Regularized Directions of Maximal Outlyingness

Michiel Debruyne

Dept. of mathematics and computer science, *Universiteit Antwerpen*

COMPSTAT 2010 August 23, 2010

Motivation

Nowadays many robust methods are available to detect outliers in a multivariate, possibly high-dimensional data set (e.g. robust covariance estimators, robust PCA methods, . . .).

Once an observation is flagged as an outlier, it is often interesting to know which variables contribute most to this outlyingness.

Motivation

Nowadays many robust methods are available to detect outliers in a multivariate, possibly high-dimensional data set (e.g. robust covariance estimators, robust PCA methods, $\ldots$).

Once an observation is flagged as an outlier, it is often interesting to know which variables contribute most to this outlyingness.

Given observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ with $\boldsymbol{x}_i \in \mathbb{R}^p$. Given weights $w_i > 0$ determining the outlyingness of $\boldsymbol{x}_i$ (e.g. based on robust Mahalanobis distances). Suppose w_i is small (so $\boldsymbol{x}_i$ is outlying). Let $k < p$.

Goal: select k variables out of p that contribute most to the outlyingness of $\boldsymbol{x}_i$.
$\longmapsto$ Variable selection for outliers.

Overview

1. A simple idea.
   (a) Outline.
   (b) Problems.
2. Main proposal.
3. Two algorithms
   (a) Moderate dimension.
   (b) High dimension.
4. Example.

1. A simple idea

Denote $\bar{\boldsymbol{x}}_w$ the weighted sample mean and and S_w the weighted sample covariance matrix.

A typical measure of the outlyingness of $\boldsymbol{x}_i$ is its squared robust Mahalanobis distance:

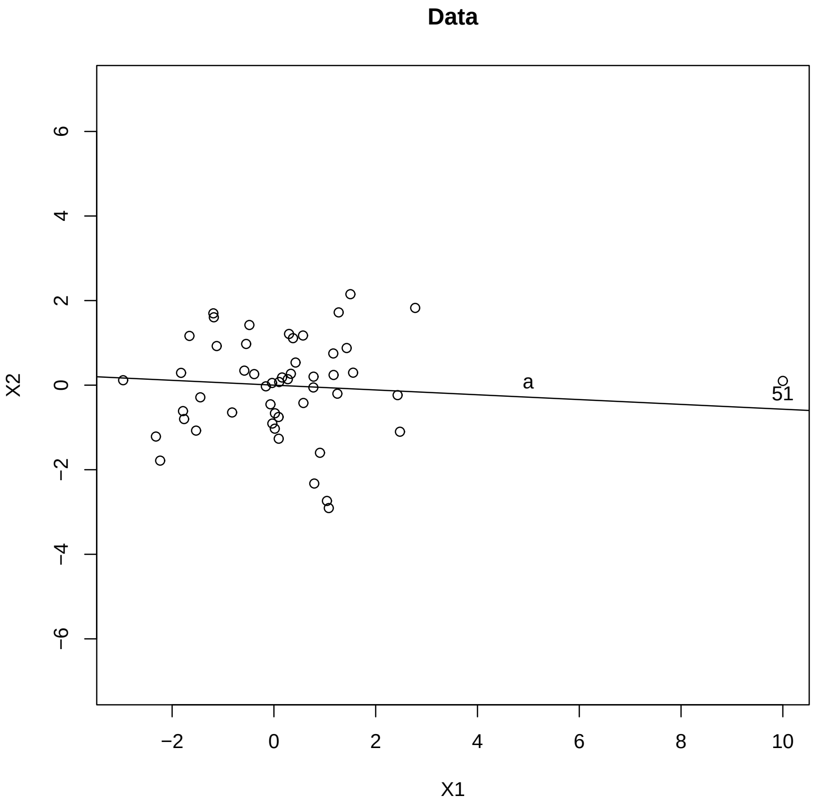
$$(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_w)^t S_w^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_w).$$

It is well known that this also equals the maximal standardized distance between the projection of $\boldsymbol{x}_i$ and the projection of the weighted sample mean:

$$(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_w)^t S_w^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_w) = \max_{\boldsymbol{a} \in \mathbb{R}^p, \|\boldsymbol{a}\|=1} \frac{\left(\boldsymbol{a}^t \boldsymbol{x}_i - \boldsymbol{a}^t \bar{x}_w\right)^2}{\boldsymbol{a}^t S_w \boldsymbol{a}}.$$

A simple idea is to check the coefficients of the direction $\boldsymbol{a}$ for which the maximum on the right hand side is attained.

1. A simple idea: example



$\boldsymbol{a} = (0.99, 0.14) \Rightarrow X_1$ contributes most to the outlyingness of observation 51.

1. A simple idea: problems

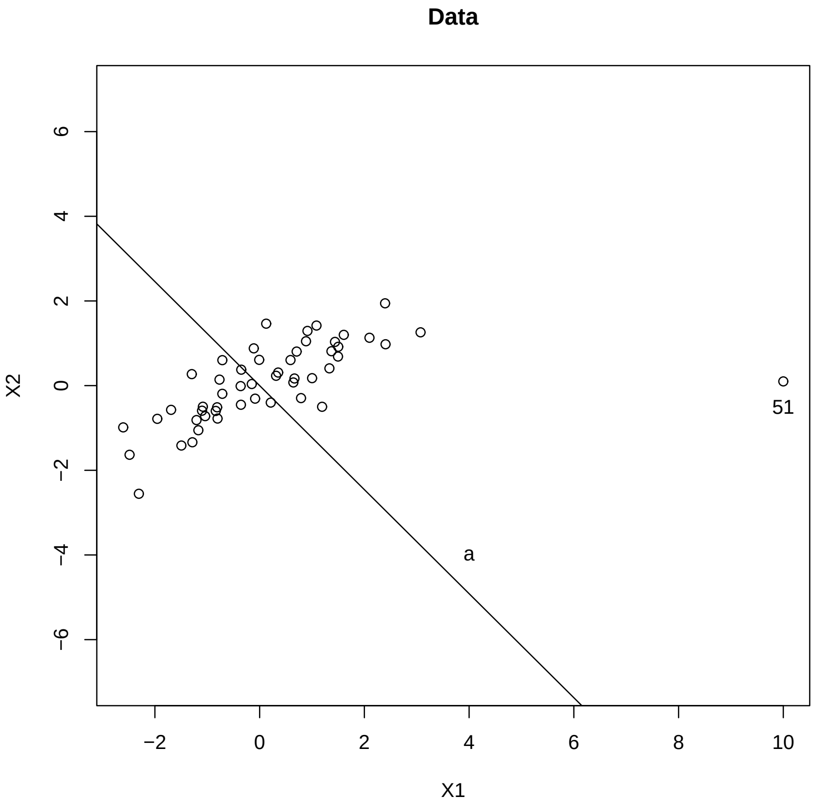
Note that

$$\arg \max_{\boldsymbol{a} \in \mathbb{R}^p, \|\boldsymbol{a}\|=1} \frac{\left(\boldsymbol{a}^t \boldsymbol{x}_i - \boldsymbol{a}^t \bar{\boldsymbol{x}}_w\right)^2}{\boldsymbol{a}^t S_w \boldsymbol{a}} = \frac{S_w^{-1} \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_w\right)}{\left\| S_w^{-1} \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_w\right) \right\|}.$$

This direction of maximal outlyingness can be computed very easily, but

- Does not work in high dimensions (p>n).
- Even in moderate dimensions the curse of dimensionality causes trouble.
- Very dependent on the covariance structure.

1. A simple idea: problems



2. Main proposal

Result

Let $X_w = (w_1(\boldsymbol{x}_1^t - \bar{\boldsymbol{x}}_w^t), \ldots, w_n(\boldsymbol{x}_n^t - \bar{\boldsymbol{x}}_w^t))^t$.

Let $\boldsymbol{y}_w = (n-1)\frac{\boldsymbol{e}_i}{w_1}$ with $\boldsymbol{e}_i$ the ith canonical basis vector.

Then the direction of maximal outlyingness can be written as a normed LS solution.

$$\arg \max_{\boldsymbol{a}\in\mathbb{R}^p, \|\boldsymbol{a}\|=1} \frac{\left(\boldsymbol{a}^t\boldsymbol{x}_i - \boldsymbol{a}^t\bar{x}_w\right)^2}{\boldsymbol{a}^t S_w \boldsymbol{a}} = \frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|} \text{ with } \boldsymbol{\theta} = \arg \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{y}_w - X_w\boldsymbol{\beta}\|^2$$

Proposal

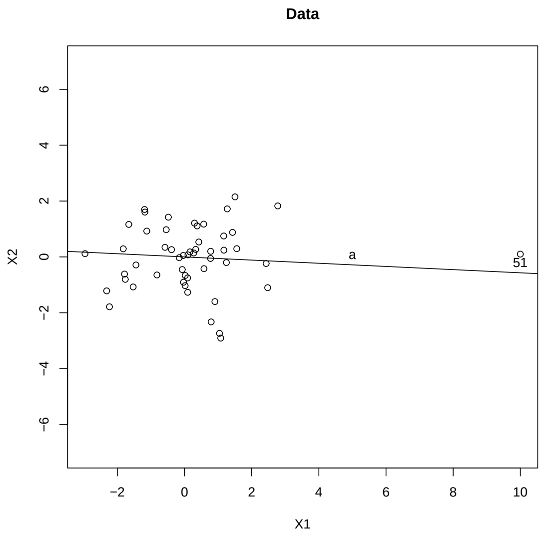
Add a L_1 type penalty:

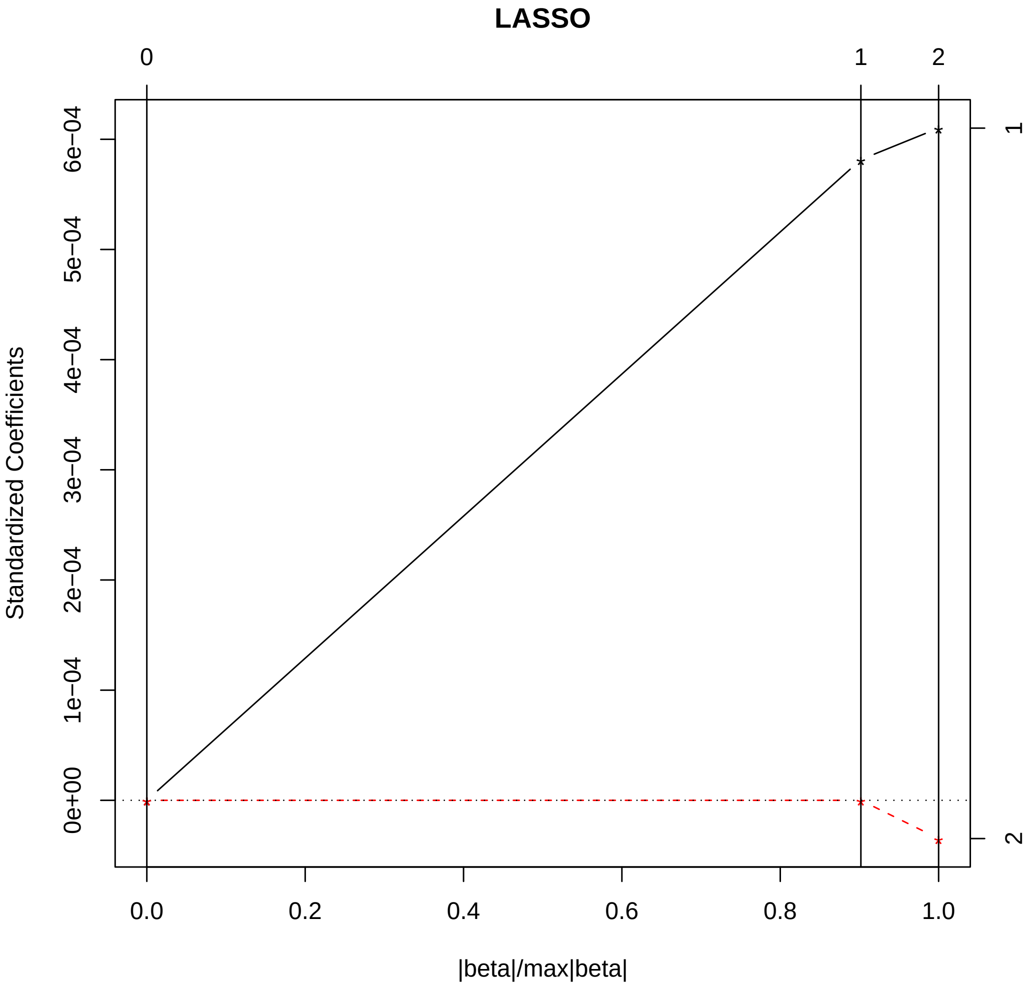
$$\boldsymbol{a}(t) = \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|} \text{ with } \boldsymbol{\theta}(t) = \arg \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{y}_w - X_w\boldsymbol{\beta}\|^2 \text{ subject to } \sum_{j=1}^p |\beta_j| \le t.$$

This yields a path of sparse directions of maximal outlyingness.

2. Examples revisited

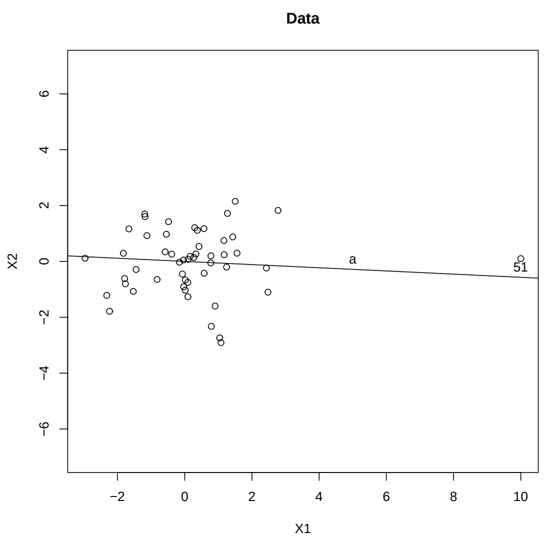
2 dimensions

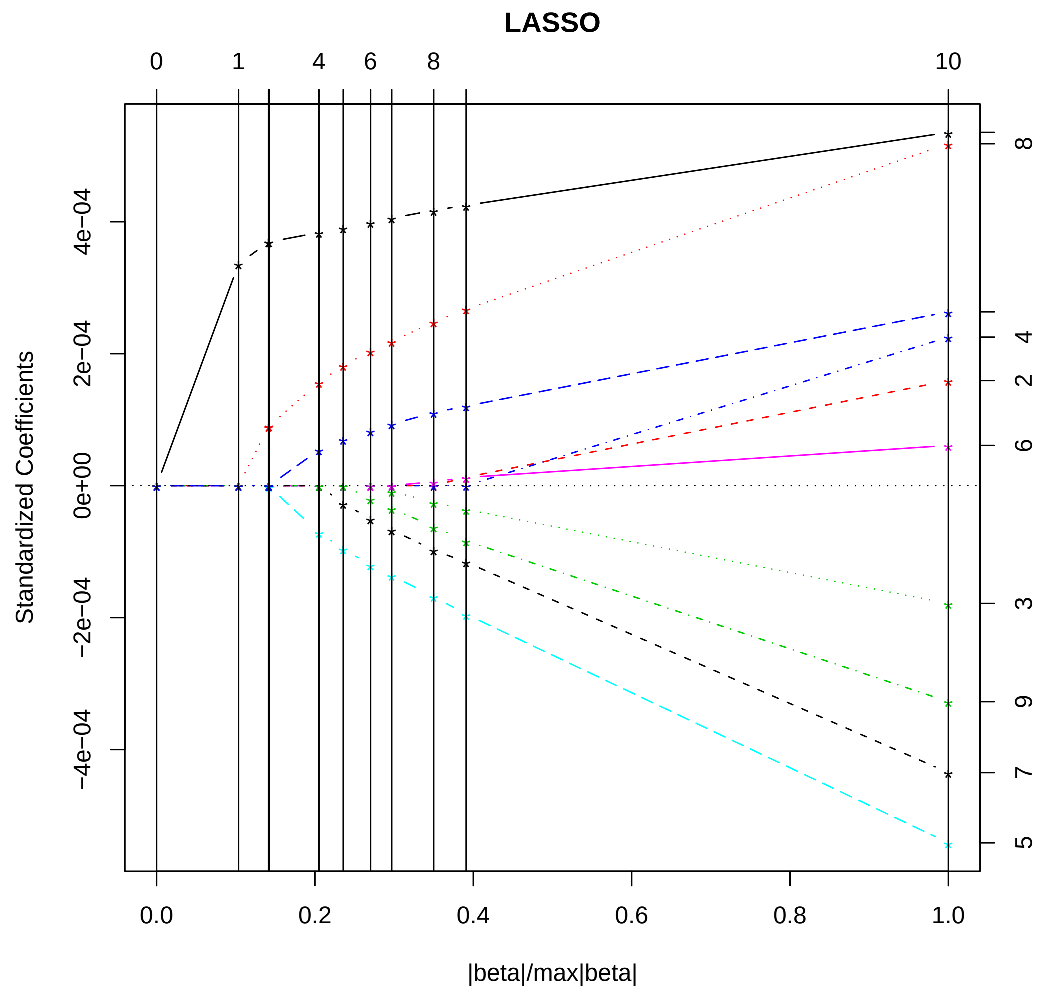




2. Examples revisited

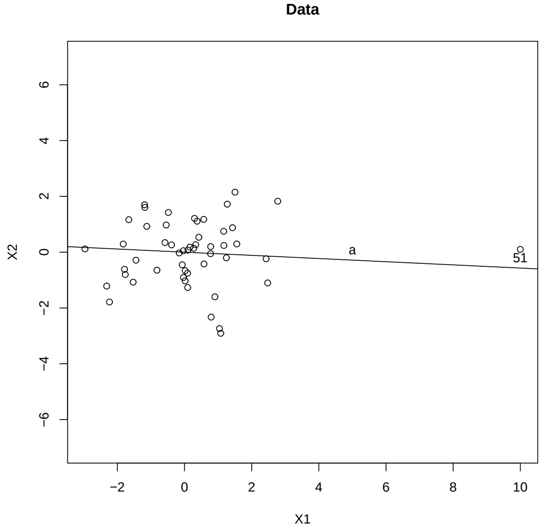
10 dimensions

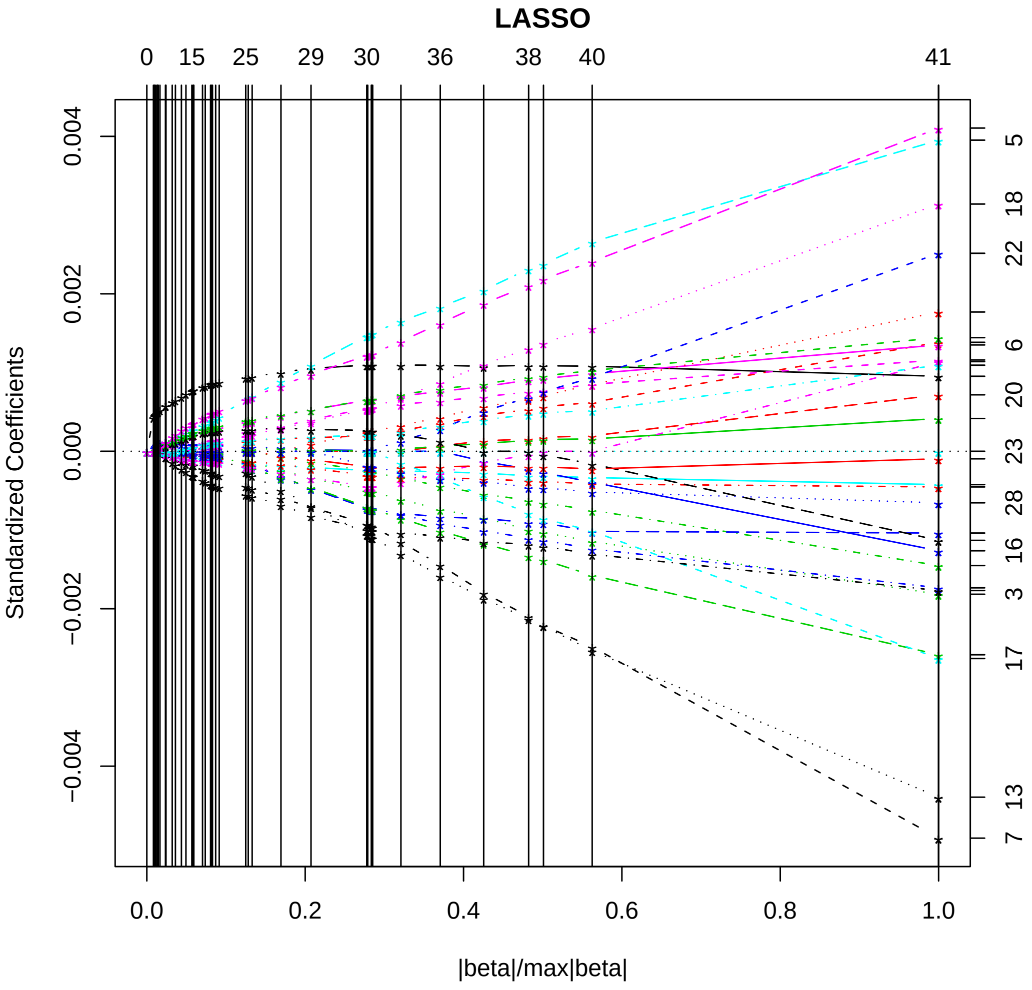




2. Examples revisited

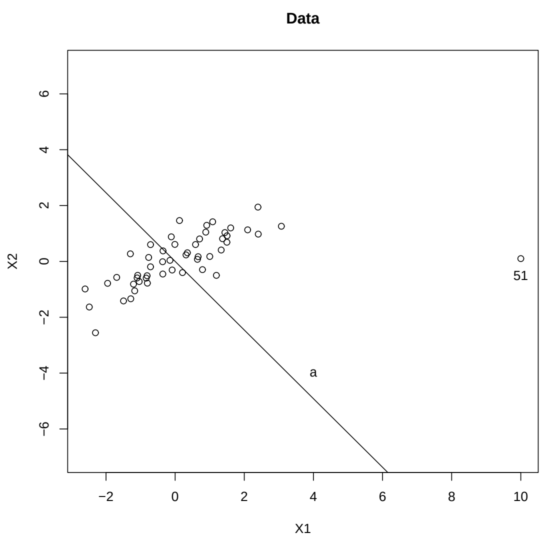
30 dimensions

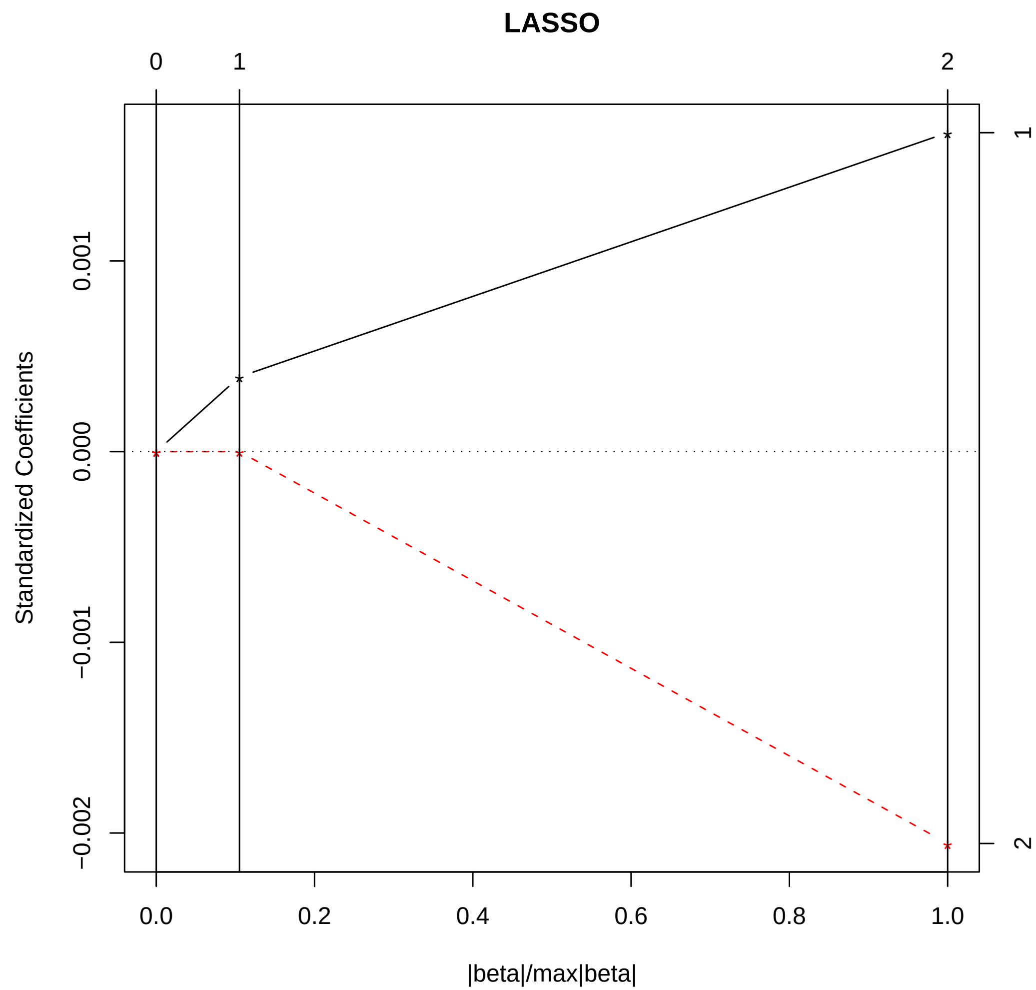




2. Examples revisited

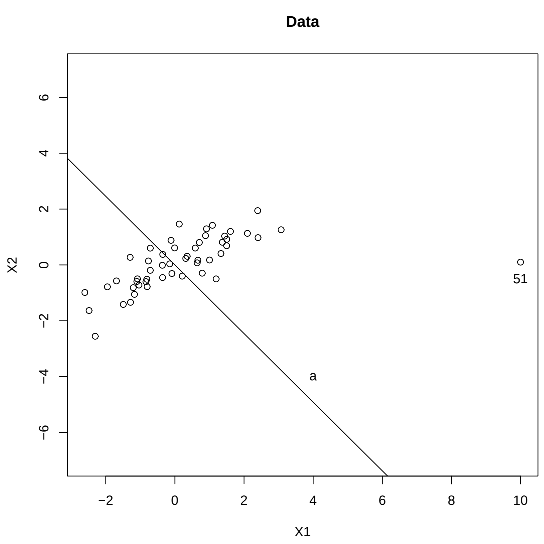
2 dimensions

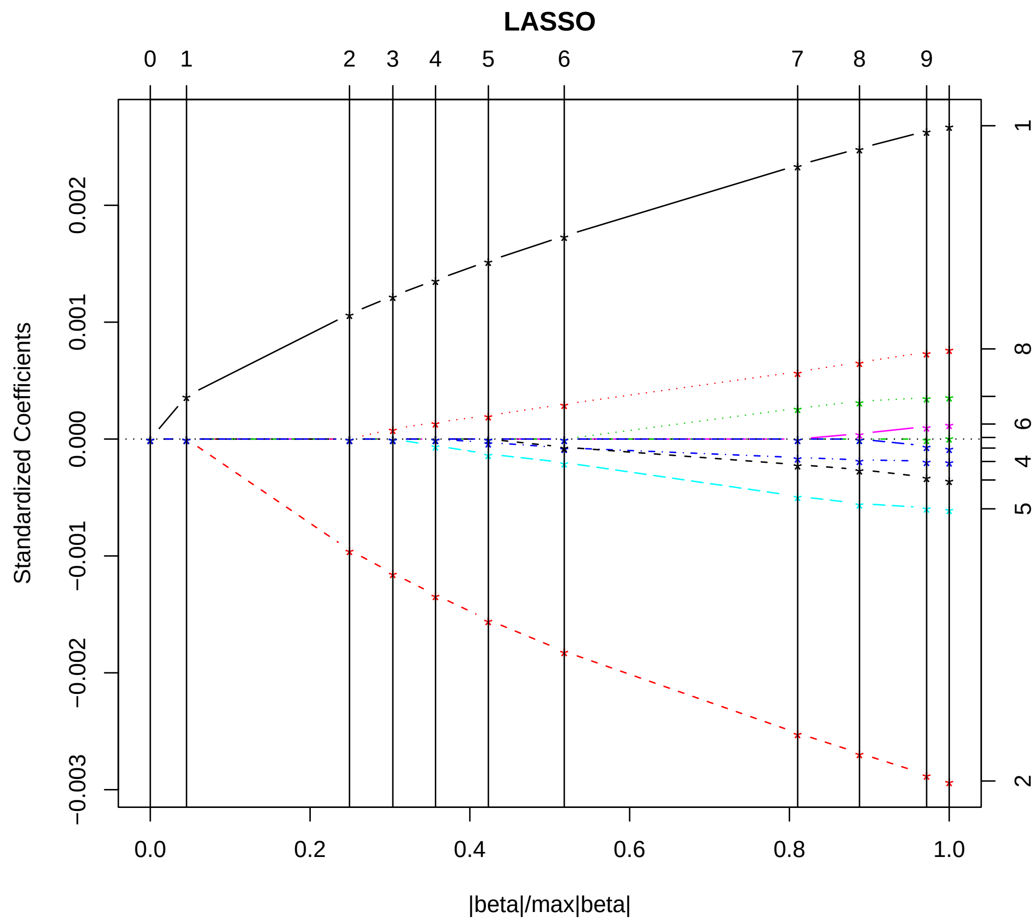




2. Examples revisited

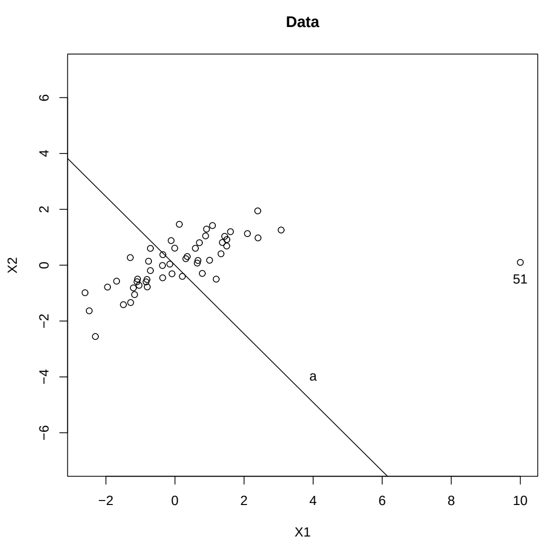
10 dimensions

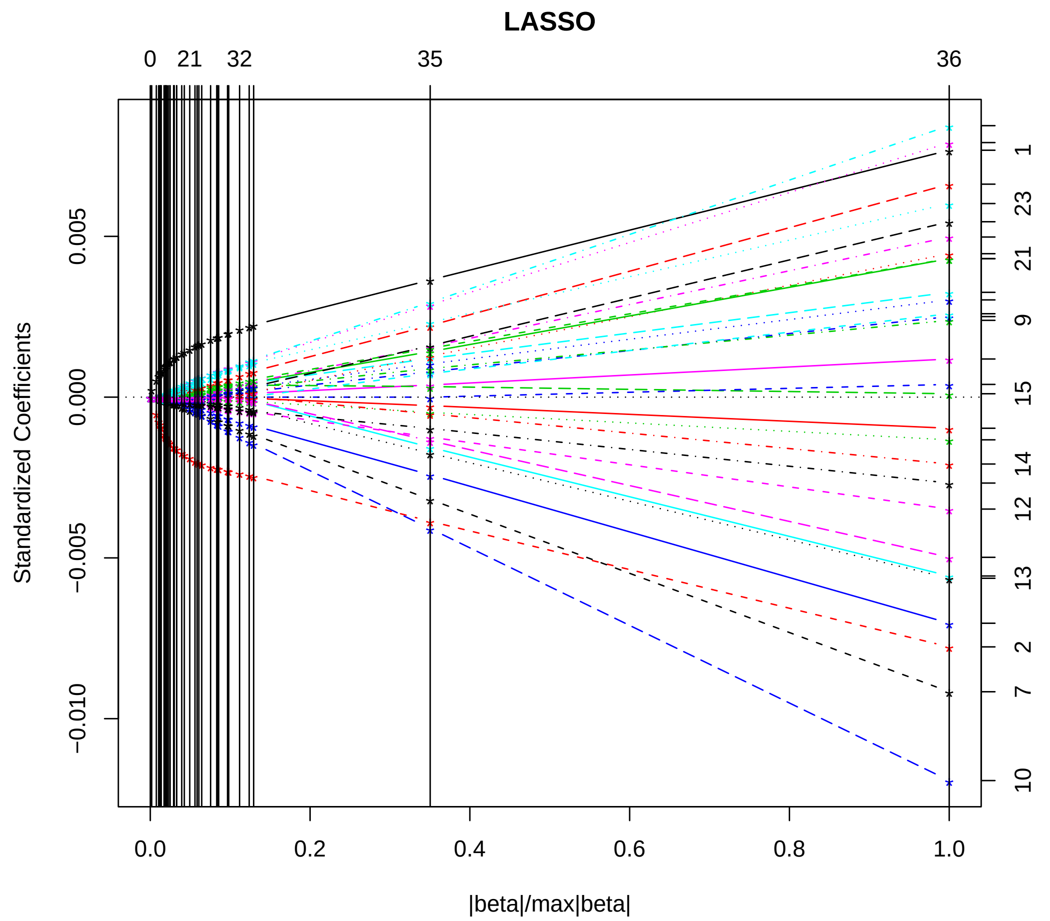




2. Examples revisited

30 dimensions





2. Forward versus backward

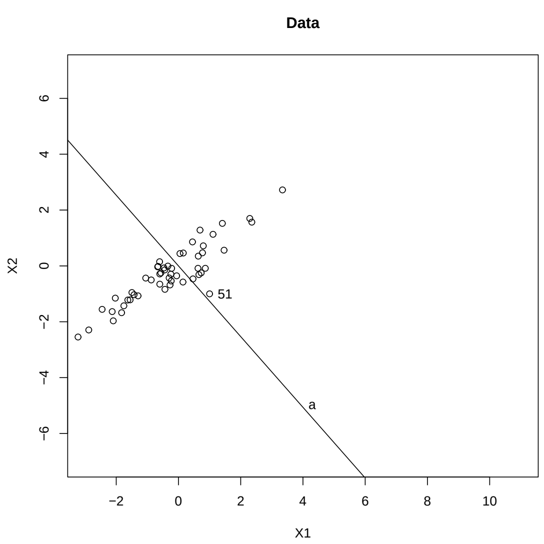
LASSO is essentially a forward method: starting from scratch variables are added to the model.

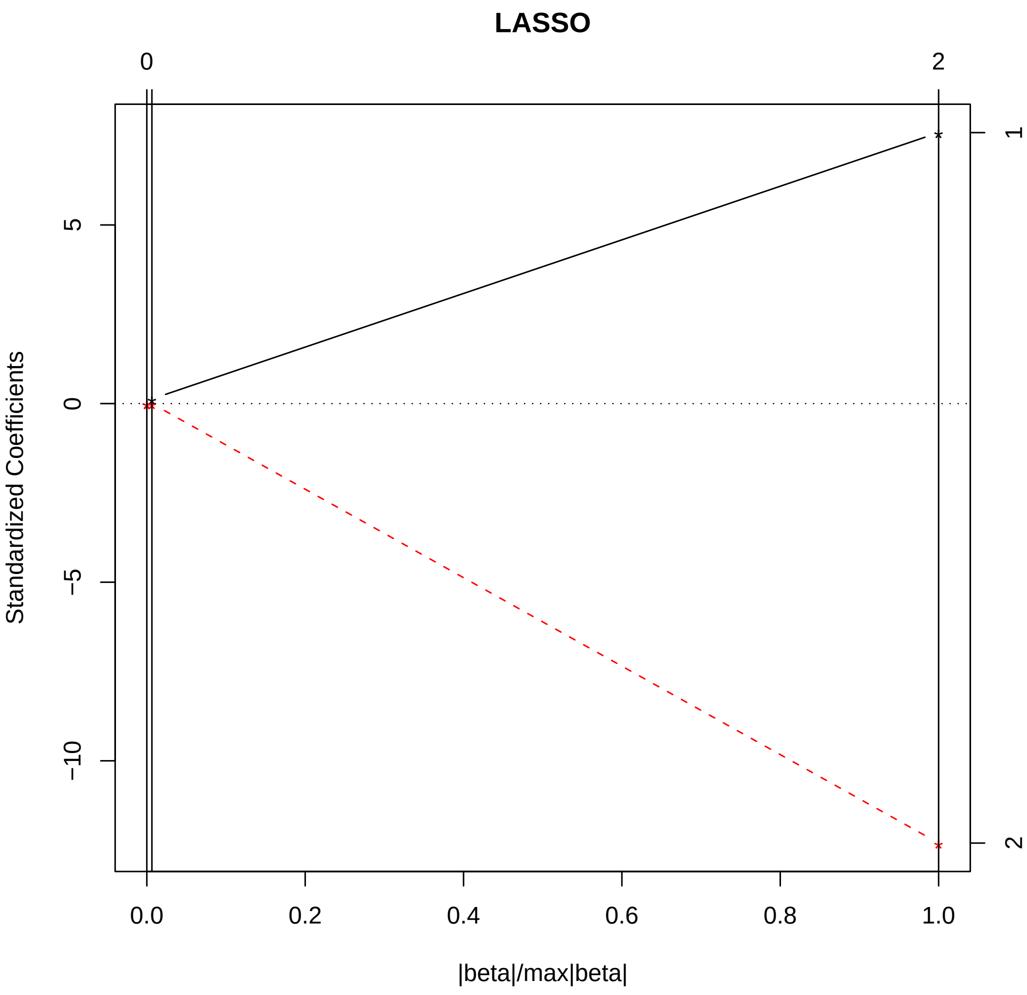
This might lead to difficulties in situations where variables only contribute to the outlyingness in combination with other highly correlated variables.

In that case the simple backward approach might be better.

2. Forward versus backward

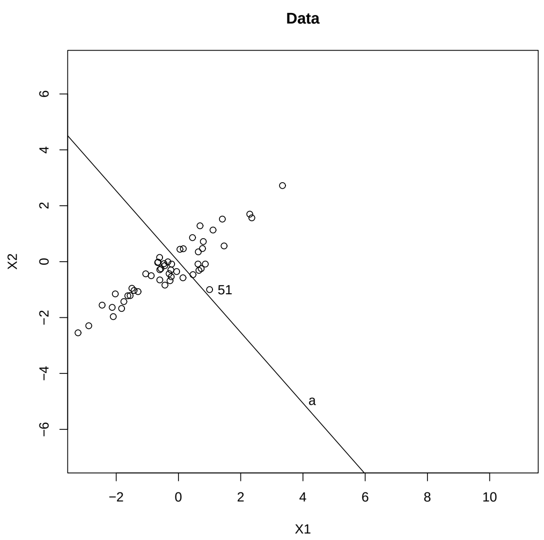
2 dimensions

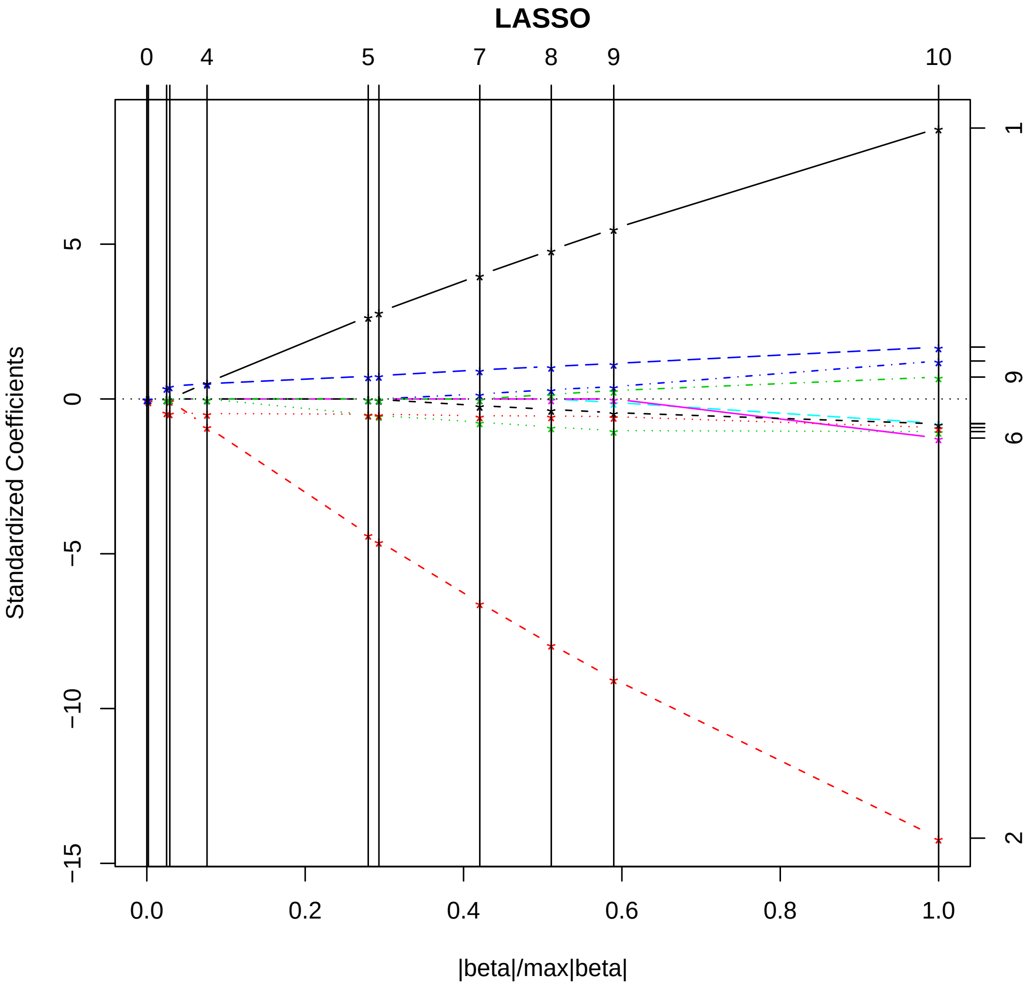




2. Forward versus backward

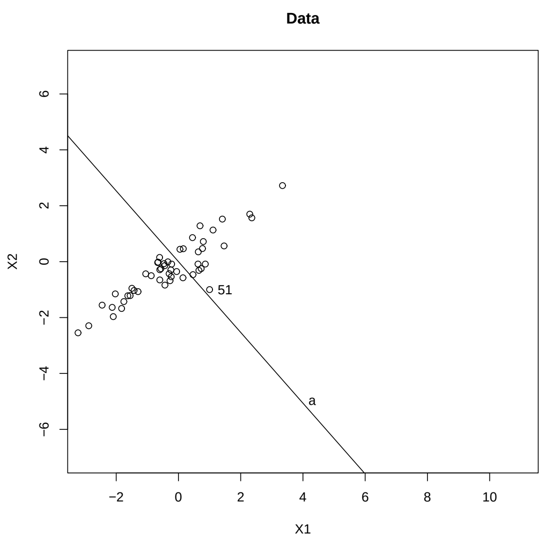
10 dimensions

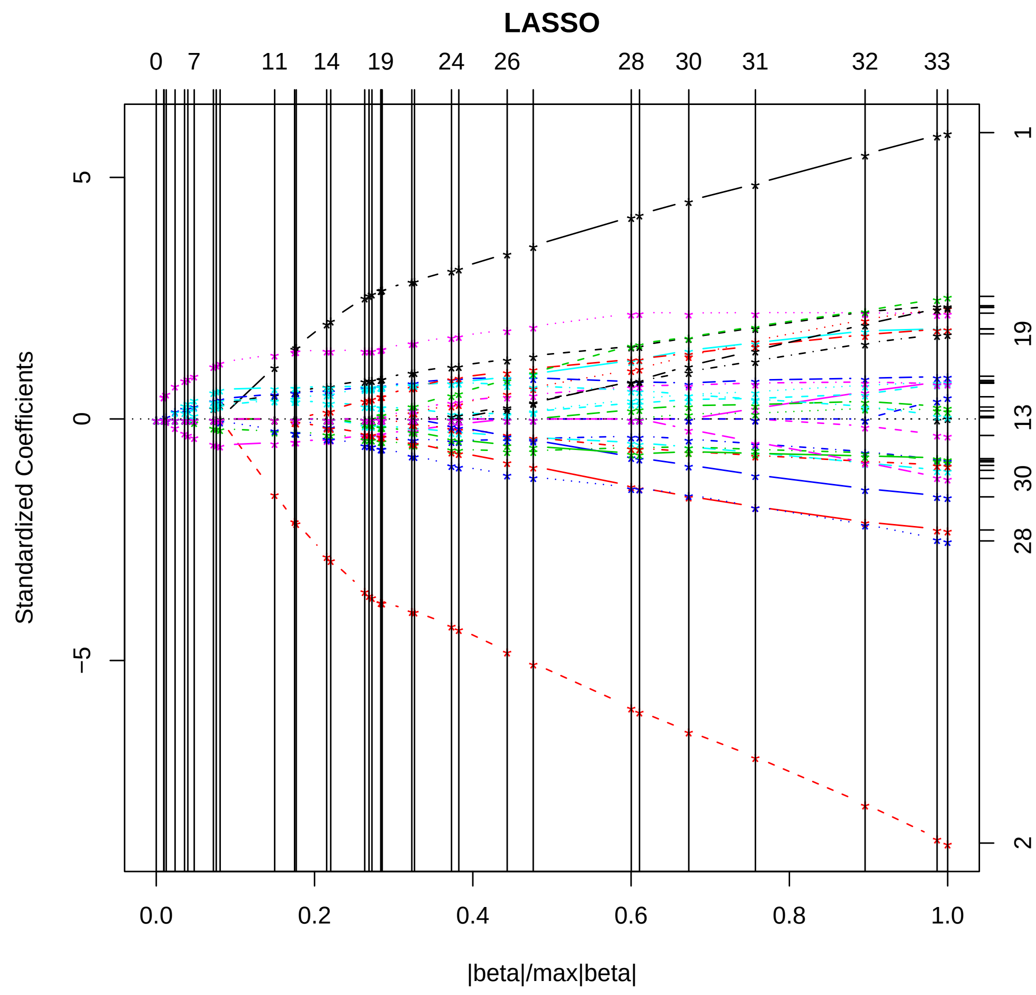




2. Forward versus backward

30 dimensions





3. An algorithm in moderate dimensions

If $p < n$ we can combine the forward and backward approach to select $k << p$ variables contributing most to the outlyingness of $\boldsymbol{x}_i$.

1. Compute the full LASSO path.
2. For $j \in \{0, \ldots, k\}$
   Let $\mathcal{S}_j$ be the set of the j variables taken first into the model by LASSO and the $k - j$ variables with largest coefficients in the unregularized solution.
3. Retain the set $\mathcal{S}_j$ for which the robust Mahalanobis distance of $\boldsymbol{x}_i$ is the largest.

This turns out to work very well.

3. An algorithm in high dimensions

If $p > n$ a backward approach is impossible, so the previous algorithm cannot be used.

An interesting extension of the LASSO is the elastic net (Zou, Hastie, 2005) adding an additional L_2 type penalty. This can be useful e.g. in data with a lot of correlation between the variables.

1. Compute the path

$$a(t) = \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|} \text{ with } \boldsymbol{\theta}(t) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{y}_w - X_w\boldsymbol{\beta}\|^2 + \lambda_j\|\boldsymbol{\beta}\|^2 \quad \text{subject to } \sum_{j=1}^{p} |\boldsymbol{\theta}_j| \leq t.$$

   Let $\mathcal{S}_j$ be the set of k variables selected by this elastic net for λ_j, $j = 1, \ldots, M$.

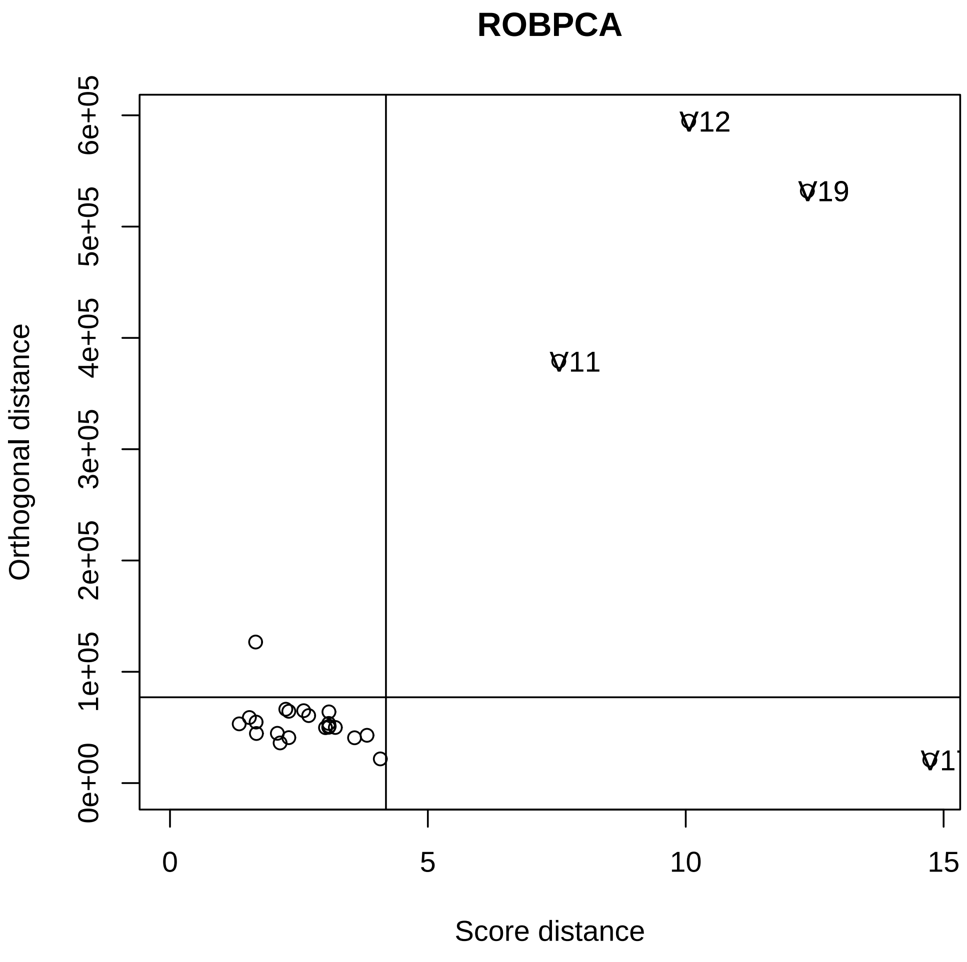
2. Select the set $\mathcal{S}_j$ for which the outlyingness of $\boldsymbol{x}_i$ is the largest.

4. Example

The breast cancer data set by West et al. (2001) contains $p = 7129$ gene expression profiles for 49 breast cancer patients. There are 25 ER+ cases and 24 ER- cases. Here we only consider the ER+ cases.

A robust PCA algorithm reveals 4 outliers.

4. Example



ROBPCA
Orthogonal distance
Score distance
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
0
5
10
15
V12
V19
V11
V17

4. Example

The breast cancer data set by West et al. (2001) contains $p = 7129$ gene expression profiles for 49 breast cancer patients. There are 25 ER+ cases and 24 ER- cases. Here we only consider the ER+ cases.

A robust PCA algorithm reveals 4 outliers.

For each outlier we can search for the 10 genes that are contributing most to its outlyingness.

4. Example

The breast cancer data set by West et al. (2001) contains $p = 7129$ gene expression profiles for 49 breast cancer patients. There are 25 ER+ cases and 24 ER- cases. Here we only consider the ER+ cases.

A robust PCA algorithm reveals 4 outliers.

For each outlier we can search for the 10 genes that are contributing most to its outlyingness.

For 11, 12 and 19 we find genes that have no immediate biological interpretation.

For 17 it turns out that 6 out of 10 selected variables also appear in the list of 20 genes by West et al. most relevant for differentiating between ER+ and ER-.

4. Example

The breast cancer data set by West et al. (2001) contains $p = 7129$ gene expression profiles for 49 breast cancer patients. There are 25 ER+ cases and 24 ER- cases. Here we only consider the ER+ cases.

A robust PCA algorithm reveals 4 outliers.

For each outlier we can search for the 10 genes that are contributing most to its outlyingness.

For 11, 12 and 19 we find genes that have no immediate biological interpretation.

For 17 it turns out that 6 out of 10 selected variables also appear in the list of 20 genes by West et al. most relevant for differentiating between ER+ and ER-.

This confirms West et al.: for 11, 12 and 19 array hybridization failed, whereas 17 is a mislabeled observation.

5. Conclusion

Summary:

- Given a robust procedure that detects outliers. How to select variables most relevant for the outlyingness of an outlier?
- The direction of maximal outlyingness is a normed solution of a least squares problem. By adding a LASSO type penalty a regularized path of sparse directions can be defined.
- In moderate dimensions:
  - Graphical display.
  - An automatic algorithm is proposed combining forward and backward selection.
- In high dimensions:
  - Elastic net.
  - Essentially forward.