Sparse and Robust Methods for Discrimination in High Dimensions

V. Todorov[1] P. Filzmoser[2]

[1]United Nations Industrial Development Organization (UNIDO)

[3]Vienna University of Technology

Österreichische Statistiktage 2011
Graz 7-9 September, 2011

Outline

Classification in High Dimensions

Feature Selection and Sparse Methods

Robust Algorithms for Classification

Examples with Real Data Sets

Simulation Study

Summary and Conclusions

1 Classification in High Dimensions

2 Feature Selection and Sparse Methods

3 Robust Algorithms for Classification

4 Examples with Real Data Sets

5 Simulation Study

6 Summary and Conclusions

Introduction

... the main goals of high dimensional regression and classification are:

- to construct as effective a method as possible to predict future observations;
- to gain insight into the relationship between features and response for scientific purposes, as well as, hopefully, to construct an improved prediction model.

P. J. Bickel (2008)

Introduction

Sparse and Robust Methods for Discrimination in High Dimensions

Todorov, Filzmoser

Classification in High Dimensions

Feature Selection and Sparse Methods

Robust Algorithms for Classification

Examples with Real Data Sets

Simulation Study

Summary and Conclusions

- Classification (or discrimination) is a supervised technique (as opposed to the automatic identification of group structure or cluster analysis which are unsupervised techniques).
- Conventional classification methods - inapplicable (or produce poor results when applied) to high dimensional data - $n \ll p$
- The case with $n \ll p$ is not anomalous but rather it is the generic one - examples in different research applications ...

Algorithms for Classification

DIMENSION REDUCTION + LOW-DIM. CLASSIFIER, E.G. LDA

- **PCA**: Principal component analysis - linear dimension reduction method based on decomposition of the empirical covariance matrix of the data matrix **X**.
- **PLS**: Partial Least Squares (Wold, 1966, 1975) - linear dimension reduction method based on decomposition of the empirical cross-covariance matrix of both the data matrix **X** and the grouping variable **y**.
- Other methods
  - **SIMCA**: Soft independent modeling of class analogues (Wold, 1976) - incorporates PCA for dimension reduction in each group separately
  - **SIR**: Sliced inverse regression (Duan and Li, 1991; Li, 1991)
  - **LDA PP**: Projection pursuit LDA (Pires and Branco, 2010)
  - Feature selection

Sparse and Robust Methods for Discrimination in High Dimensions

Todorov, Filzmoser

Classification in High Dimensions

Feature Selection and Sparse Methods

Robust Algorithms for Classification

Examples with Real Data Sets

Simulation Study

Summary and Conclusions

PLS-DA: Partial Least Squares for Discrimination

- PLS regression is used to predict a set of dependent variables **Y** from a (very) large set of predictors **X** and to describe their common structure
- Preferred algorithm is SIMPLS - more consistent with multivariate statistics (solving eigenstructure problems), fast, results are easier to interpret
- Inherently not designed for discrimination but routinely used (if more than two groups, the grouping variable is transformed to dummy binary variables)
- Alternatively classical discrimination method (logistic regression, LDA, QDA) can be applied to the PLS components (Nguyen and Rocke, 2002; Boulesteix, 2004)

Why should PLS be preferred to PCA

PLS-DA OR PCA?

- **PLS** - the dimension reduction is guided explicitly by between-group variability
- **PCA** - does not take into account the group structure (considers only the total variability but not the within-groups and between-groups variability)
- When the within-groups variability dominates the between-groups variability, PLS will necessarily perform better (than PCA)
- Formal statistical explanation of the relation of PLS to LDA exists (Barker and Rayens, 2002) which suggests that in discrimination context PLS is the preferred one in presence of high collinearity

Feature Selection and Sparse Methods

- Although PLS can deal with more predictors than samples often variable filtering/selection is utilized as a preprocessing step.
- Chun and Keles (2010) show that the performance of PLS is ultimately affected by the large number of predictors (existence of high number of irrelevant variables leads to inconsistency of the parameter estimates).

Sparse and Robust Methods for Discrimination in High Dimensions

Todorov, Filzmoser

Classification in High Dimensions

Feature Selection and Sparse Methods

Robust Algorithms for Classification

Examples with Real Data Sets

Simulation Study

Summary and Conclusions

Feature Selection techniques

- Two sample t-statistic commonly used in binary classification (Nguyen and Rocke, 2002)
- Ad-hoc robust version by replacing the means by medians and standard deviations by MADs (Pires and Branco, 2010)
- Ratio of between-sum-of-squares to within-sum-of-squares (BSS/WSS) in multi-category classification (Boulesteix, 2004)
- Ordering of the variables by the absolute values of the coefficients for the first PLS component - equivalent to the above (Boulesteix, 2004)

Sparse and Robust Methods for Discrimination in High Dimensions

Todorov, Filzmoser

Classification in High Dimensions

Feature Selection and Sparse Methods

Robust Algorithms for Classification

Examples with Real Data Sets

Simulation Study

Summary and Conclusions

FEATURE SELECTION TECHNIQUES

... But:

- All are univariate and ignore the correlation between variables
- No established way of deciding what number of top ranking variables to take

Sparse and Robust Methods for Discrimination in High Dimensions

Todorov, Filzmoser

Classification in High Dimensions

Feature Selection and Sparse Methods

Robust Algorithms for Classification

Examples with Real Data Sets

Simulation Study

Summary and Conclusions

SPARSE PLS

Variable selection within the course of PLS dimension reduction

- easy interpretation
- correlations among covariates are considered
- computationally efficient
- a tunable sparsity parameter

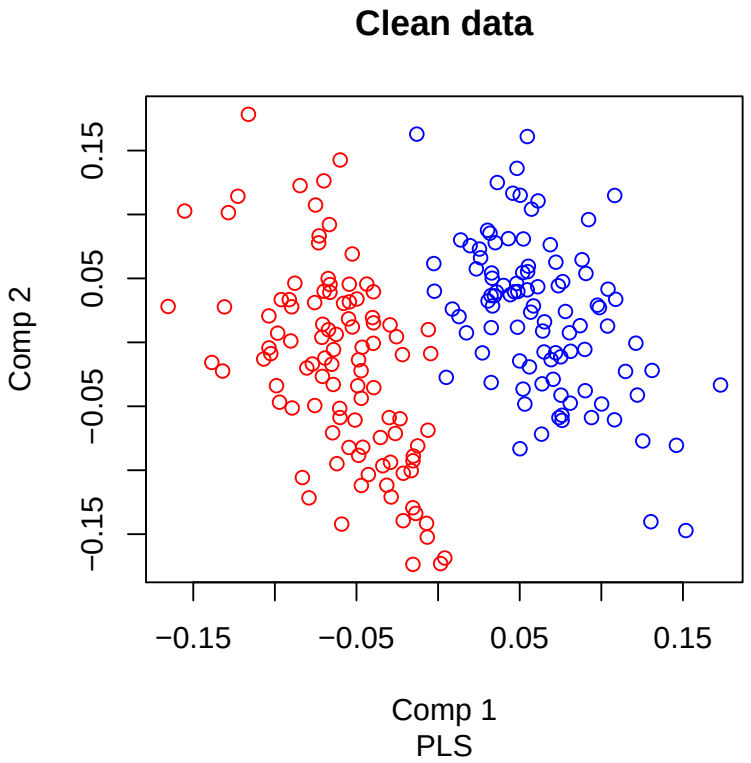
- Le Cao et al (2008) - R package **mixOmics**
- Chung and Keles (2010) - R package **spls**

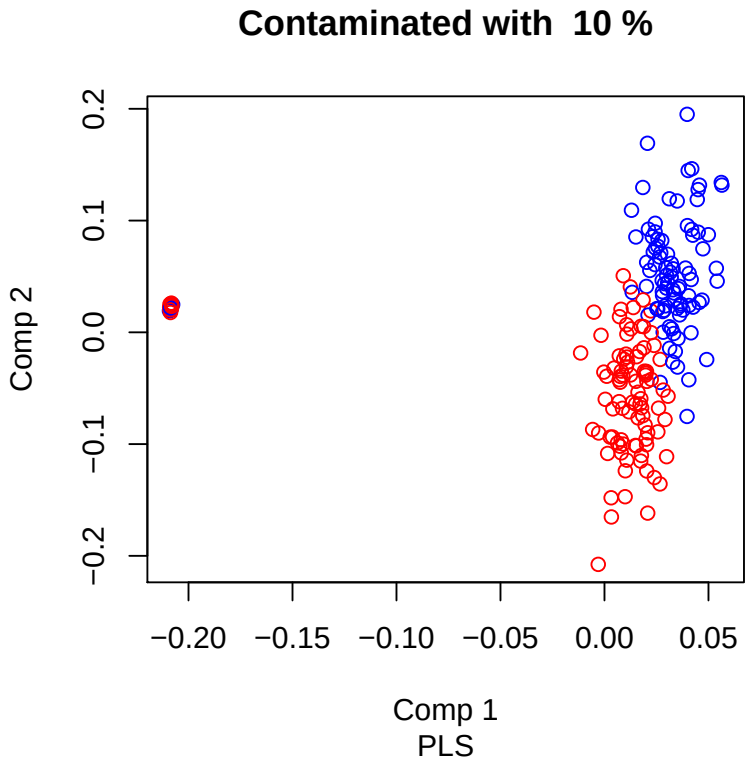
Outliers in Classification Context

- All the methods considered so far are based on a classical covariance matrix and thus will be very sensitive to the presence of outliers in the data
- In classification context we could encounter two main types of outliers
  1. An observation that belongs to a different group present in the data (mislabeled observation, incorrectly classified). These are relatively easily recognized by classification methods
  2. Abnormal observations that do not belong to any group present in the data (can be caused by many reasons)

Simulated outliers and PLS-DA

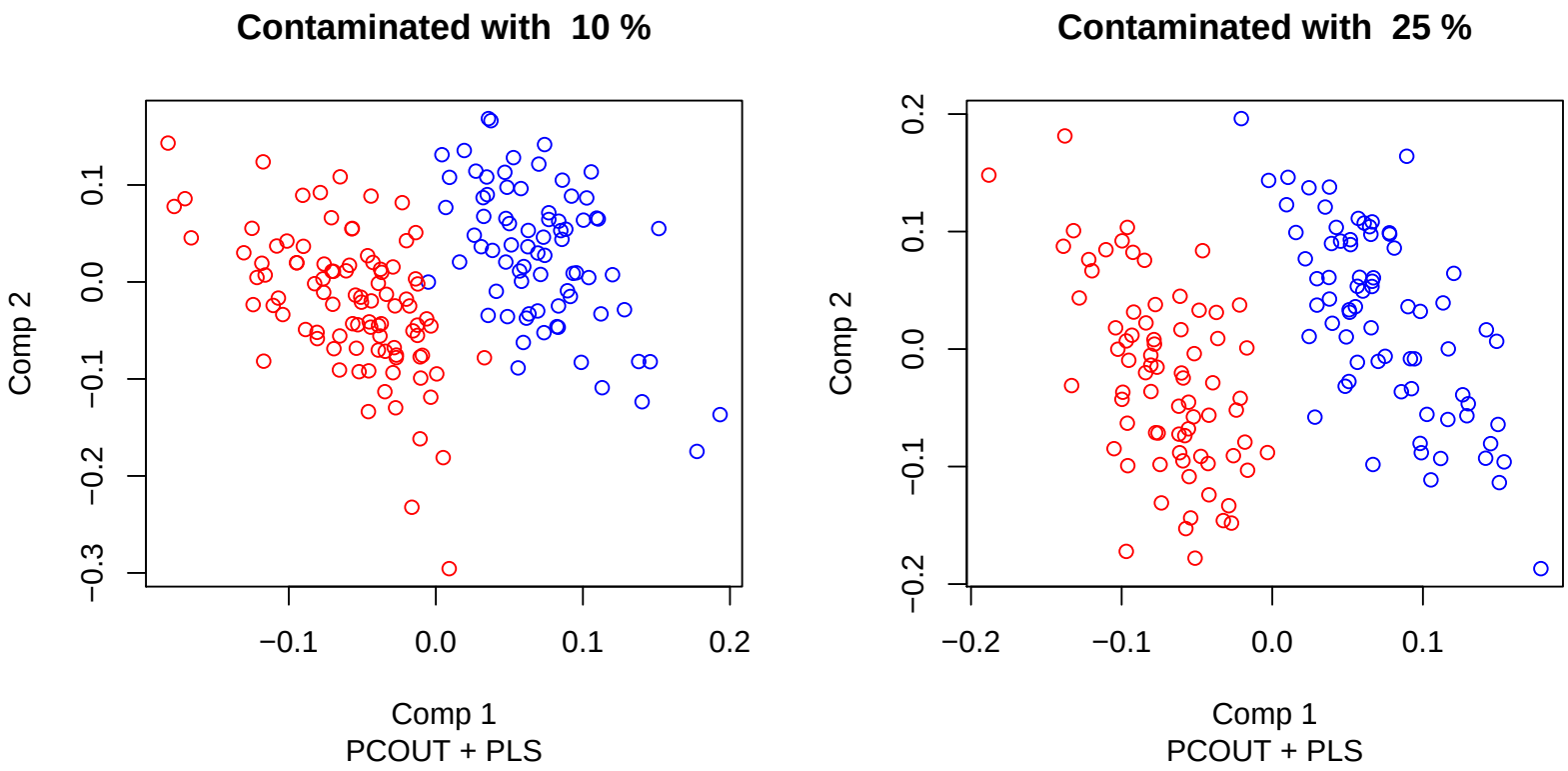
FIRST TWO PLS COMPONENTS





Simulated outliers - Robust PLS-DA

First two Robust-PLS components



Robust PLS

Sparse and Robust Methods for Discrimination in High Dimensions

Todorov, Filzmoser

Classification in High Dimensions

Feature Selection and Sparse Methods

Robust Algorithms for Classification

Examples with Real Data Sets

Simulation Study

Summary and Conclusions

- There are many attempts for robustification of PLS but most of them are either not resistant to leverage points or cannot handle large data sets with $n \ll p$
- Two methods could be considered
  1. Hubert and Vanden Branden (2003)
     - Based on robust covariance matrices for high dimensional data (Hubert et al, 2005)
     - The data matrix contains a categorical variable and the robust covariance matrix cannot be computed

     NOT APPLICABLE FOR DISCRIMINATION
  2. Robust partial M-regression (Serneels et al,2005)
     - Computed by a modification of the IRPLS algorithm (Wold, 1973)
     - Does not explicitly compute a covariance matrix but nevertheless

     NOT APPLICABLE FOR DISCRIMINATION

so far no robust PLS-DA available

Robust PLS-DA and SPLS-DA

Sparse and Robust Methods for Discrimination in High Dimensions

Todorov, Filzmoser

Classification in High Dimensions

Feature Selection and Sparse Methods

Robust Algorithms for Classification

Examples with Real Data Sets

Simulation Study

Summary and Conclusions

ROBUST PLS-DA: OUTLIER DETECTION + PLS/SPLS

- A three-step procedure:
  1. Identify the outliers that are possibly present in the data (using a robust method)
  2. Perform classical PLS or sparse PLS to reduce the dimensionality
  3. Apply robust LDA (or an other robust classifier) to the low dimensional data
- We need outlier detection methods which work in high dimensions
  1. PCOUT: Filzmoser et al. (2008)
  2. PCDIST: Shieh and Hung (2009)
  3. SIGN1: Locantore et al. (1999), Filzmoser et al. (2008)

Robust PLS-DA and SPLS-DA

PCOUT: Filzmoser et al. (2008)

A two phase procedure targeting different types of outliers in each phase; utilizes inherent properties of PCA

- Phase 1: Location outliers
  1. Robustly sphere the data - $\mathbf{X}^*$ - and calculate $\mathbf{C} = Cov(\mathbf{X}^*)$
  2. PCA decomposition of $\mathbf{C}$ (semi-robust). Retain p^* components ($\geq$ 99% of the total variance)
  3. Compute robust kurtosis weights for the components
  4. Determine weights w^L using T-Biweight function (Rocke, 1996)
- Phase 2: Scatter outliers
  1. Robustly sphere the data - as in Step 2 above, obtain (transformed) Euclidean distances in PC space
  2. Determine weights w^S using T-Biweight function
- Combine the weights w^L and w^S

Robust PLS-DA and SPLS-DA

PCDIST: Shieh and Hung (2009)
Specifically developed for microarray data - thousands of variables; a large number of the genes are non informative, i.e. many of the dimensions are meaningless; contain class information

- Treat each class independently and try to find outliers with regard to each class.
  1. Dimension reduction by classical PCA
  2. Automatic selection of the number of PC components to retain (Zhu and Ghodsi, 2006)
  3. Compute S-estimates $(\mathbf{T}, \mathbf{C})$ of location and scatter in the selected low-dimensional space
  4. Use $(\mathbf{T}, \mathbf{C})$ to construct robust distances and compare them to a quantile of the χ^2 distribution $(\chi^2_{p,0.975})$

Robust PLS-DA and SPLS-DA

SIGN1: Locantore et al. (1999), Filzmoser et al. (2008)
Based on a type of robust PCA (spherical)

- Project the data on to a sphere. Thus the effects of outlying observations in the data is limited (they would be placed on the boundary of the sphere)
- Standard PCA on the projected data(without the undue influence of any outlier)
- The resulting mean and covariance matrix are robust. Use them to compute robust distances
- Use the 0.975 quantil of the χ^2 distribution as an outlier cutoff ($\chi^2_{p,0.975}$).

Examples with Real Data Sets

Breast Cancer Data: van't Veer et al. (2002)

- A study of expression levels of $p = 22483$ genes from 78 samples of breast cancer patients: 44 have good prognosis and 34 poor.
- A second (test) data set contains 19 samples: 7 have good prognosis and 12 poor.
- The data were filtered as described in the original article - remain only 4348 genes.
- No known outliers

Breast Cancer Data

Sparse and Robust Methods for Discrimination in High Dimensions

Todorov, Filzmoser

Classification in High Dimensions

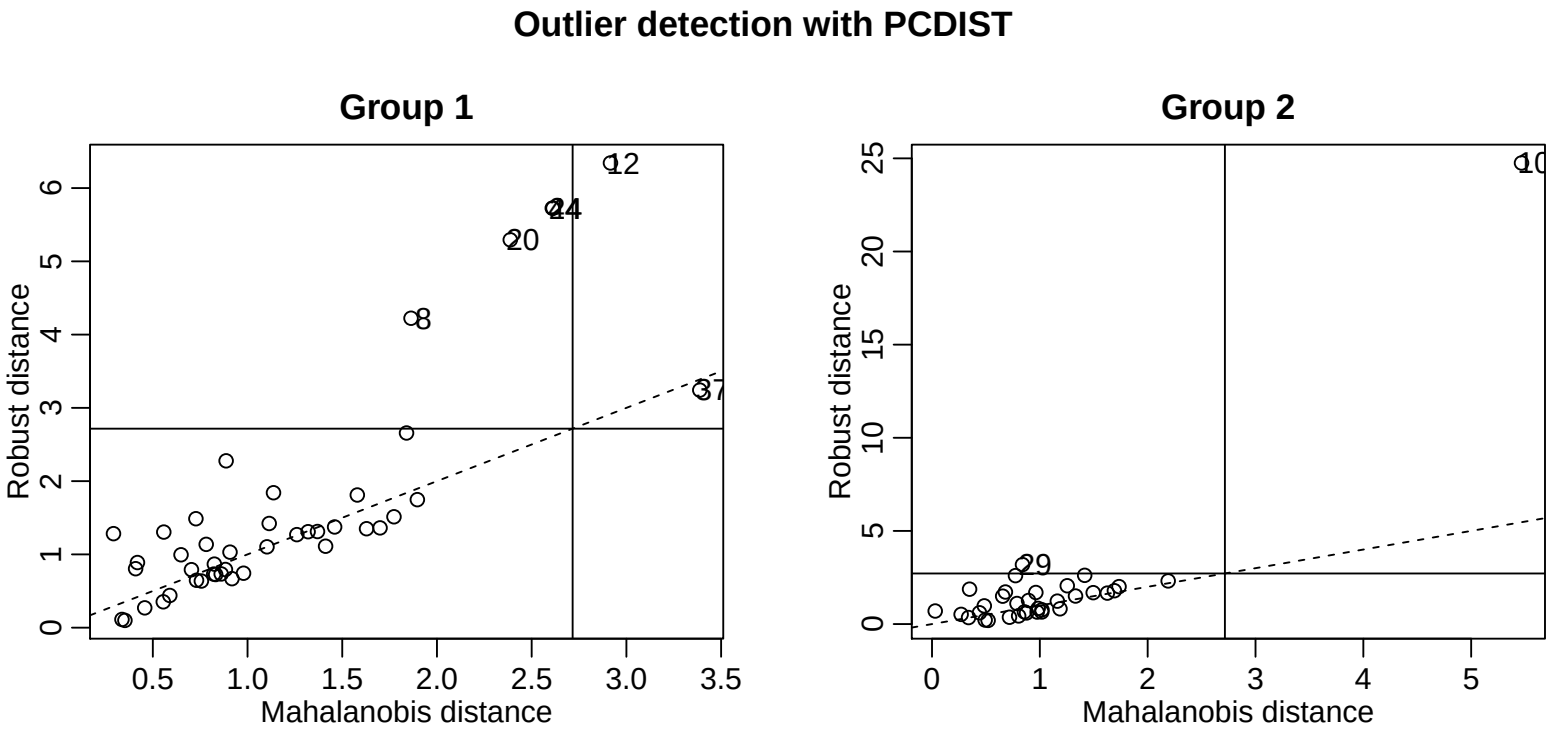
Feature Selection and Sparse Methods

Robust Algorithms for Classification

Examples with Real Data Sets

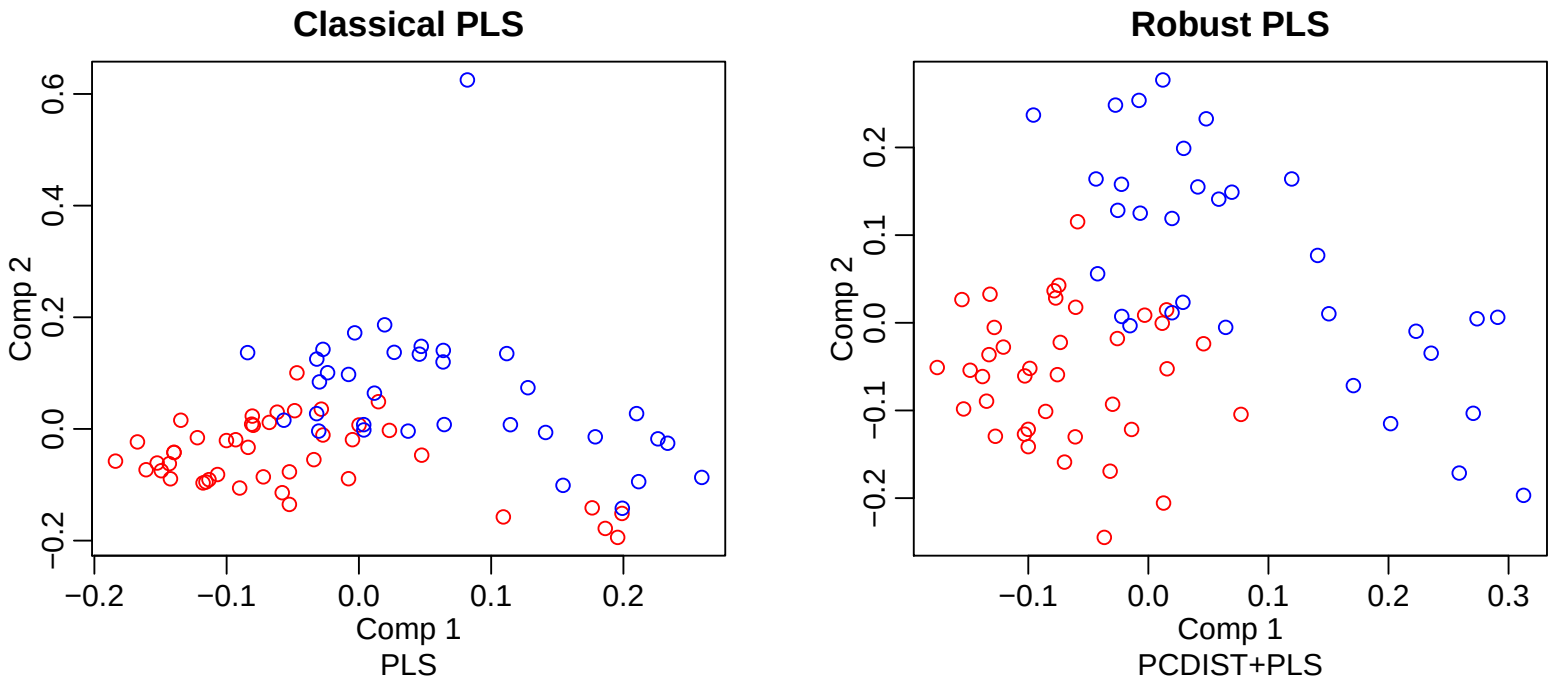
Simulation Study

Summary and Conclusions



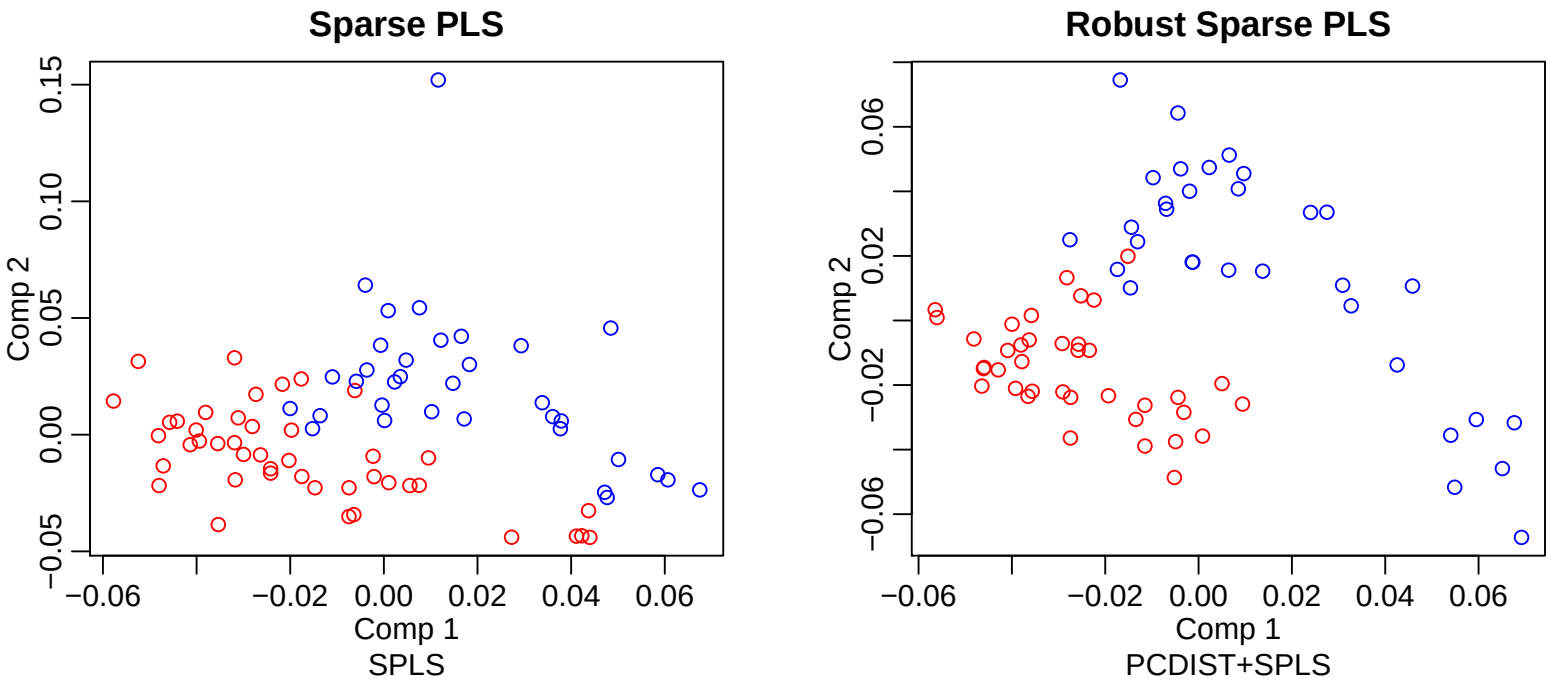
Breast Cancer Data

Improved groups separation through robust PLS



Breast Cancer Data

Improved sroups separation through sparse robust PLS



Breast Cancer Data

CV Misclassification Errors for the different estimators and 1 to 6 components

	1	2	3	4	5	6
PLS	36.84	26.32	26.32	26.32	36.84	36.84
PCOUT	18.75	25.00	43.75	43.75	37.50	43.75
PCOUT01	12.50	18.75	37.50	37.50	37.50	37.50
PCDIST	16.67	25.00	25.00	33.33	33.33	25.00
SPLS	21.05	21.05	21.05	36.84	31.58	26.32
SPCOUT	12.50	25.00	31.25	12.50	43.75	37.50
SPCOUT01	37.50	31.25	25.00	25.00	31.25	31.25
SPCDIST	16.67	**8.33**	25.00	16.67	25.00	16.67
PCA	42.11	42.11	36.84	42.11	31.58	31.58
HUBERT	25.00	25.00	25.00	25.00	31.25	25.00
LOCANTORE	50.00	25.00	25.00	25.00	25.00	25.00

Simulation Setup I

DATA STRUCTURE

- Data constructed for binary classification as follows:
- Three latent variables H_1, H_2 and H_3 from $N(0, 5^2)$.
- A matrix of covariates **X** scaled to zero mean and unit variance with
  - $x_1, \ldots, x_5 = H_1 + N(0, 1)$
  - $x_6, \ldots, x_{10} = H_2 + N(0, 1)$
  - $x_{11}, \ldots, x_{15} = H_3 + N(0, 1)$
  - $x_{16}, \ldots, x_p = N(0, 1)$
- A response variable **Y** such that
  - $prob = P(Y = 1|H_1, H_2, H_3) = g(3H_1 - 14H2)$ where g is the inverse of the link function (the cumulative distribution function for the logistic distribution).
  - $Y \sim Bernoulli(prob)$

Sparse and Robust Methods for Discrimination in High Dimensions

Todorov, Filzmoser

Classification in High Dimensions

Feature Selection and Sparse Methods

Robust Algorithms for Classification

Examples with Real Data Sets

Simulation Study

Summary and Conclusions

GROUPS AND OUTLIERS

- Training set consisting of two groups with size $n = n_1 + n_2 = 160$
- Outliers are generated by replacing εn observations by $\mathbf{Y} \sim N(\mathbf{b}_p, \kappa \mathbf{I}_p)$ with $\mathbf{b} = (b, \ldots, b,\ b = 4,\ \kappa = 0.1$ and $\varepsilon = \{0, 0.1, 0.25\}$. Equal percentage of outliers in both groups.
- Clean (outlier free) validation data set with the same structure as the training one and group size proportional to the group size of the training data set

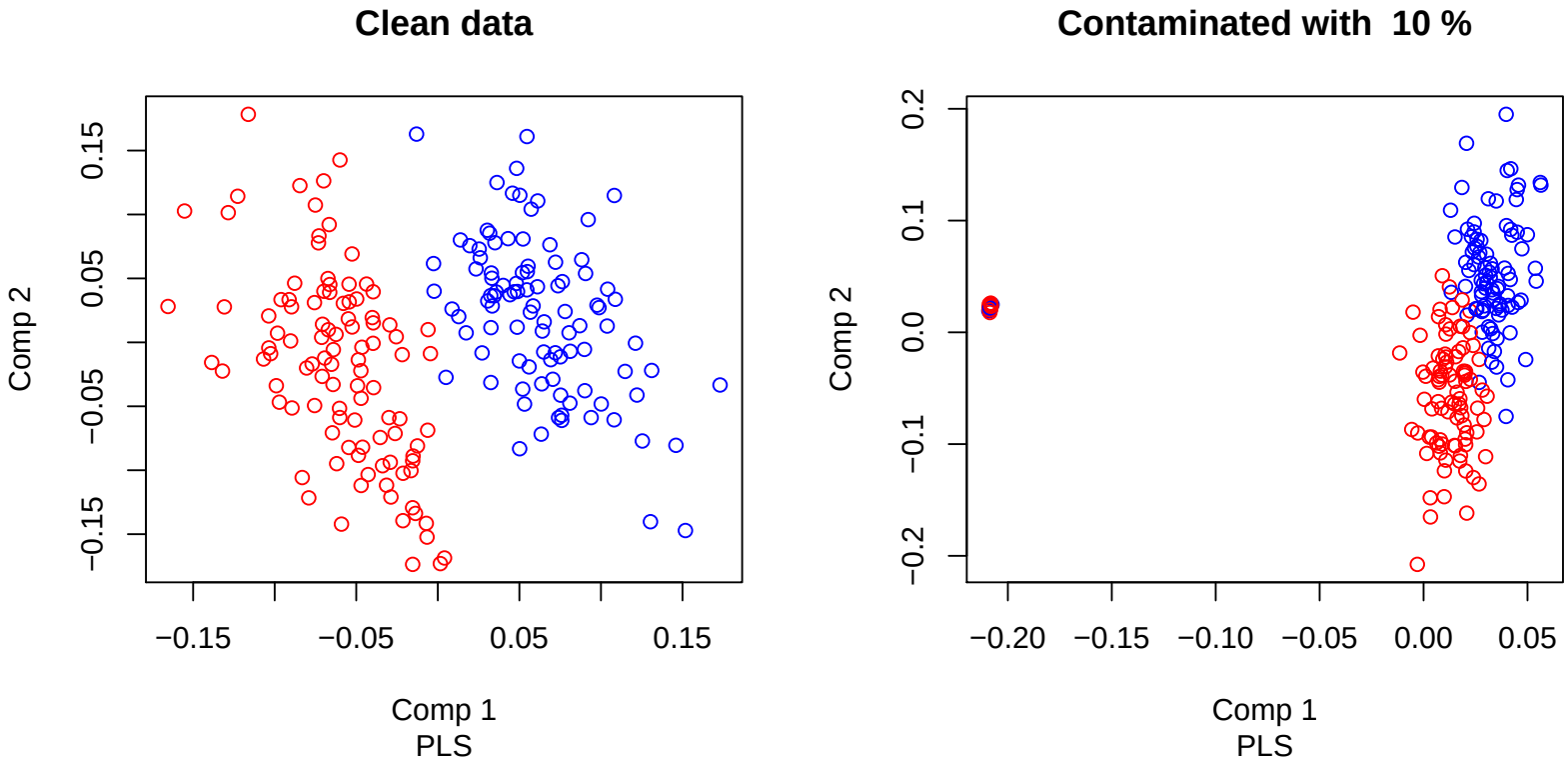
Simulation Setup III

Methods and evaluation

- Methods: **PCA, PLS, PCDIST, PCOUT, PCOUT01, SPLS, SPCDIST, SPCOUT, SPCOUT01**
- For each data set each method is run and the misclassification error is estimated. This is repeated $m = 100$ times and the average misclassification error is calculated

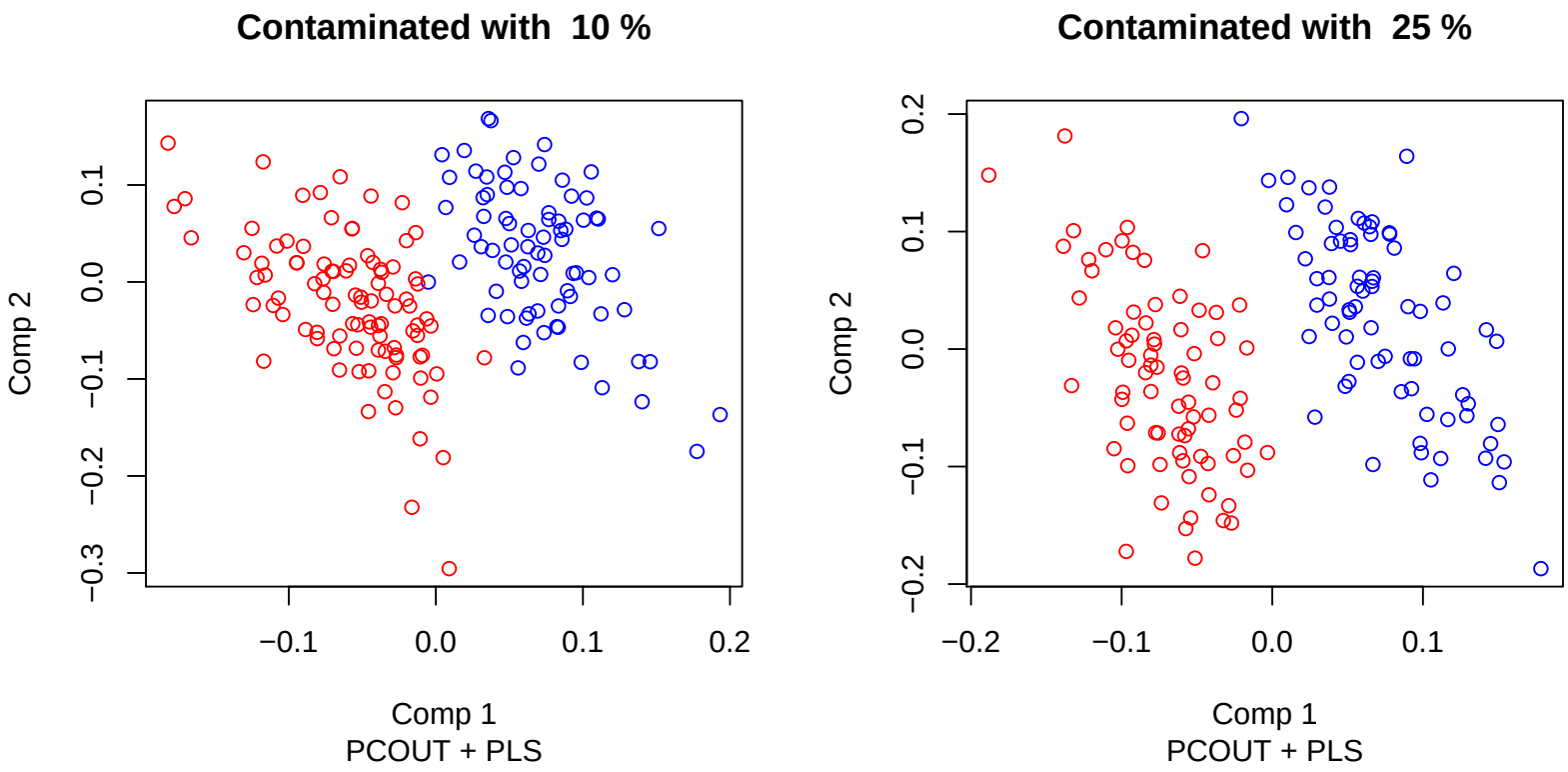
Simulation Setup IV

First two PLS components



Simulation Setup V

First two Robust-PLS components



Simulation results

Average Misclassification Errors for different levels of contamination

	0%	10%	25%
PLS	8.25	19.40	20.70
PCOUT	8.83	9.57	9.69
PCOUT01	8.75	9.32	12.42
PCDIST	8.79	9.01	10.62
SPLS	**6.46**	38.15	45.58
SPCOUT	8.21	6.62	**7.15**
SPCOUT01	6.78	**6.47**	12.46
SPCDIST	6.83	6.69	8.24
PCA	8.22	19.41	20.04
HUBERT	8.41	8.91	9.76
LOCANTORE	8.25	7.91	9.33

Conclusions and Outlook

- We considered methods for discrimination in **high dimensional data sets**. In this context the following new methods were proposed:
  - **PCOUT-PLS**, **PCDIST-PLS** and **SIGN1-PLS**: Outlier detection followed by classical PLS
  - **PCOUT-SPLS**, **PCDIST-SPLS** and **SIGN1-SPLS**: Outlier detection followed by sparse PLS
- The methods were compared in terms of computation time and discrimination performance on example data sets and simulation study
- The considered methods are implemented in an `R` package `rrcovHD` $\Rightarrow$ soon available on CRAN.
- The considered HD data sets (and many more) are available in the `R` package `rrcovHDData` (not available on CRAN).

References I

Shieh AD, Hung YS.
Detecting Outlier Samples in Microarray Data. *Statistical Applications in Genetics and Molecular Biology*. 8(1), 2009

H. Chun and S. Keles
Sparse partial least squares regression for simultaneous dimension reduction and variable selection,
Statistical Applications in Genetics and Molecular Biology, **9**, Article 17, 2010.

M. Hubert, P.J. Rousseeuw and S. van Aelst
High-Breakdown Robust Multivariate Methods,
Statistical Science, **23**, 92–119, 2008.

V. Todorov and P. Filzmoser
Multivariate Robust Statistics: Methods and Computation, Südwestdeutscher Verlag für Hochschulschriften, Saarbrüucken, 2009