Selecting Variables in Two-Group Robust Linear Discriminant Analysis

Stefan Van Aelst and Gert Willems

Department of Applied Mathematics and Computer Science
Ghent University, Belgium

COMPSTAT'2010

Linear discriminant analysis setting

- p-dimensional data set
- Group 1: $\mathbf{x}_{11} \ldots, \mathbf{x}_{1n_1} \in \Pi_1 \sim F_1 = F_{\boldsymbol{\mu}_1, \Sigma}$
- Group 2: $\mathbf{x}_{21} \ldots, \mathbf{x}_{2n_2} \in \Pi_2 \sim F_2 = F_{\boldsymbol{\mu}_2, \Sigma}$
- Common covariance matrix Σ
- $P(X \in \Pi_1) = P(X \in \Pi_2)$
- $d_j^L(\mathbf{x}) = \boldsymbol{\mu}_j^t \Sigma^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_j^t \Sigma^{-1} \boldsymbol{\mu}_j;\ j = 1, 2$

Linear Bayes rule: Classify $\mathbf{x} \in \mathbb{R}^p$ into Π_1 if

$$d_1^L(\mathbf{x}) > d_2^L(\mathbf{x})$$

and into Π_2 otherwise.

Linear discriminant analysis setting

- p-dimensional data set
- Group 1: $\mathbf{x}_{11} \ldots, \mathbf{x}_{1n_1} \in \Pi_1 \sim F_1 = F_{\boldsymbol{\mu}_1,\Sigma}$
- Group 2: $\mathbf{x}_{21} \ldots, \mathbf{x}_{2n_2} \in \Pi_2 \sim F_2 = F_{\boldsymbol{\mu}_2,\Sigma}$
- Common covariance matrix Σ
- $P(X \in \Pi_1) = P(X \in \Pi_2)$
- $d_j^L(\mathbf{x}) = \boldsymbol{\mu}_j^t \Sigma^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_j^t \Sigma^{-1}\boldsymbol{\mu}_j;\ j = 1, 2$

Linear Bayes rule: Classify $\mathbf{x} \in \mathbb{R}^p$ into Π_1 if

$$d_1^L(\mathbf{x}) > d_2^L(\mathbf{x})$$

and into Π_2 otherwise.

Discriminant coordinate

Direction $\mathbf{a}$ that best separates the two populations:

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

The projection $\mathbf{a}^t\mathbf{x}$ is called the canonical variate or discriminant coordinate

Sample LDA

- Estimate the centers μ_1 and μ_2 and the scatter Σ from the data
- Standard LDA uses the sample means $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$, and the pooled sample covariance matrix

$$S_n = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

Robust LDA

- Use robust estimators of the centers μ_1 and μ_2 and the common scatter Σ

  $\longrightarrow$ S-estimators
  $\longrightarrow$ MM-estimators

One-sample S-estimators

- Observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^p$
- $\rho_0 : [0, \infty[\to [0, \infty[$ is bounded, increasing and smooth

S-estimates of the location $\widetilde{\boldsymbol{\mu}}_n$ and scatter $\widetilde{\Sigma}_n$ minimize $|C|$ subject to

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0 \left(\left[(\mathbf{x}_i - T)^t C^{-1} (\mathbf{x}_i - T) \right]^{\frac{1}{2}} \right) = b$$

among all $T \in \mathbb{R}^p$ and $C \in \mathrm{PDS}(p)$

(Davies 1987, Rousseeuw and Leroy 1987, Lopuhaä 1989)

ρ functions

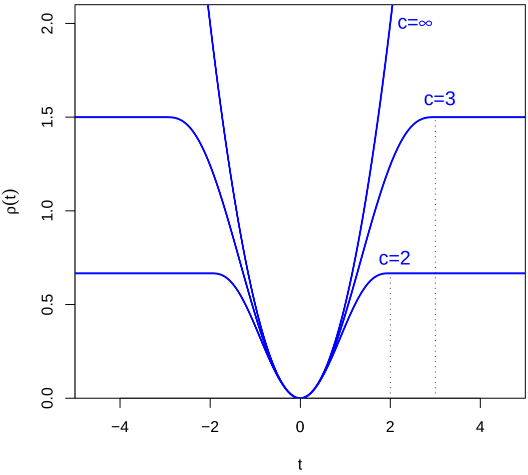
A popular family of loss functions is the Tukey biweight (bisquare) family of ρ functions:

$$\rho_c(t) = \begin{cases} \frac{t^2}{2} - \frac{t^4}{2c^2} + \frac{t^6}{6c^4} & \text{if } |t| \leq c \\ \frac{c^2}{6} & \text{if } |t| \geq c. \end{cases}$$

- The constant c can be tuned for robustness (breakdown point)
- The choice of c also determines the efficiency of the S-estimator

$\rightarrow$ Trade-off robustness vs efficiency

Tukey biweight ρ functions



One-sample MM-estimates

Put $\tilde{\sigma}_n = \det(\widetilde{\Sigma}_n)^{1/2p}$, the S-estimate of scale

Then the MM-estimates of the location $\widehat{\mu}_n$ and shape $\widehat{\Gamma}_n$ minimize

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1 \left(\left[(\mathbf{x}_i - T)^t G^{-1} (\mathbf{x}_i - T) \right]^{\frac{1}{2}} / \tilde{\sigma}_n \right)$$

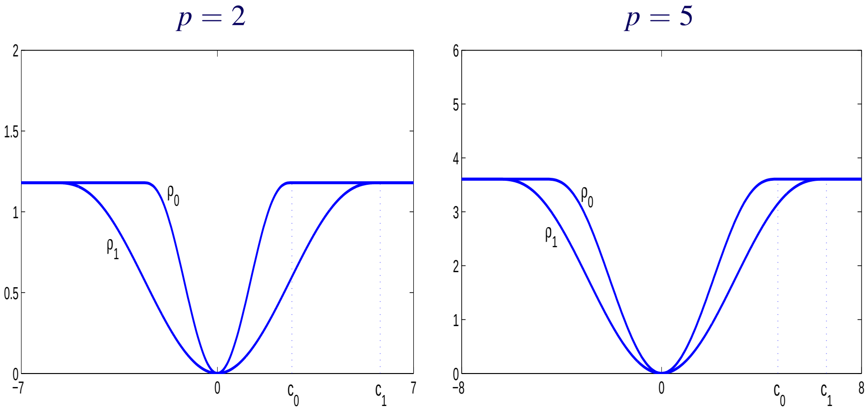
among all $T \in \mathbb{R}^p$ and $G \in \mathrm{PDS}(p)$ for which det(G)=1

(Tatsuoka and Tyler 2000)

ρ functions

- Both ρ_0 and ρ_1 are taken from the same family
- The constant c in ρ_0 can be tuned for robustness (breakdown point)
- MM-estimator inherits its robustness from the S-scale
- The constant c in ρ_1 can be tuned for efficiency of locations

Tukey biweight ρ functions



Robust two-sample estimates

- Pool the scatter estimates $\widehat{\Sigma}_{1n_1}$ and $\widehat{\Sigma}_{2n_2}$ of both groups:

$$\widehat{\Sigma}_n = \frac{n_1\widehat{\Sigma}_{1n_1} + n_2\widehat{\Sigma}_{2n_2}}{n_1 + n_2}$$

- Calculate simultaneous S-estimates of the two locations and the common scatter matrix:

$\widehat{\boldsymbol{\mu}}_{1n}$, $\widehat{\boldsymbol{\mu}}_{2n}$ and $\widehat{\Sigma}_n$ minimize $|C|$ subject to

$$\frac{1}{n_1+n_2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}\rho_0\left([(\mathbf{x}_{ji}-T_j)^t C^{-1}(\mathbf{x}_{ji}-T_j)]^{\frac{1}{2}}\right) = b$$

among all $T_1, T_2 \in \mathbb{R}^p$ and $C \in \mathrm{PDS}(p)$

(He and Fung 2000)
Similarly, simultaneous MM-estimates can be calculated

Bootstrap inference

- Advantages of bootstrap
  - Few assumptions
  - Wide range of applications
- Bootstrapping robust estimators
  - High computational cost
  - Robustness not guaranteed

Bootstrap inference

- Advantages of bootstrap
  - Few assumptions
  - Wide range of applications
- Bootstrapping robust estimators
  - High computational cost
  - Robustness not guaranteed

Fast and robust bootstrap principle

For each bootstrap sample

- Calculate an approximation for the estimates
- Use the estimating equations
- Fast to compute approximations
- Inherit robustness of initial solution

Fast and robust bootstrap

- Consider estimates that are the solution of a fixed point equation $\widehat{\Theta}_n = \mathbf{g}_n(\widehat{\Theta}_n)$
- For a bootstrap sample $\widehat{\Theta}_n^* = \mathbf{g}_n^*(\widehat{\Theta}_n^*)$ consider the one-step approximation

$$\widehat{\Theta}_n^{1\star} = \mathbf{g}_n^*(\widehat{\Theta}_n)$$

- Take a Taylor expansion about estimands Θ:

$$\widehat{\Theta}_n = \mathbf{g}_n(\Theta) + \nabla \mathbf{g}_n(\Theta)(\widehat{\Theta}_n - \Theta) + O_P(n^{-1})$$

which can be rewritten as:

$$\sqrt{n}(\widehat{\Theta}_n - \Theta) = [\mathbf{I} - \nabla \mathbf{g}_n(\Theta)]^{-1} \sqrt{n}(\mathbf{g}_n(\Theta) - \Theta) + O_P(n^{-1/2})$$

- We then obtain

$$\sqrt{n}(\widehat{\Theta}_n^* - \widehat{\Theta}_n) = [\mathbf{I} - \nabla \mathbf{g}_n(\widehat{\Theta}_n)]^{-1} \sqrt{n}(\mathbf{g}_n^*(\widehat{\Theta}_n) - \widehat{\Theta}_n) + O_P(n^{-1/2})$$

which yields the FRB estimate

$$\widehat{\Theta}_n^{R\star} = \widehat{\Theta}_n + [\mathbf{I} - \nabla \mathbf{g}_n(\widehat{\Theta}_n)]^{-1}(\widehat{\Theta}_n^{1\star} - \widehat{\Theta}_n)$$

Fast and robust bootstrap

- Consider estimates that are the solution of a fixed point equation $\widehat{\Theta}_n = \mathbf{g}_n(\widehat{\Theta}_n)$
- For a bootstrap sample $\widehat{\Theta}_n^* = \mathbf{g}_n^*(\widehat{\Theta}_n^*)$ consider the one-step approximation

$$\widehat{\Theta}_n^{1\star} = \mathbf{g}_n^*(\widehat{\Theta}_n)$$

- Take a Taylor expansion about estimands Θ:

$$\widehat{\Theta}_n = \mathbf{g}_n(\Theta) + \nabla\mathbf{g}_n(\Theta)(\widehat{\Theta}_n - \Theta) + O_P(n^{-1})$$

which can be rewritten as:

$$\sqrt{n}(\widehat{\Theta}_n - \Theta) = [\mathbf{I} - \nabla\mathbf{g}_n(\Theta)]^{-1}\sqrt{n}(\mathbf{g}_n(\Theta) - \Theta) + O_P(n^{-1/2})$$

- We then obtain

$$\sqrt{n}(\widehat{\Theta}_n^* - \widehat{\Theta}_n) = [\mathbf{I} - \nabla\mathbf{g}_n(\widehat{\Theta}_n)]^{-1}\sqrt{n}(\mathbf{g}_n^*(\widehat{\Theta}_n) - \widehat{\Theta}_n) + O_P(n^{-1/2})$$

which yields the FRB estimate

$$\widehat{\Theta}_n^{R\star} = \widehat{\Theta}_n + [\mathbf{I} - \nabla\mathbf{g}_n(\widehat{\Theta}_n)]^{-1}(\widehat{\Theta}_n^{1\star} - \widehat{\Theta}_n)$$

Properties of fast robust bootstrap

Computational efficiency: The FRB estimates are solutions of a system of linear equations

Robustness: The FRB estimates use the weights of the MM-estimates at the original sample

Consistency: Under regularity conditions, the FRB distribution of $\widehat{\Theta}_n$ and the sample distribution of $\widehat{\Theta}_n$ converge to the same limiting distribution

Smooth mappings: FRB commutes with smooth functions, such as $\mathbf{a} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

Properties of fast robust bootstrap

Computational efficiency: The FRB estimates are solutions of a system of linear equations

Robustness: The FRB estimates use the weights of the MM-estimates at the original sample

Consistency: Under regularity conditions, the FRB distribution of $\widehat{\Theta}_n$ and the sample distribution of $\widehat{\Theta}_n$ converge to the same limiting distribution

Smooth mappings: FRB commutes with smooth functions, such as $\mathbf{a} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

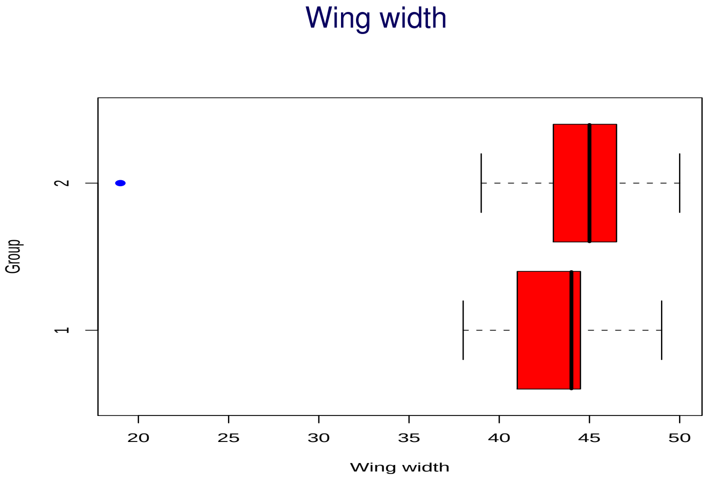
Variable selection in robust LDA

- Two group robust LDA
- Selection criterion: test for significance of the discriminant coordinate coefficients
- Use FRB distribution to estimate p-values

Example: Biting Flies

- Two groups of 35 flies (Leptoconops torrens and Leptoconops carteri)
- Measurements of
  - wing length
  - wing width
  - third palp length
  - third palp width
  - fourth palp length

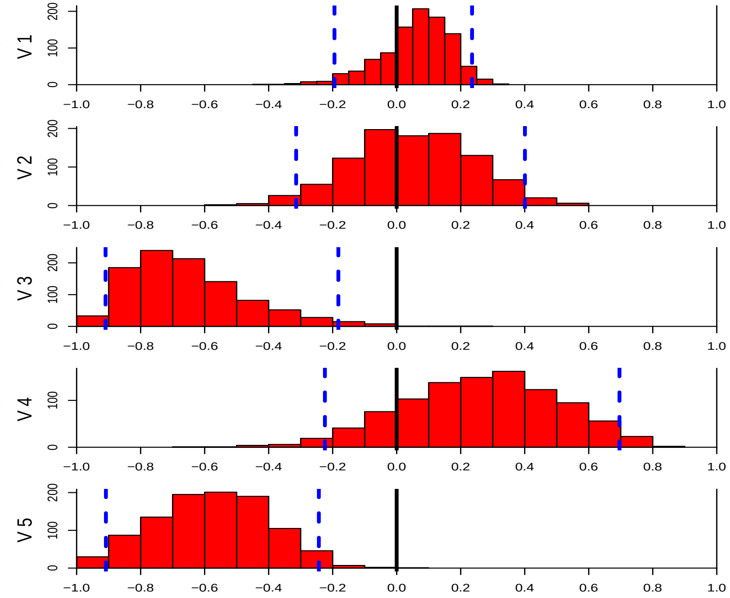
Biting Flies: outliers



Biting Flies: LDA

- Robust LDA
- Simultaneous two-sample MM-estimates
- Backward elimination variable selection

Biting Flies: FRB



Biting Flies: Backward elimination

	Variable				
Model	1	2	3	4	5
1	0.490	0.817	0.006	0.296	0.002
2	0.306	-	0.016	0.216	0.000
3	-	-	0.016	0.096	0.000
4	-	-	0.006	-	0.000

Conclusions and outlook

UNIVERSITEIT GENT

- Robust LDA based on S/MM-estimators
- Inference based on fast robust bootstrap
- Simulations confirm its good performance
- Variable selection based on contributions to discriminant coordinate
- More than two groups: Use a robust likelihood ratio type test statistics as selection criterion

Robust likelihood ratio type test statistics

UNIVERSITEIT GENT

$$\Lambda_n^R = \frac{|\widetilde{\Sigma}_n^{(g)}|}{|\widetilde{\Sigma}_n^{(1)}|} \equiv \frac{\tilde{\sigma}_n^{(g)}}{\tilde{\sigma}_n^{(1)}} = \frac{S_n(\widetilde{\boldsymbol{\mu}}_{1,n}^{(g)}, \ldots, \widetilde{\boldsymbol{\mu}}_{g,n}^{(g)}, \tilde{\Gamma}_n^{(g)})}{S_n(\widetilde{\boldsymbol{\mu}}_n^{(1)}, \tilde{\Gamma}_n^{(1)})}$$

$$\Lambda_n^R = \frac{\sum_{j=1}^{g}\sum_{i=1}^{n_j} \rho_0([(\mathbf{x}_{ji} - \widetilde{\boldsymbol{\mu}}_{j,n}^{(g)})^t (\tilde{\Gamma}_n^{(g)})^{-1} (\mathbf{x}_{ji} - \widetilde{\boldsymbol{\mu}}_{j,n}^{(g)})]^{\frac{1}{2}} / \tilde{\sigma}_n^{(g)})}{\sum_{j=1}^{g}\sum_{i=1}^{n_j} \rho_0([(\mathbf{x}_{ji} - \widetilde{\boldsymbol{\mu}}_n^{(1)})^t (\tilde{\Gamma}_n^{(1)})^{-1} (\mathbf{x}_{ji} - \widetilde{\boldsymbol{\mu}}_n^{(1)})]^{\frac{1}{2}} / \tilde{\sigma}_n^{(g)})}$$

$$\Lambda_n^R = \frac{\sum_{j=1}^{g}\sum_{i=1}^{n_j} \rho_0([(\mathbf{x}_{ji} - \widetilde{\boldsymbol{\mu}}_{j,n}^{(g)})^t (\widetilde{\Sigma}_n^{(g)})^{-1} (\mathbf{x}_{ji} - \widetilde{\boldsymbol{\mu}}_{j,n}^{(g)})]^{\frac{1}{2}})}{\sum_{j=1}^{g}\sum_{i=1}^{n_j} \rho_0([(\mathbf{x}_{ji} - \widetilde{\boldsymbol{\mu}}_n^{(1)})^t (\widetilde{\Sigma}_n^{(g)})^{-1} (\mathbf{x}_{ji} - \widetilde{\boldsymbol{\mu}}_n^{(1)})]^{\frac{1}{2}})}$$

References

- He, X. and Fung, W.K. (2000).
  High breakdown estimation for multiple populations with applications to discriminant analysis.
  Journal of Multivariate Analysis, 72, 151–162.
- Lopuhaä, H. (1989).
  On the relation between S-estimators and M-estimators of multivariate location and covariance.
  The Annals of Statistics, 17, 1662-1683.
- Salibian-Barrera, M., Van Aelst, S., and Willems, G. (2006).
  PCA based on multivariate MM-estimators with fast and robust bootstrap.
  Journal of the American Statistical Association, 101, 1198–1211.
- Tatsuoka, K.S. and Tyler, D.E. (2000).
  The uniqueness of S and M-functionals under non-elliptical distributions.
  The Annals of Statistics, 28, 1219–1243.
- Van Aelst, S. and Willems, G. (2010).
  Inference for robust canonical variate analysis.
  Advances in Data Analysis and Classification, to appear.