$$d_i = \frac{1}{2}\ln\left|\mathbf{S}_i\right| + \frac{1}{2}\left(\mathbf{x} - \overline{\mathbf{x}}_i\right)' \mathbf{S}_i^{-1} \left(\mathbf{x} - \overline{\mathbf{x}}_i\right) - \log\left(p_i\right)$$

Classification with 2 or more Populations

Then the allocation rule that minimizes TPM is:

Allocate $\mathbf{x}$ to π_j if $d_j = \max_i \left(d_i \right)$

Variable Selection Procedures

- With each case measured on p variables, it is frequently the case that not all of the p variables are necessary for discrimination.
- Variable Selection Procedures commonly available in Discriminant Analysis include:
  - Forward Selection
  - Backward Elimination
  - Stepwise Procedure

Variable Selection Procedures

- Similar to their implementation in Multiple Regression, these stepwise procedures are based on partial F-tests.

Forward Selection

1. Determine that variable that produces the largest F value in a 1-way ANOVA. If its associated p-value < α, select that variable. Otherwise, stop.
2. Of the remaining variables, select the one that produces the largest F in a 1-way ANCOVA, using the first variable as the covariate. If its associated p-value < α, select that variable. Otherwise, stop.
3. Iterate until the algorithm stops.

Backward Elimination

1. Determine that variable that produces the smallest F value in a 1-way ANCOVA with the other p-1 variables as covariates. If the associated p-value > α, eliminate that variable. Otherwise, stop.
2. Iterate, removing one variable at a time, until the algorithm stops.

Stepwise Selection

- Forward Selection with Backward Elimination also done after each step.

Choice of α in Stepwise Procedures

SAS Procedure	**SLEntry**	**SLStay**
STEPDISC	0.15	0.15
LOGISTIC	0.05	0.05
REG	0.50 FORWARD 0.15 STEPWISE	0.10 BACKWARD 0.15 STEPWISE
Recommended	0.20 FORWARD 0.15 STEPWISE	0.01 BACKWARD 0.15 STEPWISE

Stepwise Algorithms Other Points:

- For STEPWISE, SLE $\leq$ SLS
- The final set of variables selected by the algorithm is generally NOT the best final set.
  - Some additional fine-tuning is generally necessary, removing one or two variables that have a large p-value.

Stepwise Variable Selection
Wilk's Λ

- Wilks' Λ is the likelihood ratio statistic for testing the hypothesis that the means of the classes on the selected variables are equal in the population.
- Wilks' Λ is close to zero if any two groups are well separated.

Stepwise Variable Selection
Average Squared Canonical Correlation

- The ASCC is Pillai's trace divided by the number of groups minus 1.
- The ASCC is close to 1 if all groups are well separated and if all or most directions in the discriminant space show good separation for at least two groups.
- If ASCC is small, but the associated p-value < α, then the groups means are significantly different, but the groups are not well separated (i.e., considerable overlap of individual points).

Canonical Discriminant Analysis

Discrimination among several populations

Assume $\pi_i : (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \qquad i = 1, \cdots, g$

Also, assume we have N_i p-variate observations from population π_i with x_{ij} = j^{th} observation from population i

Canonical Discriminant Analysis

Let

$$\overline{\mathbf{x}}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}$$

$$\overline{\mathbf{x}} = \frac{\sum_{i=1}^{g} N_i \overline{\mathbf{x}}_i}{\sum_{i=1}^{g} N_i} = \frac{\sum_{i=1}^{g} N_i \overline{\mathbf{x}}_i}{N}$$

Canonical Discriminant Analysis

Let

$$\mathrm{B} = \sum_{i=1}^{g} \left(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}\right)\left(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}\right)'$$

$$\mathrm{W} = \sum_{i=1}^{g} \left(N_i - 1\right)\mathbf{S}_i = \left(N - 1\right)\mathbf{S}_p$$

Canonical Discriminant Analysis

Consider the eigenvalues

$\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_s > 0$

of BW^{-1} or $\mathrm{B}(\mathrm{B}+\mathrm{W})^{-1}$

with corresponding eigenvectors

$\hat{\mathbf{l}}_1 \geq \cdots \geq \hat{\mathbf{l}}_s$

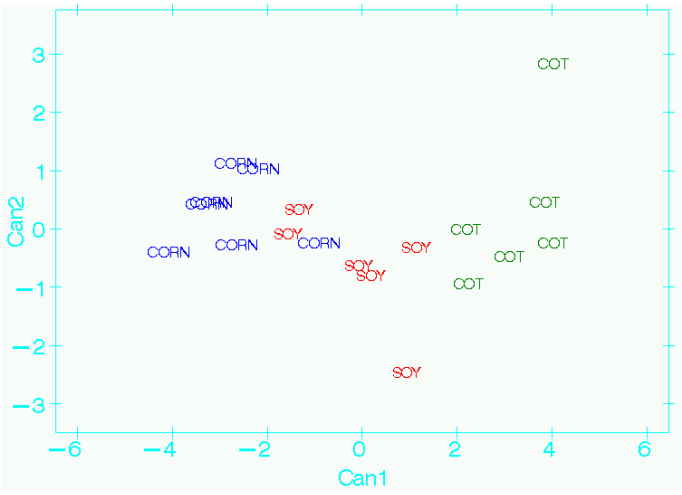
scaled so that

$\hat{\mathbf{l}}_i' \, \mathbf{S}_p \, \hat{\mathbf{l}}_i = 1$

Canonical Discriminant Analysis

$\hat{\mathbf{l}}_i' \mathbf{x}$ is called the i^{th} sample discriminant.

How Does Canonical Discriminant Analysis Work?



Overall Test of Significance

The overall test that the discriminant functions successfully detect separation of the populations is equivalent to the test that the mean vectors in the various populations are different:

$$\mathrm{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g$$

$$\mathrm{H}_1 : \sim \mathrm{H}_0$$

Overall Test of Significance

A test statistic for testing that the discriminant functions as a whole significantly separate groups is Wilk's Λ:

$$\Lambda_1 = \prod_{i=1}^{s} \frac{1}{1+\lambda_i} \sim \Lambda_{p,g-1,N-g}$$

$$N = \sum_{i=1}^{g} N_i$$

Overall Test of Significance

An F approximation to Λ_1 is given by:

$$F = \frac{1-\Lambda_1^{1/t}}{\Lambda_1^{1/t}} \frac{\nu_2}{\nu_1} \sim F_{\nu_1,\nu_2}$$

$$t = \sqrt{\frac{p^2(g-1)^2-4}{p^2+(g-1)^2-5}} \qquad \nu_1 = p(g-1)$$

$$\nu_2 = wt - \frac{1}{2}\left[p(g-1)-2\right]$$

$$w = N - 1 - \frac{1}{2}(p+g)$$

Determining the number of Canonical Discriminant Functions Needed

To test the significance of λ_2, ⋯ ,λ_s , we delete λ_1 from Wilk's Λ, to obtain:

$$\Lambda_2 = \prod_{i=2}^{s} \frac{1}{1+\lambda_i} \sim \Lambda_{p-1,g-2,N-g-1}$$

Determining the number of Canonical Discriminant Functions Needed

In general, to test the significance of λ_m, ⋯ ,λ_s , we delete λ_1, ⋯ ,λ_{m-1} from Wilk's Λ, to obtain:

$$\Lambda_2 = \prod_{i=m}^{s} \frac{1}{1+\lambda_i} \sim \Lambda_{p-m+1,\,g-m,\,N-g-m+1}$$

Determining the number of Canonical Discriminant Functions Needed

An F approximation to Λ_m is given by:

$$\mathrm{F} = \frac{1-\Lambda_m^{1/t}}{\Lambda_m^{1/t}} \frac{\nu_2}{\nu_1} \sim \mathrm{F}_{\nu_1,\nu_2}$$

$$t = \sqrt{\frac{(p-m+1)^2(g-m)^2-4}{(p-m+1)^2+(g-m)^2-5}}$$

$$\nu_1 = (p-m+1)(g-m) \qquad w = N-1-\frac{1}{2}(p+g)$$

$$\nu_2 = wt - \frac{1}{2}\left[(p-m+1)(g-m)-2\right]$$

Using the Sample Discriminants to Classify

There are $s = \min(g\text{-}1, p)$ discriminants.

Let $y_j = \mathbf{l}_j' \mathbf{x} = j^{th}$ discriminant $\quad j = 1, \cdots, s$

Let $\mathbf{y}' = \begin{pmatrix} y_1 & y_2 & \cdots & y_s \end{pmatrix}$

Then $\mathrm{E}(\mathbf{y}' \mid \pi_i) = \begin{pmatrix} \mathbf{l}_1' \boldsymbol{\mu}_i & \cdots & \mathbf{l}_s' \boldsymbol{\mu}_i \end{pmatrix} = \boldsymbol{\mu}_{iy}'$

$\mathrm{Var}(\mathbf{y} \mid \pi_i) = \mathbf{I}$

Using the Sample Discriminants to Classify

The squared distance of $\mathbf{y}$ from $\boldsymbol{\mu}_{iy}$ is

$$(\mathbf{y} - \boldsymbol{\mu}_{iy})' (\mathbf{y} - \boldsymbol{\mu}_{iy}) = \sum_{j=1}^{s} (y_j - \mu_{iy_j})^2$$

Canonical Classification Rule

A reasonable classification rule would be:

Assign $\mathbf{y}$ to π_j if the squared distance from $\mathbf{y}$ to $\boldsymbol{\mu}_{jy}$ is smaller than the squared distance from $\mathbf{y}$ to $\boldsymbol{\mu}_{iy}$ for $i \neq j$.

Canonical Classification Rule

If only $r \leq s$ of the discriminants are used, then, using the sample statistics to estimate the population parameters, the allocation rule is:

Canonical Classification Rule

Allocate $\mathbf{x}$ to π_k if

$$\sum_{j=1}^{r}\left(\hat{y}_j - \overline{y}_{kj}\right)^2 = \sum_{j=1}^{r}\left[\hat{\mathbf{l}}_j'\left(\mathbf{x} - \overline{\mathbf{x}}_k\right)\right]^2$$

$$\leq \sum_{j=1}^{r}\left[\hat{\mathbf{l}}_j'\left(\mathbf{x} - \overline{\mathbf{x}}_i\right)\right]^2 \qquad \text{for } i \neq k$$

Rule Equivalence

When $p_1 = \cdots = p_g = \dfrac{1}{g}$ and $r = s$,

these rules are equivalent to the previously stated rules for minimizing the Total Probability of Misclassification (TPM).

Nearest Neighbor Discriminant Analysis

- Examine the k nearest neighbors (points) to $\mathbf{x}$.
- Assign $\mathbf{x}$ to the group to which the majority of these k points belong.

Example with 2 Groups

- Set k
- Of these k points, let k_1 belong to Group 1 (with sample size N_1) and k_2 belong to Group 2 (with sample size N_2).
- Let $k_1 + k_2 = k$.
- If $N_1 = N_2$, assign $\mathbf{x}$ to Group 1 if $k_1 > k_2$.

Example with 2 Groups

If $N_1 = N_2$, assign $\mathbf{x}$ to Group 1 if $k_1 > k_2$.

If $N_1 \neq N_2$, an alternative rule could be:

Assign $\mathbf{x}$ to Group 1 if $\frac{k_1}{N_1} > \frac{k_2}{N_2}$

Example with 2 Groups

If prior probabilities are not equal, a corresponding rule could be:

Assign $\mathbf{x}$ to Group 1 if $\frac{k_1/N_1}{k_2/N_2} > \frac{p_1}{p_2}$

Choice of k

1. $k = \sqrt{N_i}$
2. Select k that minimizes the error rate.

Classification based on Density Estimators

Parametric Classification assumes

$$\mathbf{x} \sim \mathrm{N}_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right)$$

and assigns x to the group for which

$p_i \; f\left(\mathbf{x} \mid \pi_i\right)$ is a maximum

where $f\left(\mathbf{x} \mid \pi_i\right)$ is the probability density function of $\mathrm{N}_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right)$

Classification based on Density Estimators

If $f(\mathbf{x} \mid \pi_i)$ is nonnormal or unknown,

the pdf can be estimated directly from the data.

This is known as kernel estimation.

The DISCRIM Procedure

PROC DISCRIM can be used for many different types of analysis including:

- canonical discriminant analysis
- assessing and confirming the usefulness of the functions (empirical validation and crossvalidation)
- predicting group membership on new data using the functions (scoring)
- linear and quadratic discriminant analysis
- nonparametric discriminant analysis