A Robust Approach to Regularized Discriminant Analysis

Moritz Gschwandtner

Department of Statistics and Probability Theory
Vienna University of Technology, Austria

Österreichische Statistiktage, Graz, Austria

September 08, 2011

Vienna University of Technology

Joint work with

Peter Filzmoser, Vienna University of Technology, Austria

Christophe Croux, ORSTAT and University Center of Statistics, K. U. Leuven, Belgium

Gentiane Haesbroeck, University of Liege, Liege, Belgium

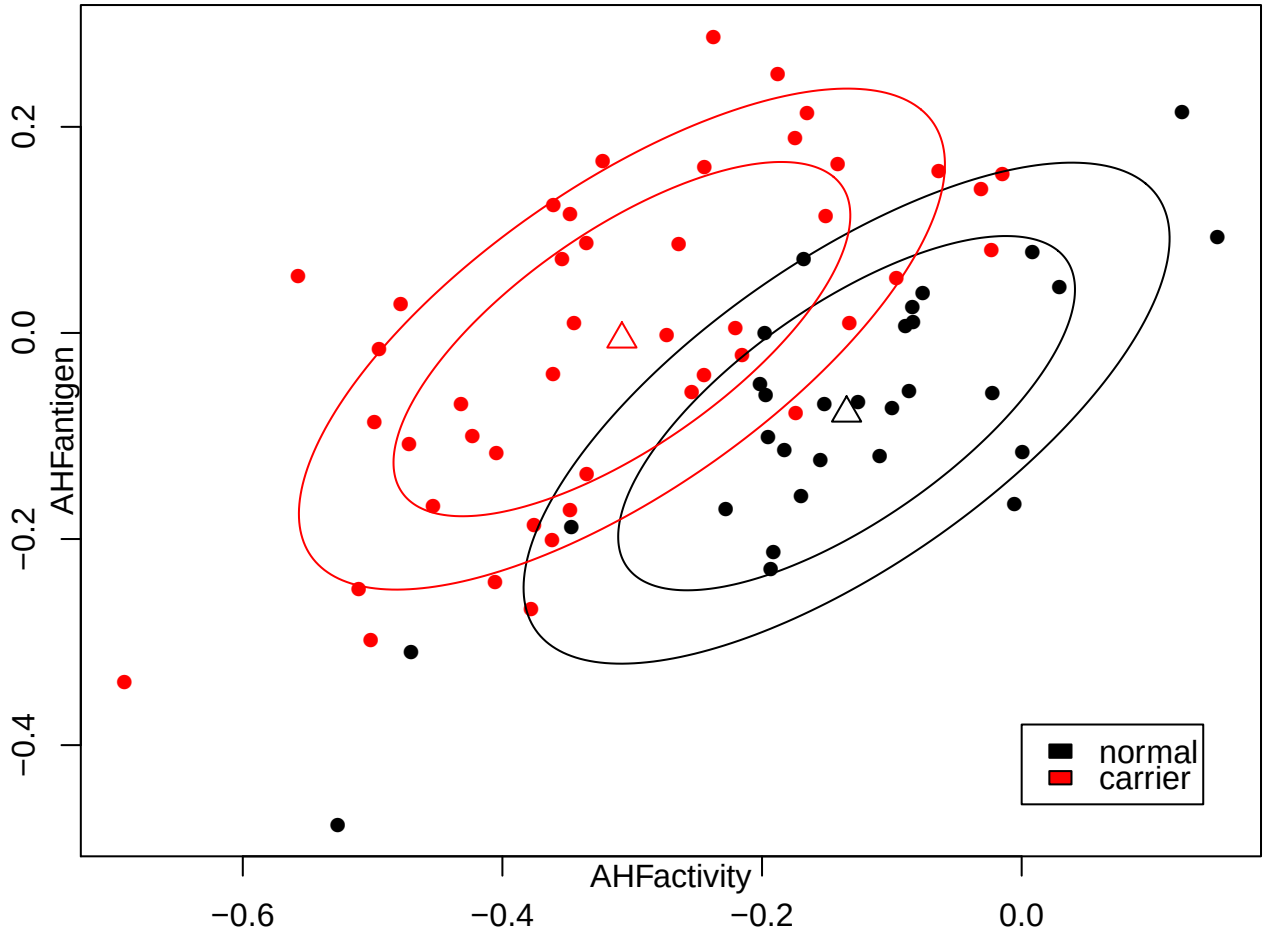
moritz.gschwandtner@chello.at

Contents

moritz.gschwandtner@chello.at

Discriminant Analysis (DA): Example

Haemophilia data: 30 normal persons and 22 obligatory carriers of hemophilia A



moritz.gschwandtner@chello.at

Notation

Given n observations of training data, measured at p variables.

Observations originate from

- k different populations $G_1, \ldots, G_k$,
- according to prior probabilities $\pi_1, \ldots, \pi_k$, where $\sum_{j=1}^{k} \pi_j = 1$,
- with sample sizes $n_1, \ldots, n_k$, where $\sum_{j=1}^{k} n_j = n$.
- **Usual assumption:** Observations are distributed according to a normal distribution $\mathbf{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$, $j = 1, \ldots, k$.

moritz.gschwandtner@chello.at

DA: Classification and Prediction

Find a classification function f based on the training data that assigns a new, unlabelled observation to one (and only one) of the k groups:

$$f : \Omega^p \to \{1, \ldots, k\}$$

Bayes Rule: Given an observation $\boldsymbol{x}$, the posterior probability for group G_j equals

$$P(G_j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|G_j) \cdot \pi_j}{\sum_{i=1}^{k} p(\boldsymbol{x}|G_i) \cdot \pi_i}$$

moritz.gschwandtner@chello.at

Bayesian DA Rule

A **test set observation** $\boldsymbol{x}$ is assigned to that population G_j, for which $\ln P(G_j|\boldsymbol{x})$ is a maximum over all groups $j = 1, \ldots, k$.

$$\Rightarrow f(\boldsymbol{x}) = \arg\max_j \Big(\ln(P(G_j|\boldsymbol{x}))\Big) = \arg\max_j \Big(\ln(p(\boldsymbol{x}|G_j) \cdot \pi_j)\Big)$$

Quadratic Discriminant Analysis:

$$f_{QDA}(\boldsymbol{x}) = \arg\max_j \left(-\frac{1}{2}\ln(\det \boldsymbol{\Sigma}_j) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j) + \ln \pi_j\right)$$

Linear Discriminant Analysis: assume $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$, and use

$$f_{LDA}(\boldsymbol{x}) = \arg\max_j \left(\boldsymbol{\mu}_j^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_j^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \ln \pi_j\right)$$

moritz.gschwandtner@chello.at

Parameter Estimation

The essential elements of the LDA rule are the group centers and the common group covariance matrix.

Estimate **group centers** and the **common group covariance matrix** by

- the sample means and pooled sample covariance matrix.
- **robust estimators** of location and covariance, like the MCD estimators.
- **regularized (sparse)** estimators of location and covariance.

Robust estimators lead to robust DA rules!

moritz.gschwandtner@chello.at

Maximum Likelihood

Given a data sample $\mathbf{X}$, the log-likelihood function of joint location $\boldsymbol{\mu}$ and inverse scatter $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$ is given by

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Theta}) = \log(\det(\boldsymbol{\Theta})) - \frac{1}{n} \sum_{\boldsymbol{x} \in \mathbf{X}} (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Theta} (\boldsymbol{x} - \boldsymbol{\mu})$$

$\Rightarrow$ Maximization leads to **classical estimators:**

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{\boldsymbol{x} \in \mathbf{X}} \boldsymbol{x}$$

$$\hat{\boldsymbol{\Theta}} = \hat{\boldsymbol{\Sigma}}^{-1} = (\frac{1}{n} \sum_{\boldsymbol{x} \in \mathbf{X}} (\boldsymbol{x} - \hat{\boldsymbol{\mu}})(\boldsymbol{x} - \hat{\boldsymbol{\mu}})^\top)^{-1}$$

moritz.gschwandtner@chello.at

Regularization

Problem: If $n < p$, $\hat{\Sigma}$ is singular and the maximum likelihood estimator for Θ does not exist!

Solution: Penalization of log-likelihood function based on penalty term $\lambda > 0$ and L1 Norm $\|.\|_1$:

$$\mathcal{L}(\boldsymbol{\mu}, \Theta) = \log(\det(\Theta)) - \frac{1}{n} \sum_{\boldsymbol{x} \in \mathbf{X}} (\boldsymbol{x} - \boldsymbol{\mu})^\top \Theta (\boldsymbol{x} - \boldsymbol{\mu}) - \lambda \|\Theta\|_1$$

$$\|\Theta\|_1 = \sum_{l,m} |\theta_{lm}|$$

The maximization problem can be solved by an algorithm called graphical lasso. λ governs sparseness of $\hat{\Sigma}$ and $\hat{\Theta}$!

package: `glasso` (Friedman, Hastie, Tibshirani, 2007).

moritz.gschwandtner@chello.at

Example

Simulated three-dimensional data, $\mathbf{X} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_3), n = 100$:

```
> solve(cov(X))
       [,1]   [,2]   [,3]
[1,]  1.179 -0.039 -0.088
[2,] -0.039  0.835 -0.147
[3,] -0.088 -0.147  1.100

> glasso(cov(X), rho=0.2)$wi
      [,1] [,2]  [,3]
[1,] 0.948  0.0 0.000
[2,] 0.000  0.7 0.000
[3,] 0.000  0.0 0.879
```

$\lambda = 0.2$ leads to **sparse estimate** of concentration matrix!

moritz.gschwandtner@chello.at

Example

Problem: `glasso` not robust!

Adding 10 outliers distributed according to $\mathbf{N}(\mathbf{10}, \mathbf{I}_3)$ leads to

```
> glasso(cov(X), rho=0.2)$wi
       [,1]   [,2]   [,3]
[1,]  0.504 -0.224 -0.283
[2,] -0.224  0.504 -0.216
[3,] -0.283 -0.216  0.552
```

Idea: Combine regularization of `glasso` with robust techniques!

moritz.gschwandtner@chello.at

Regularized MCD estimator

Croux and Haesbroeck (2010):

Improvement: Adapt MCD idea and integrate it into log-likelihood function:

$$\mathcal{L}(H, (\boldsymbol{\mu}, \boldsymbol{\Theta})) = \log(\det(\boldsymbol{\Theta})) - \frac{1}{h} \sum_{i \in H} (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Theta} (\boldsymbol{x}_i - \boldsymbol{\mu}) - \lambda \|\boldsymbol{\Theta}\|_1$$

with

$$H \subseteq \{1, \ldots, n\}, \ |H| = h < n$$

Maximization of $\mathcal{L}(H, (\boldsymbol{\mu}, \boldsymbol{\Theta}))$ means to find an index subset H_{opt} for which

$$\max_{(\boldsymbol{\mu}, \boldsymbol{\Theta})} \mathcal{L}(H_{opt}, (\boldsymbol{\mu}, \boldsymbol{\Theta})) \geq \max_{(\boldsymbol{\mu}, \boldsymbol{\Theta})} \mathcal{L}(H, (\boldsymbol{\mu}, \boldsymbol{\Theta}))$$

$$\forall H \subseteq \{1, \ldots, n\} : |H| = h$$

moritz.gschwandtner@chello.at

Regularized MCD estimator

Problem: $\binom{n}{h}$ subsets to check. Not applicable to large n.

Improvement: C-Step Algorithm: Let H_k be the subset derived at iteration k and $(\hat{\boldsymbol{\mu}}_{H_k}, \hat{\boldsymbol{\Theta}}_{H_k})$ be the corresponding estimates maximizing $\mathcal{L}(H_k, (\boldsymbol{\mu}, \boldsymbol{\Theta}))$.

Compute Mahalanobis distances with respect to $(\hat{\boldsymbol{\mu}}_{H_k}, \hat{\boldsymbol{\Theta}}_{H_k})$:

$$d_i^{(k)}(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_{H_k}, \hat{\boldsymbol{\Theta}}_{H_k}) = \sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{H_k})^\top \hat{\boldsymbol{\Theta}}_{H_k} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{H_k})}$$

Define next subset H_{k+1} as

$$H_{k+1} = \left\{ i \in \{1, \ldots, n\} : d_i^{(k)} \in \{d_{(1)}^{(k)}, \ldots, d_{(h)}^{(k)}\} \right\}$$

where $d_{(j)}^{(k)}$ are the ordered distances.

$$\Rightarrow \mathcal{L}(H_k, (\boldsymbol{\mu}_{H_k}, \boldsymbol{\Theta}_{H_k})) \leq \mathcal{L}(H_{k+1}, (\boldsymbol{\mu}_{H_{k+1}}, \boldsymbol{\Theta}_{H_{k+1}}))$$

moritz.gschwandtner@chello.at

Algorithm

The regularized MCD estimator is computed using the following algorithm:

1. Draw **initial subset** H_0
2. Maximize **penalized likelihood function** (`glasso`) to obtain $(\hat{\boldsymbol{\mu}}_{H_0}, \hat{\boldsymbol{\Theta}}_{H_0})$
3. Compute **ordered Mahalanobis distances** w.r.t. $(\hat{\boldsymbol{\mu}}_{H_0}, \hat{\boldsymbol{\Theta}}_{H_0})$
4. Choose next subset containing h observations with **smallest distances**
5. Repeat steps 2-4 until convergence to obtain $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Theta}})$

$\Rightarrow$ A local maximum of the likelihood value is reached. Algorithm can be repeated several times with different initial subsets.

moritz.gschwandtner@chello.at

LDA and RegMCD

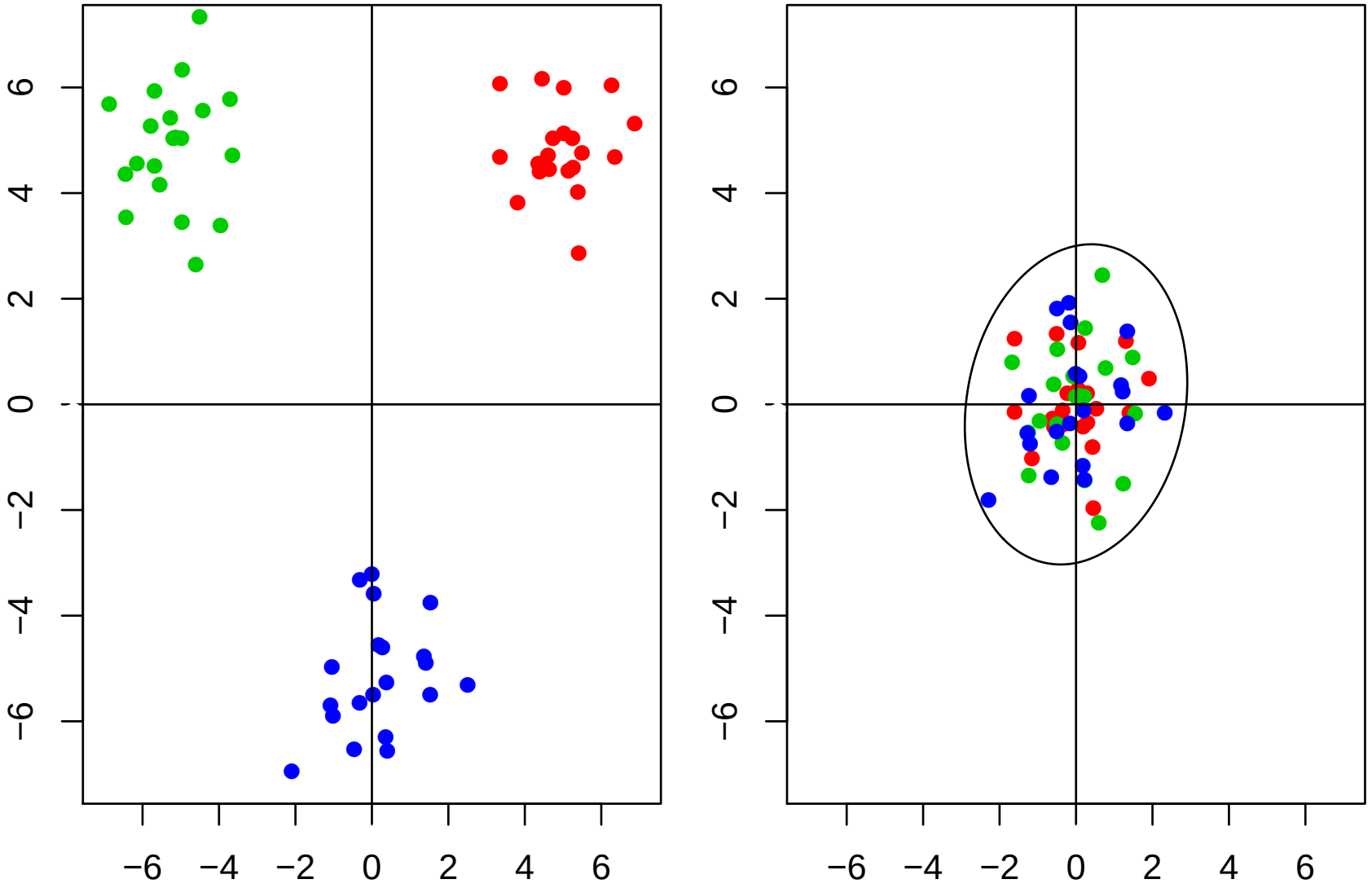
How to apply LDA with the regularized MCD estimator in the multi group setting:

$$\mathbf{X} = \{\mathbf{x}_{ij} : i = 1, \ldots, n_j; j = 1, \ldots, k\}$$

- Compute **robust location estimates** $\boldsymbol{t}_j$ for $j = 1, \ldots, k$
- Compute **centered observations** $\mathbf{Z} = \{\boldsymbol{z}_{ij}\}$ with $\boldsymbol{z}_{ij} = \boldsymbol{x}_{ij} - \boldsymbol{t}_j$
- Apply the **regularized MCD algorithm** to $\mathbf{Z}$ to obtain common estimates $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Theta}})$
- Correct location estimates: $\hat{\boldsymbol{\mu}}_j = \boldsymbol{t}_j + \hat{\boldsymbol{\mu}}$
- Apply LDA using parameters $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Theta}}$

moritz.gschwandtner@chello.at

Centering



moritz.gschwandtner@chello.at

The penalty parameter λ

How to choose the penalty parameter λ:

- Based on test error rates: Cross Validation
- Based on a model selection criterion: AIC, BIC

BIC criterion:

$$BIC(\Gamma) = -2 \cdot \log L(\Gamma) + \kappa(\Gamma) \log n$$

$L(\Gamma)$ . . . Likelihood function of the model
$\kappa(\Gamma)$ . . . Number of parameters in the model

moritz.gschwandtner@chello.at

The penalty parameter λ

$BIC(\lambda)$ is small if

- the value of the likelihood function $\mathcal{L}(H_{opt}, \hat{\boldsymbol{\theta}})$ is high
- the number of parameters in the model is small

$\Rightarrow$ Choose λ according to

$$\lambda_{opt} = \arg\min_{\lambda} BIC(\lambda)$$

Best compromise between likelihood and sparseness!

moritz.gschwandtner@chello.at

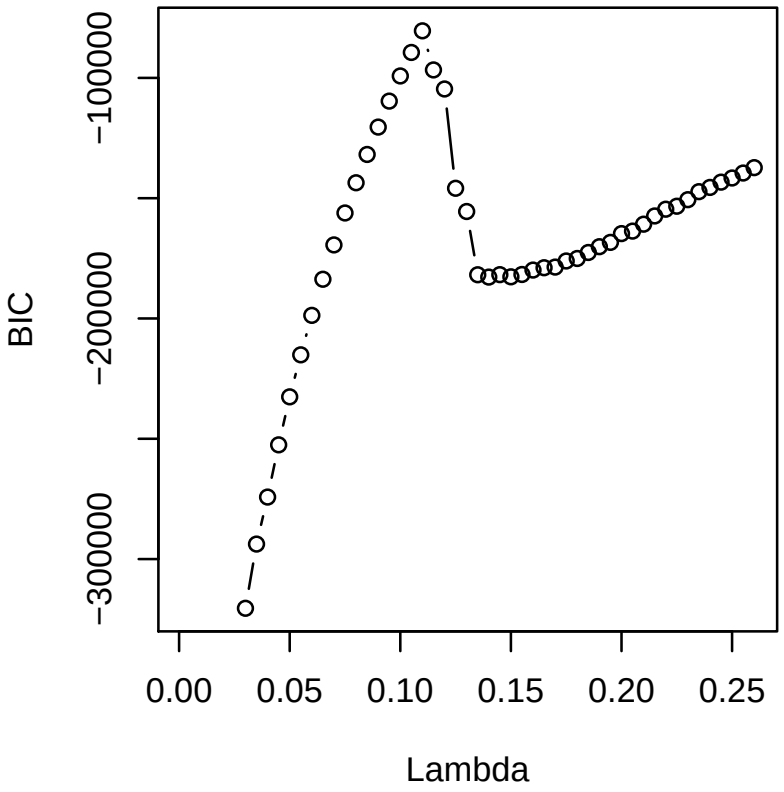
Example: Fruit Data

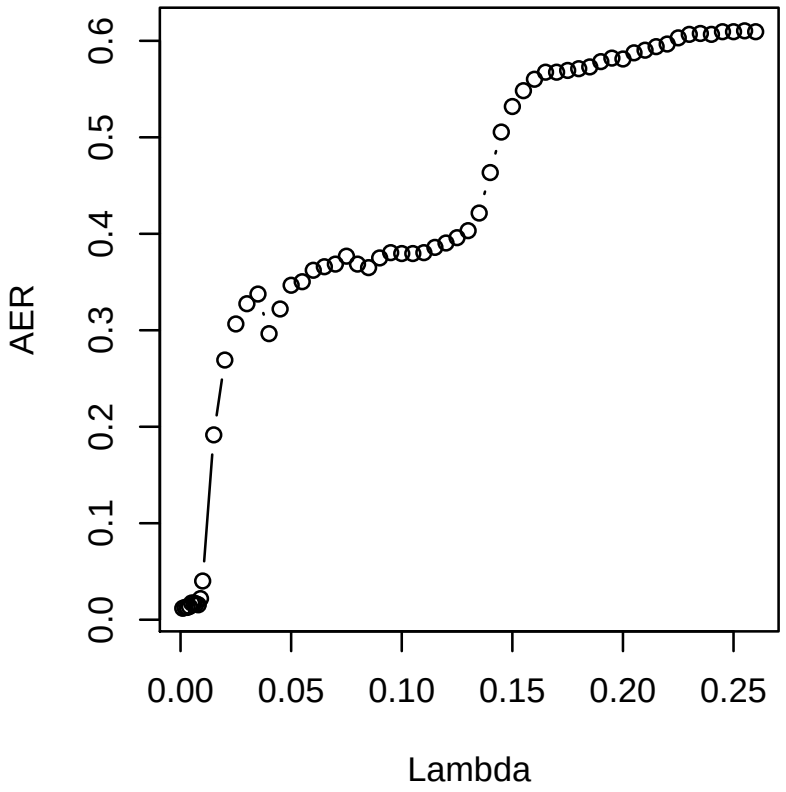
- Three different sorts of the same fruit (cucumis melo)
- 256 different spectra measured
- Outliers due to different illumination systems
- Partition of data into 60% training and 40% test set
- Test errors measured for each group separately

moritz.gschwandtner@chello.at

Example: Fruit Data, BIC

BIC and **AER** suggest a small λ value!





moritz.gschwandtner@chello.at

Example: Fruit Data, Results

- **Outliers** in the third group lead to poor results for LDA.
- **RRLDA** remains stable!

	Err_{Test_1}	Err_{Test_2}	Err_{Test_3}
RRLDA ($\lambda = 0.001$)	0.02	0.03	0.01
GLASSO ($\lambda = 0.001$)	0.04	0.03	0.03
LDA	0.01	0.01	0.14

moritz.gschwandtner@chello.at

Example: Golub Data, Results

- 38 training samples and 34 test samples from two cancer classes.
- Absolute test errors were measured for various variable subsets.
- Variable selection was done according to the nearest shrunken centroids method.

p	LDA	GLASSO	RRLDA
41	7	3	3
86	8	3	1
111	6	2	1
142	4	2	1
174	5	2	2
221	6	2	2
290	4	3	1
392	5	2	2
476	5	3	1
625	7	4	2

moritz.gschwandtner@chello.at

Simulated Example

- Two groups ($k = 2$) both consisting of 100 observations and p variables with

$$p \in \{30, 100, 300, 500, 1000\}$$

- Discrimination occurs in variables 1 and 2.

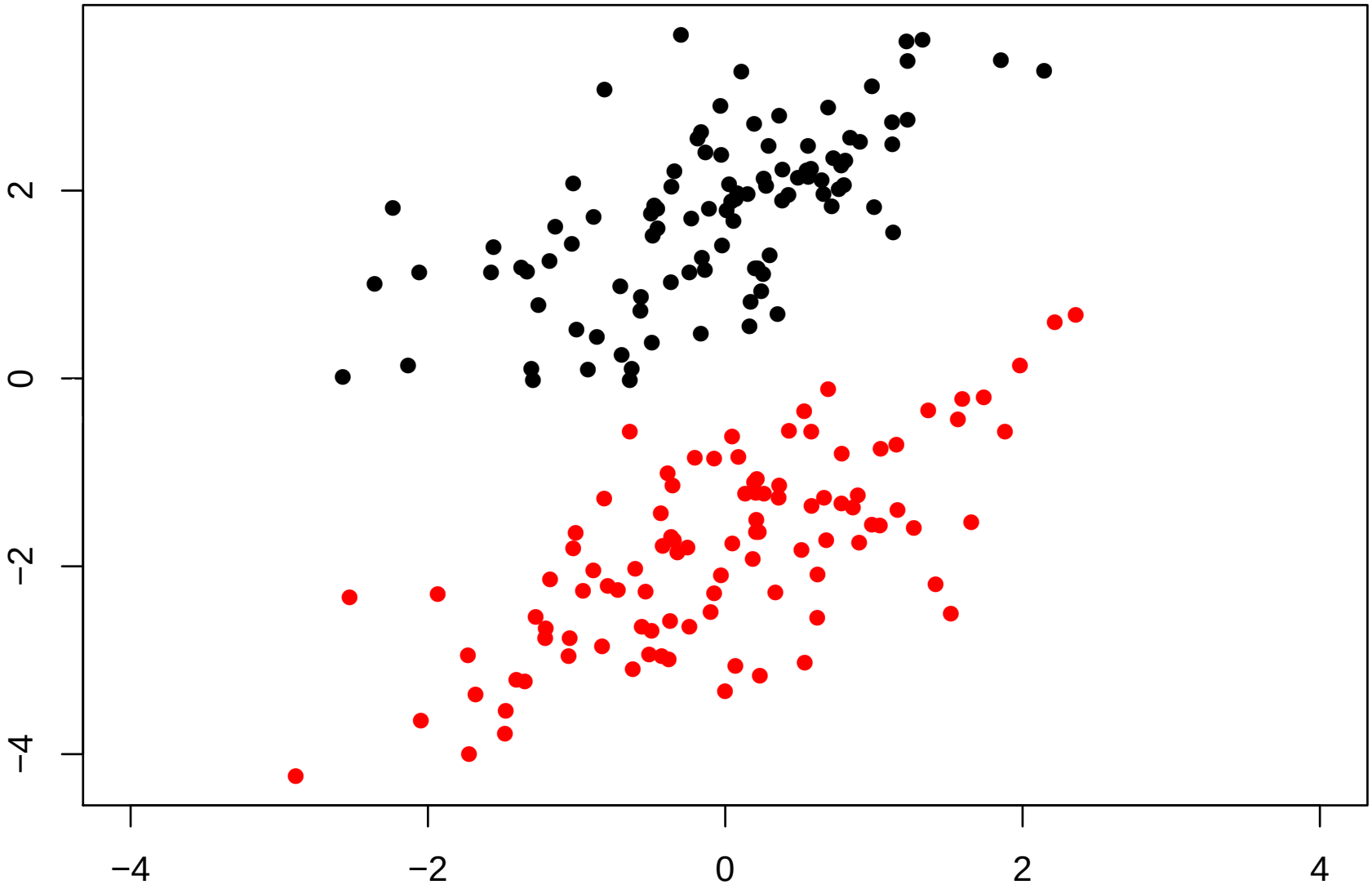
- Variables 3 - p are uncorrelated noise according to standard normal distributions.

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 1.9 \end{pmatrix} \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 0 \\ -1.9 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$$

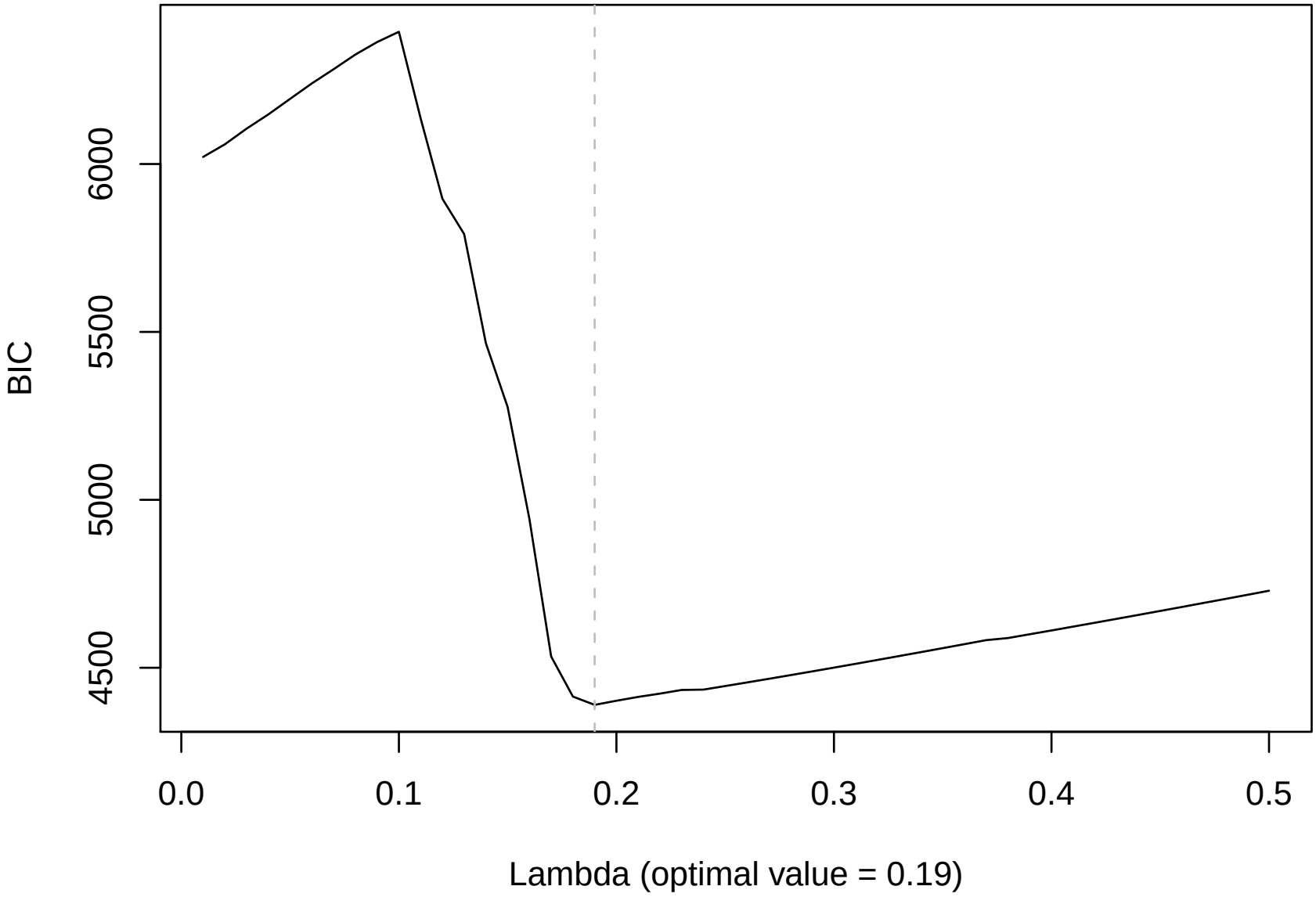
moritz.gschwandtner@chello.at

Simulated Example



moritz.gschwandtner@chello.at

Simulated Example, BIC, p=30



moritz.gschwandtner@chello.at

Simulated Example, Outliers

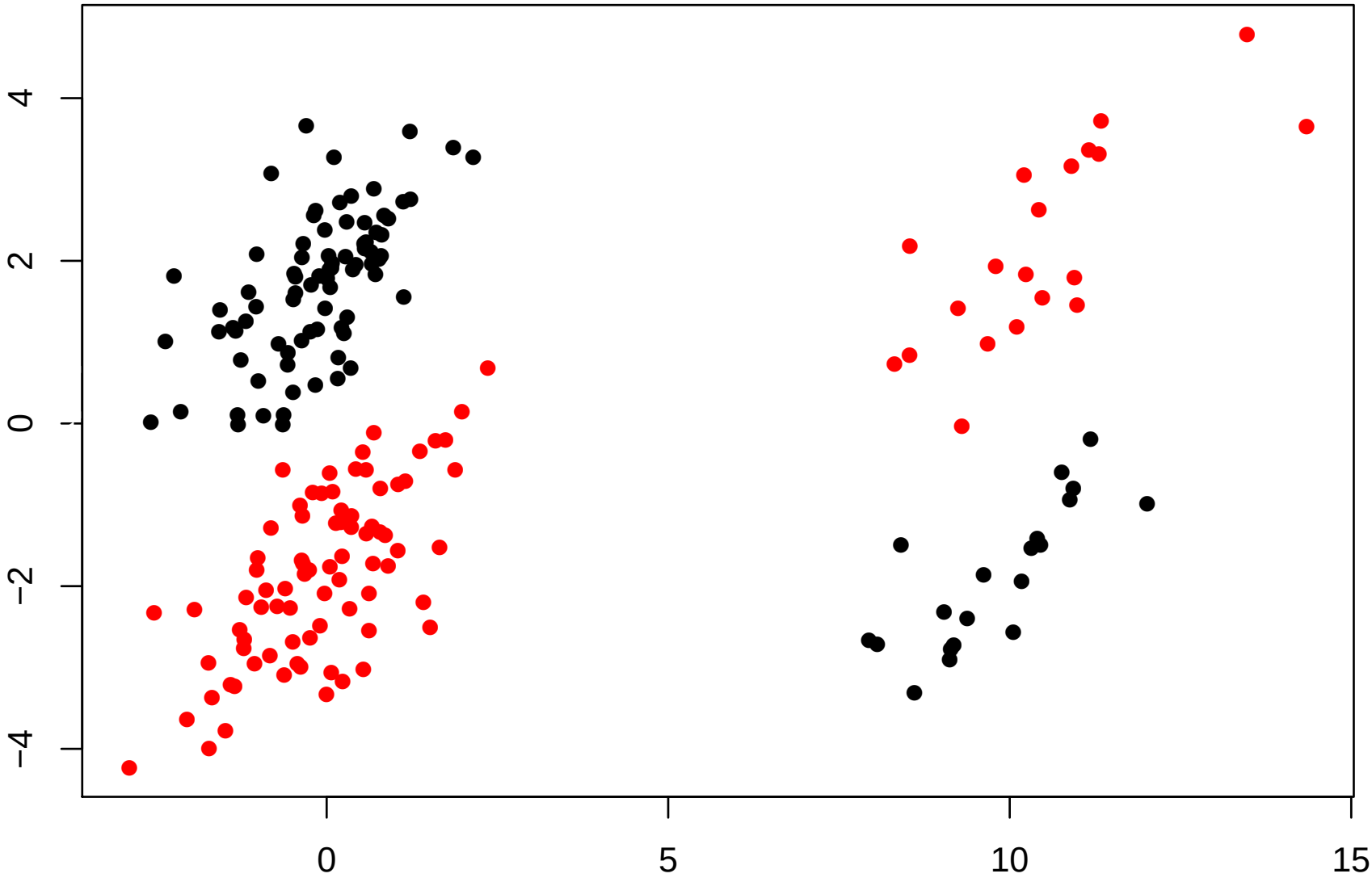
Simulate **contamination** by adding 10% **shift outliers** to the data.

- Variables 3 - p are distributed like non-outliers.
- Mean of variable 1 is shifted.
- Means of variable 2 are swapped.

$$\tilde{\mu}_1 = \begin{pmatrix} 10 \\ -1.9 \end{pmatrix} \quad \tilde{\mu}_2 = \begin{pmatrix} 10 \\ 1.9 \end{pmatrix}$$

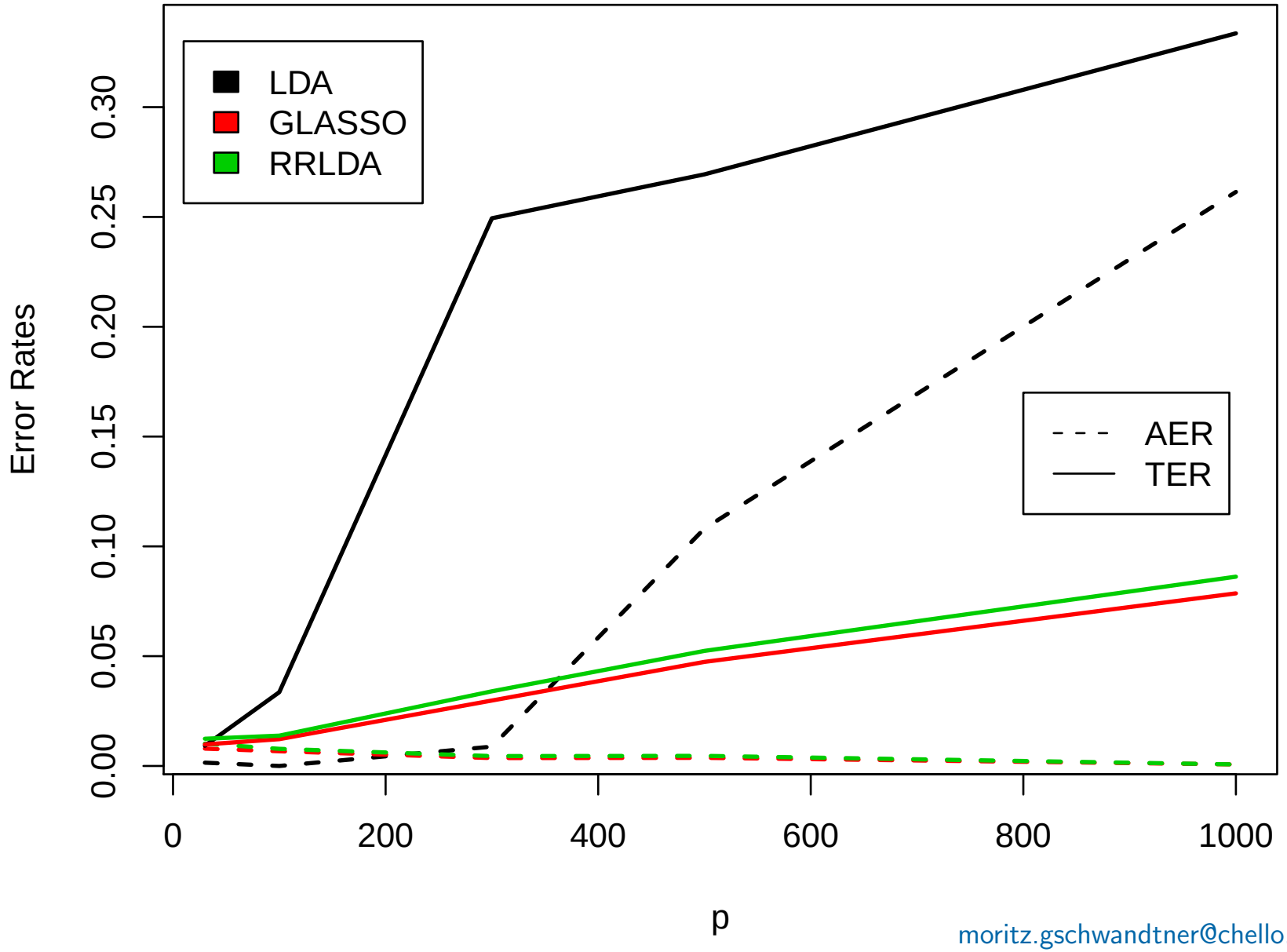
moritz.gschwandtner@chello.at

Simulated Example, Outliers



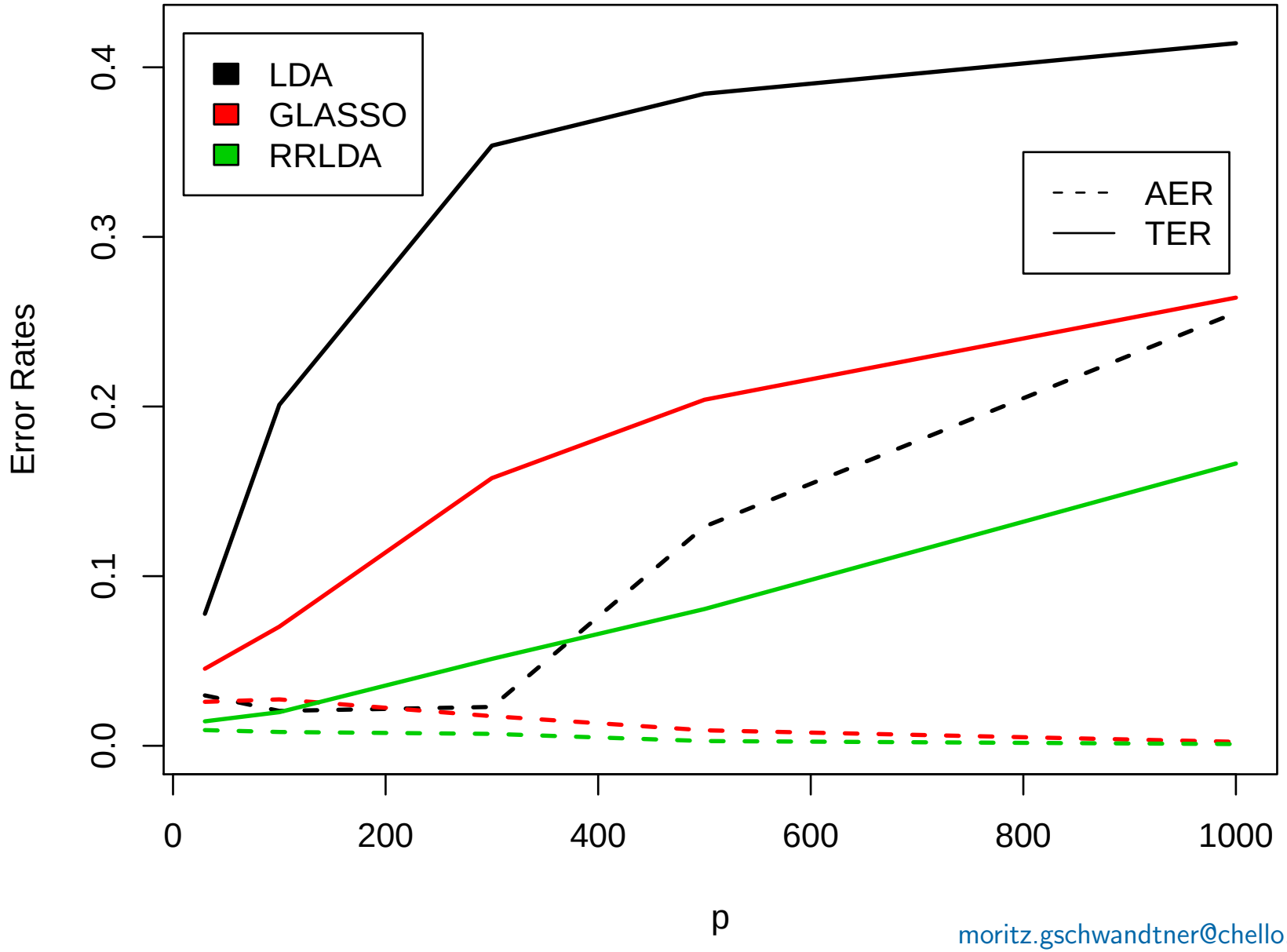
moritz.gschwandtner@chello.at

Results without contamination



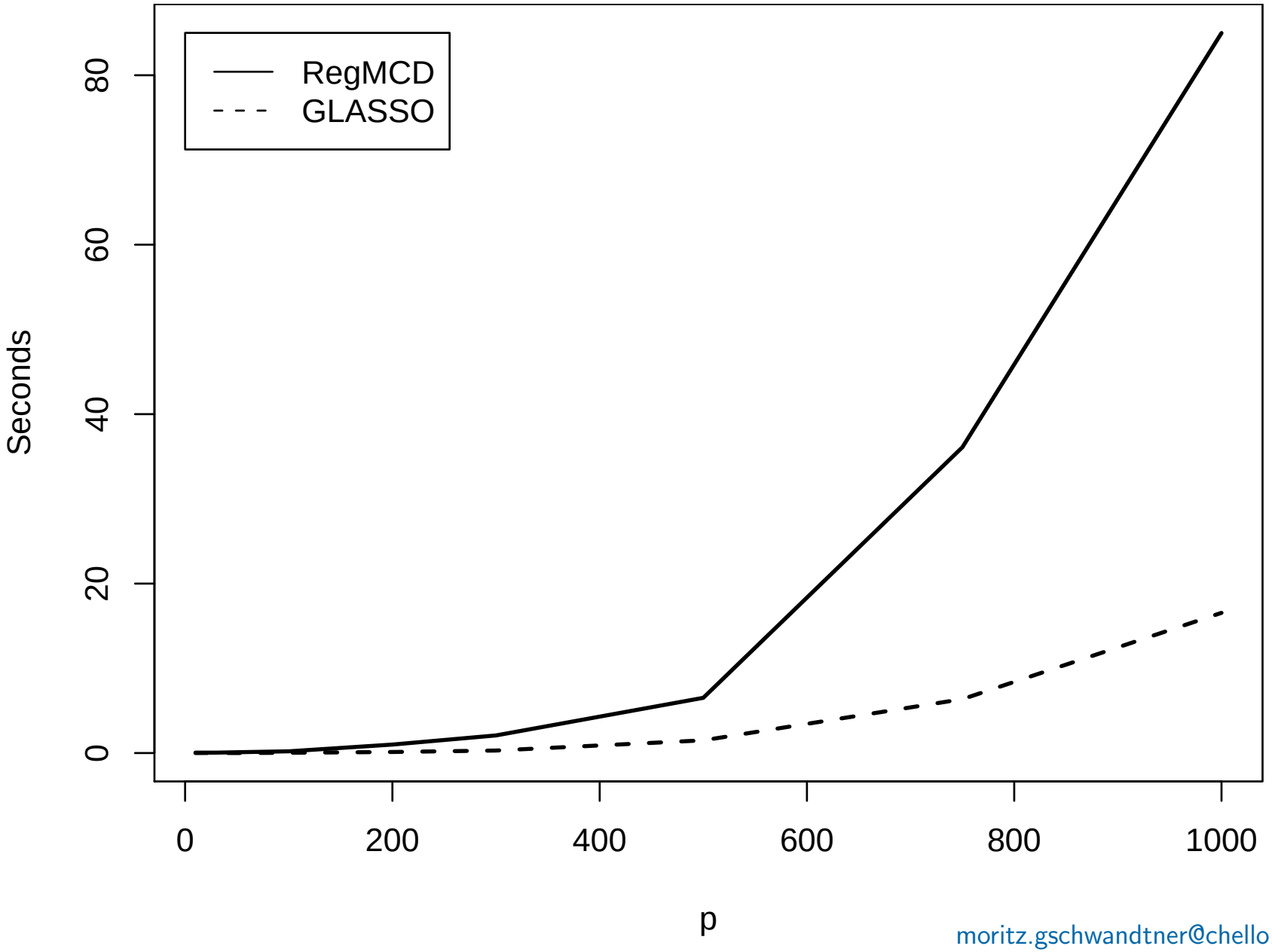
moritz.gschwandtner@chello.at

Results with contamination



moritz.gschwandtner@chello.at

Computation Times



moritz.gschwandtner@chello.at

Conclusions

- RRLDA is a combination of **regularization** and **robust methods**.
- RRLDA is a good choice if data contain either **outliers** or many **noisy variables** or both.
- Penalty parameter λ is chosen according to an **adapted BIC criterion**.

moritz.gschwandtner@chello.at

Some references

C. Croux and G. Haesbroeck. *Robust scatter regularization*. Compstat, Book of Abstracts, Paris: Conservatoire National des Arts et M´etiers (CNAM) and the French National Institute for Research in Computer Science and Control (INRIA), 2010.

J.H. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432-441, 2007.

J.H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84, 165-175, 1989.

P. Filzmoser, R. Maronna, and M. Werner. Outlier identification in high dimensions. *Computational Statistics and Data Analysis*, 52, 1694-1711, 2008.