

Improving supervised classification for high dimensional data by adding external information

Maëla Kloareg David Causeur

*Laboratoire de Mathématiques Appliquées
Agrocampus Rennes
Université Européenne de Bretagne*

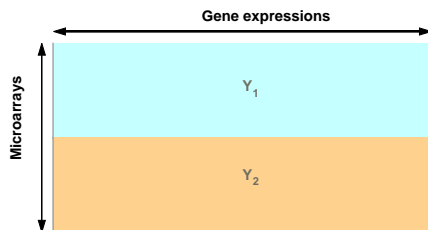
SSC-SFDS June 2008, Ottawa

Outline

- 1 Stability of model selection
- 2 Factor-adjusted procedures
- 3 External information
- 4 Concluding comments

High dimensional data

A few microarrays and a huge number of gene expressions



Differential analysis: to point out genes which mean levels differ from a group to another

Supervised classification: to affect biological samples into pre-defined groups

A major concern: Impact of dependence in multiple testing and on the stability of model selection in supervised classification

Stability of model selection

Model for supervised classification among normal populations

$$\log \left[\frac{\mathbb{P}_y(\text{Group} = 1)}{\mathbb{P}_y(\text{Group} = 2)} \right] = \beta_0 + y' \beta$$

Bayes linear classifier: $\beta^* = \Sigma^{-1}(\mu_1 - \mu_2)$

Minimizes the probability of misclassification

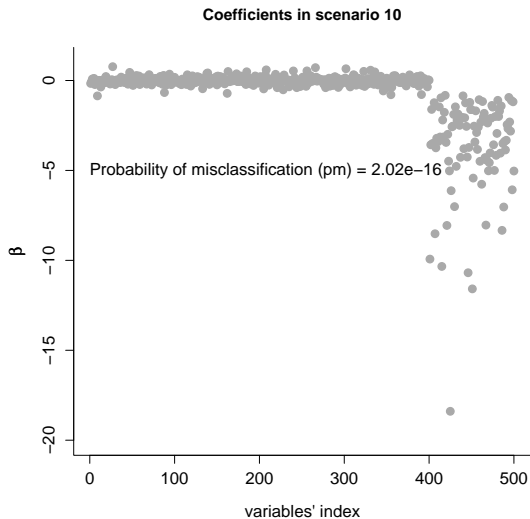
Estimation for high-dimension data:

- Regularized estimation: LASSO, SVM, ...
- **Stepwise selection**, can be implemented only forward

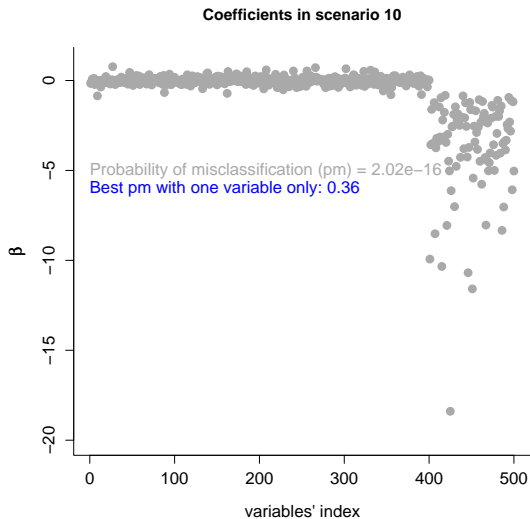
Illustration on simulated datasets

- Dimension: $m = 500$ $[m_0 = 400, m_1 = 100]$
- $Y \sim \mathcal{N}(\mu_i, \Sigma)$, $i = 1, 2$

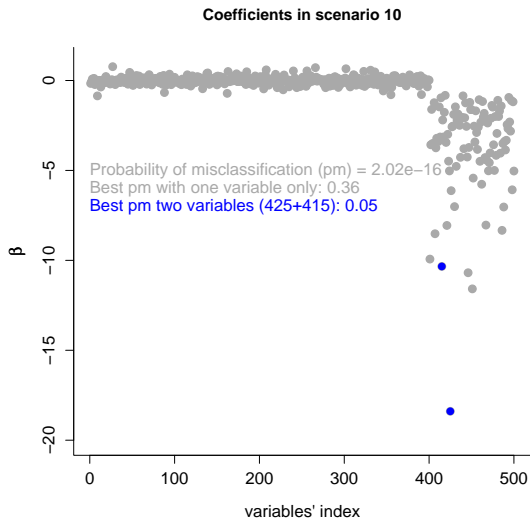
Stability of model selection



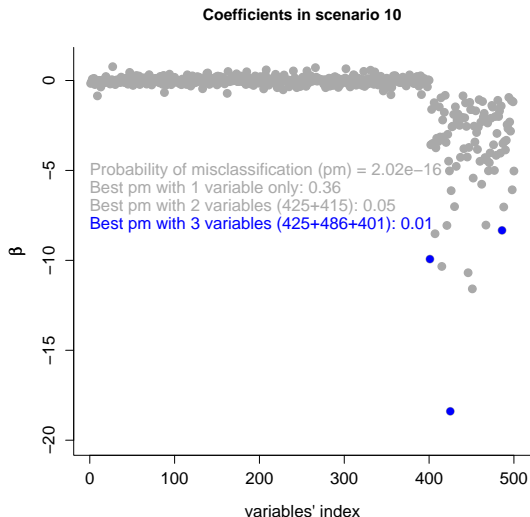
Stability of model selection



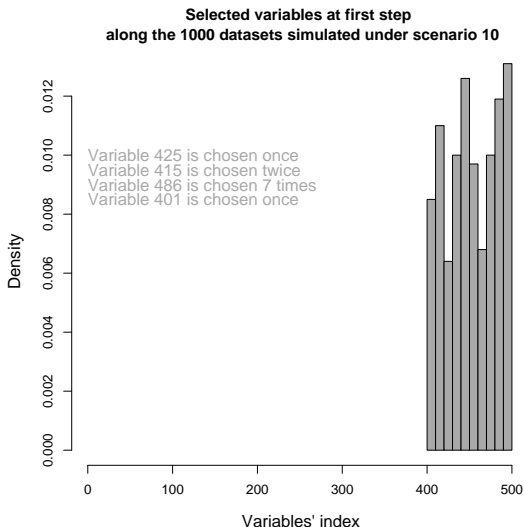
Stability of model selection



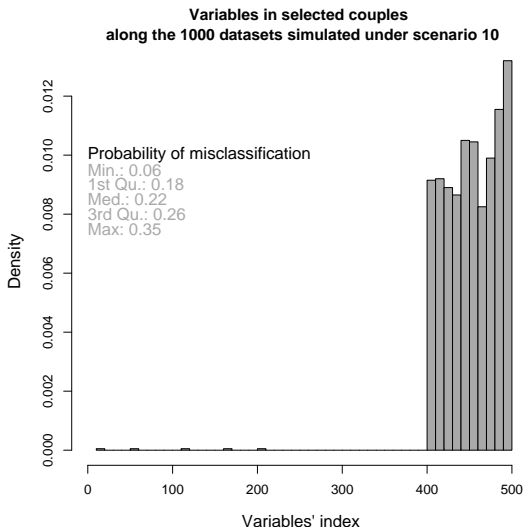
Stability of model selection



Stability of model selection

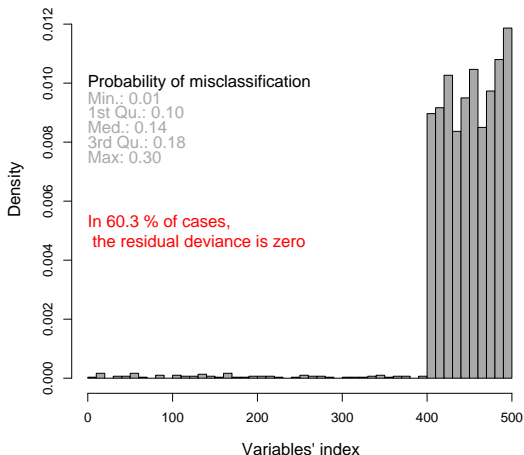


Stability of model selection



Stability of model selection

**Variables in selected triplets
along the 1000 datasets simulated under scenario 10**



Outline

- 1 Stability of model selection
- 2 Factor-adjusted procedures**
- 3 External information
- 4 Concluding comments

Factor analysis model

Linear model: $Y = (Y^{(1)}, Y^{(2)}, \dots, Y^{(m)})'$ with conditional distribution w.r.t x is normal with:

$$\mathbb{E}_x(Y) = (\beta_0^{(k)} + x' \beta^{(k)})_{k=1 \dots m}, \quad \text{Var}_x(Y) = \Sigma$$

Factor analysis model

Linear model: $Y = (Y^{(1)}, Y^{(2)}, \dots, Y^{(m)})'$ with conditional distribution w.r.t x is normal with:

$$\mathbb{E}_x(Y) = (\beta_0^{(k)} + x' \beta^{(k)})_{k=1 \dots m}, \quad \text{Var}_x(Y) = \Sigma$$

A factor structure for Σ : $\Sigma = \Psi + BB'$

Factor analysis model

Linear model: $Y = (Y^{(1)}, Y^{(2)}, \dots, Y^{(m)})'$ with conditional distribution w.r.t x is normal with:

$$\mathbb{E}_x(Y) = (\beta_0^{(k)} + x' \beta^{(k)})_{k=1 \dots m}, \quad \text{Var}_x(Y) = \Sigma$$

A factor structure for Σ : $\Sigma = \Psi + BB'$

Tests for linear contrasts:

$$T^{(k)} = \sqrt{n} \frac{\lambda' \hat{\beta}^{(k)}}{\sigma_k \sqrt{\lambda' S_{xx}^{-1} \lambda}}$$

with corresponding p-values $P^{(k)}$

Factor-adjusted test statistics

Conditional distribution of $T^{(k)}$

$$\mathbb{E}(T^{(k)} \mid Z) = \tau_k + \frac{b'_k}{\sigma_k} \tau(Z), \quad \text{Var}(T^{(k)} \mid Z) = \frac{\psi_k^2}{\sigma_k^2}.$$

Conditional centering and scaling

$$T_Z^{(k)} = \frac{\sigma_k}{\psi_k} \left[T^{(k)} - \frac{b'_k}{\sigma_k} \tau(Z) \right].$$

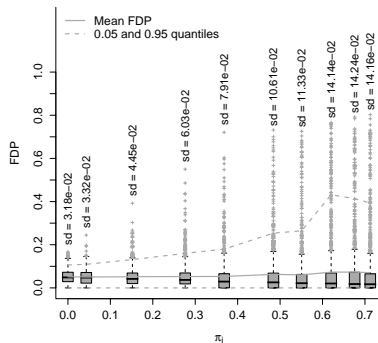
with $\mathbb{E}(T_Z^{(k)}) = \frac{\tau_k}{\sqrt{1-h_k^2}}$ and $\text{Var}(T_Z) = I_m$.

Estimation of the factor analysis model in high dimension : EM factor analysis (Rubin & Thayer, 1982)

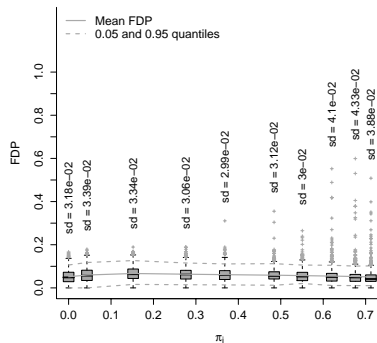
Distribution of error rates in multiple tests

Distribution of **False Discovery Proportion** on 1.000 simulated datasets/scenario (Friguet *et al.*, 2008, *submitted*)

Usual t-tests



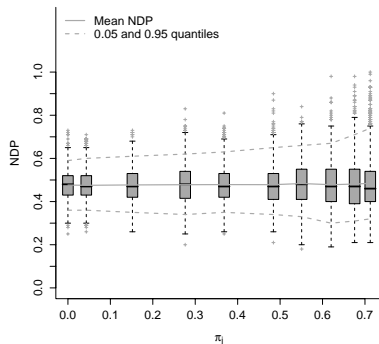
Factor-adjusted t-tests



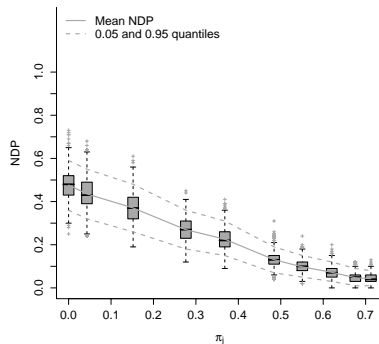
Distribution of error rates in multiple tests

Distribution of **Non-Discovery Proportion** on 1.000 simulated datasets/scenario (Friguet *et al.*, 2008, *submitted*)

Usual t-tests

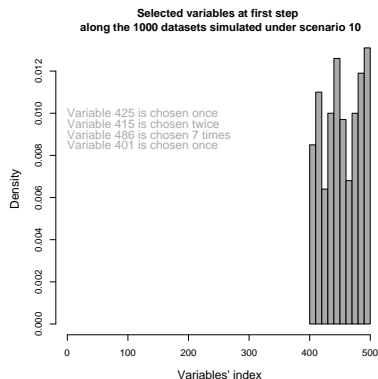


Factor-adjusted t-tests

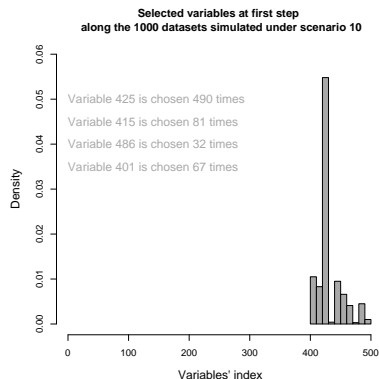


Stability in model selection

Usual t-tests

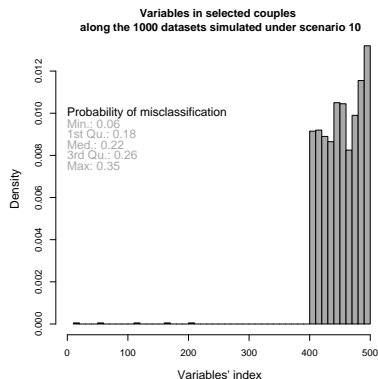


Factor-adjusted t-tests

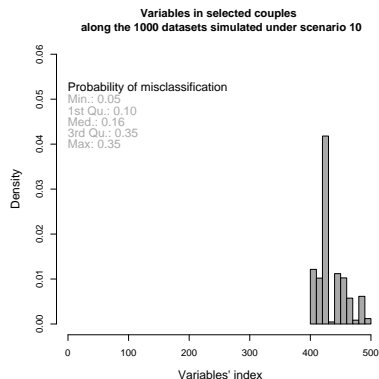


Stability in model selection

Usual t-tests



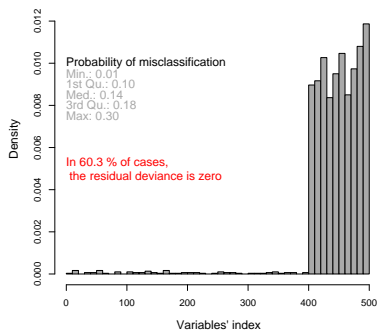
Factor-adjusted t-tests



Stability in model selection

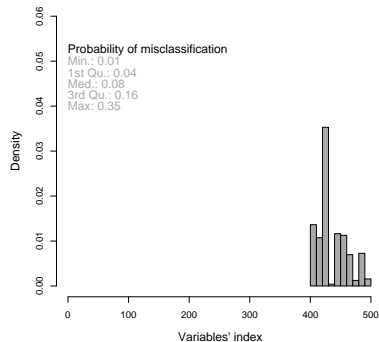
Usual t-tests

Variables in selected triplets
along the 1000 datasets simulated under scenario 10



Factor-adjusted t-tests

Variables in selected triplets
along the 1000 datasets simulated under scenario 10



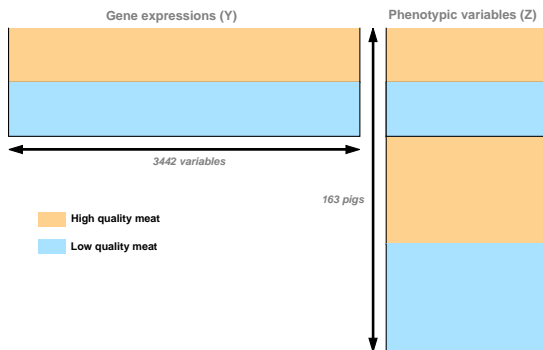
Outline

- 1 Stability of model selection
- 2 Factor-adjusted procedures
- 3 External information**
- 4 Concluding comments

Additional information in a biological context

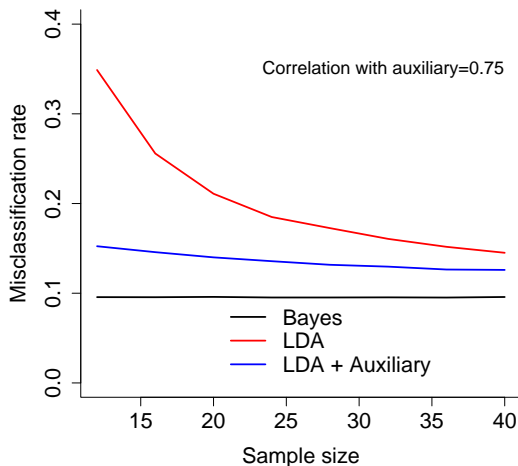
Aim : find genes involved in pork quality

Additional information : pH of the meat



Modified test statistics : improves the type II error rates of multiple testing procedures (Kloareg and Causeur, 2008, *in press*)

Involving auxiliary variables in linear discriminant analysis



Combining external information and factor analysis ?

Conditional centering and scaling

$$T_z^{(k)} = \frac{\sigma_k}{\psi_k} \left[T^{(k)} - \frac{b'_k}{\sigma_k} \tau(Z) \right]$$

Combining external information and factor analysis ?


Conditional centering and scaling **taking into account the external information**

$$T_Z^{(k)} N = \frac{\sigma_k}{\psi_k} \left[T^{(k)} - \frac{b'_k}{\sigma_k} \tau(Z_N) \right]$$


Outline

- 1 Stability of model selection
- 2 Factor-adjusted procedures
- 3 External information
- 4 Concluding comments**

Concluding comments

- **Factor-adjustment of test statistics** : large improvements in model selection
- In specific cases, **external information** should also decrease misclassification rates and improve stability of model selection
- Methods to be implemented as an  package

Concluding comments

- **Factor-adjustment of test statistics** : large improvements in model selection
- In specific cases, **external information** should also decrease misclassification rates and improve stability of model selection
- Methods to be implemented as an  package



The R User Conference 2009

July 8-10, Agrocampus Rennes, France

<http://www.agrocampus-rennes.fr/math/useR-2009/>