Session 8: High dimensional data and sparsity
Winter course, CMStatistics 2016

Mia Hubert, Peter Rousseeuw, Stefan Van Aelst

Department of Mathematics
KU Leuven, Belgium

December 6–7, 2016

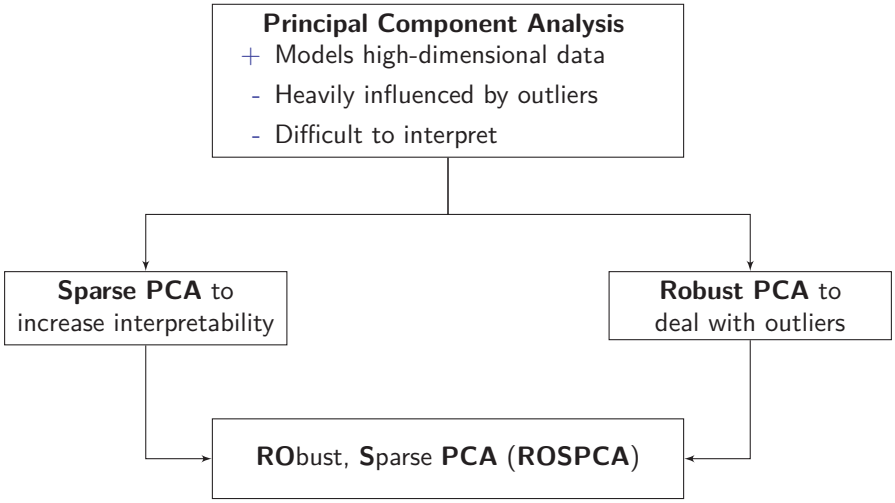
Outline of the course

- 1. General notions of robustness
- 2. Robustness for univariate data
- 3. Robust multivariate methods
- 4. Robust regression
- 5. Robust principal component analysis
- 6. Inference
- 7. Multivariate and functional depth
- 8. High dimensional data and sparsity
- 9. Cellwise outliers

High dimensional data and sparsity: Outline

1. Classical PCA
2. Sparse PCA
3. Robust PCA
4. Sparse robust PCA
5. Robust sparse PCA

Principal Component Analysis



Classical PCA

PCA searches for uncorrelated linear combinations of the original variables capturing most of the covariance structure of the original data.

For data $\boldsymbol{X}$ and loadings $\boldsymbol{P}$ ($\boldsymbol{p}_j'\boldsymbol{p}_j = 1$),

$$\max_{\boldsymbol{p}_j} \text{Var}(\boldsymbol{X}\boldsymbol{p}_j) \text{ subject to } \text{Corr}(\boldsymbol{p}_i'\boldsymbol{x}, \boldsymbol{p}_j'\boldsymbol{x}) = 0 \;\; \forall i \neq j$$

Equivalently: the jth PCA loading is given by

$$\boldsymbol{p}_j = \begin{cases} \underset{\|\boldsymbol{p}\|=1}{\text{argmax}}\, S(\boldsymbol{p}'\boldsymbol{x}_1, \ldots, \boldsymbol{p}'\boldsymbol{x}_n) & \text{if } j = 1 \\ \underset{\|\boldsymbol{p}\|=1, \boldsymbol{p}\perp\boldsymbol{p}_1, \ldots, \boldsymbol{p}\perp\boldsymbol{p}_{j-1}}{\text{argmax}}\, S(\boldsymbol{p}'\boldsymbol{x}_1, \ldots, \boldsymbol{p}'\boldsymbol{x}_n) & \text{if } 1 < j \leqslant p, \end{cases}$$

with S the standard deviation.

Sparse PCA - SCoTLASS

The goals of sparse PCA are:

- Fit PCA loadings with many elements set to zero to improve interpretability.
- Estimate a PCA model with sufficient explanatory power.

Jolliffe et al. (2003) proposed SCoTLASS:

The jth sparse PCA direction is given by

$$\tilde{\boldsymbol{p}}_j = \begin{cases} \underset{\|\boldsymbol{p}\|=1}{\text{argmax}}\, S(\boldsymbol{p}'\boldsymbol{x}_1, \ldots, \boldsymbol{p}'\boldsymbol{x}_n) - \lambda_1\|\boldsymbol{p}\|_1 & \text{if } j = 1 \\ \underset{\|\boldsymbol{p}\|=1, \boldsymbol{p}\perp\tilde{\boldsymbol{p}}_1, \ldots, \boldsymbol{p}\perp\tilde{\boldsymbol{p}}_{j-1}}{\text{argmax}}\, S(\boldsymbol{p}'\boldsymbol{x}_1, \ldots, \boldsymbol{p}'\boldsymbol{x}_n) - \lambda_j\|\boldsymbol{p}\|_1 & \text{if } 1 < j \leqslant p, \end{cases}$$

with S the standard deviation. A higher value of λ corresponds to greater sparsity, and a value of zero corresponds to no sparsity.

SCoTLASS

The following example compares PCA and SCoTLASS on simulated data.

- First two components explain most of the variability in the data.
- Ten variables. True PC 1 loads on variables 1-4. True PC 2 loads on the variables 5-8. Variables 9 and 10 are noise.

		PCA		SCoTLASS	
		1	2	1	2
Variable	1	-0.46	0.16	0.52	0
	2	-0.48	0.14	0.34	0
	3	-0.47	0.13	0.55	0
	4	-0.47	0.17	0.55	0
	5	-0.13	-0.47	0	0.49
	6	-0.2	-0.46	0	0.51
	7	-0.1	-0.5	0	0.49
	8	-0.15	-0.47	0	0.51
	9	-0.17	-0.04	0	0
	10	-0.05	-0.08	0	0

Robust PCA - ROBPCA

ROBPCA (Hubert-Rousseeuw-Vanden Branden, 2005) combines ideas of both Projection Pursuit and robust covariance estimation in two main steps:

1.
   - Use PP to find a robust subspace of dimension $k \ll p$.
   - Consider H_1 as the set of observations which are 'close enough' to this subspace.
   - Apply classical PCA to H_1, yielding loadings and k-dimensional scores.
2. Robustly estimate the scatter matrix of the scores using the MCD. The eigenvectors and eigenvalues yield the robust loadings and eigenvectors.

ROBPCA

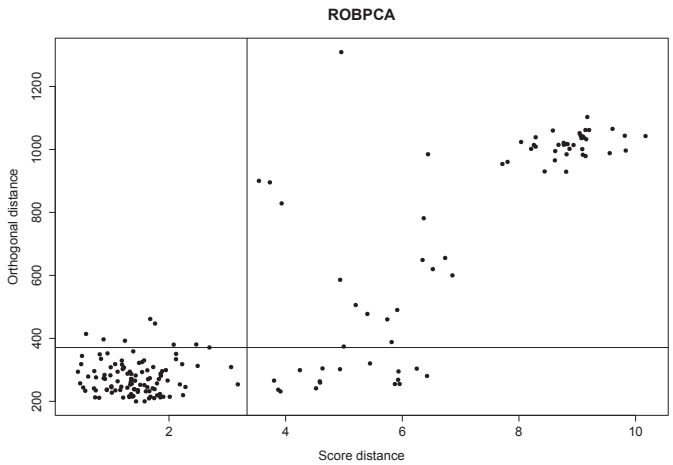
Fitting a model to the observations belonging to the majority of the data reveals three kinds of outliers:

1. Orthogonal outliers: observations that are far away from the PCA model space.
2. Good leverage points: observations that are far away on the PCA model space.
3. Bad leverage points: observations that are both far away from the model space, and on it.

These types of outlyingness can be visualized in an outlier map.

ROBPCA

Outlier map of ROBPCA applied to the glass data set



Sparse robust PCA

Croux et al. (2013) propose SRPCA. It integrates robustness and sparsity into the projection pursuit equations.

1. The jth sparse robust PCA direction is defined as in SCoTLASS with S a robust measure of scale, like the Q_n.
2. A grid algorithm is used to find appropriate directions.
3. The selection of λ is based on a BIC-criterion.

Robust sparse PCA

ROSPCA (Hubert et al. 2016) extends ROBPCA to obtain sparse loadings. Like ROBPCA, ROSPCA takes two main steps:

1.
   - Use PP to find a robust subspace of dimension $k \ll p$.
   - Consider H_1 as the set of observations which are 'close enough' to this subspace.
   - Apply SCoTLASS to H_1, yielding sparse loadings.
   - Perform a reweighting to account for the sparse structure. Set the variables with zero loadings aside and re-include observations that were only outlying on those variables. This yields the set H_2.
   - Apply SCoTLASS to H_2.

Robust sparse PCA

2.
   - Robustly estimate the eigenvalues using the Q_n^2 of the scores of the observations in H_2.
   - Robust center: the mean of the observations of H_2 with a 'regular' score distance (H_3).
   - Final eigenvalues: the sample variance of the (new) scores of the observations with indices in H_3 (the observations with low OD and high SD are not included anymore).

Selection of λ

Minimize

$$\mathrm{BIC}(\lambda) = \ln\left(\frac{1}{h_1 p}\sum_{i=1}^{h_1}\mathrm{OD}_{(i)}^2(\lambda)\right) + \mathrm{df}(\lambda)\frac{\ln(h_1 p)}{h_1 p}$$

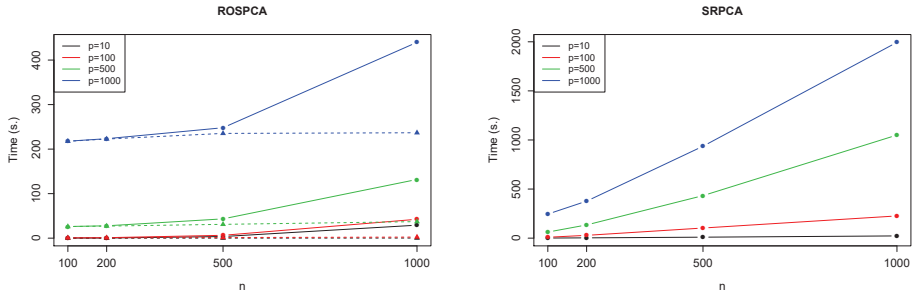
where

- h_1 is the size of H_1
- $\mathrm{OD}_{(i)}(\lambda)$ is the ith smallest orthogonal distance for the model when using λ as the sparsity parameter
- $\mathrm{df}(\lambda)$ the number of non-zero loadings

Minimize BIC within a grid $[0, \lambda_{max}]$ where λ_{max} gives full sparseness (exactly one non-zero loading per PC).

Computation of the index set H_1 in ROSPCA does not depend on the choice of the sparsity parameter!

Computation time



Computational performance of ROSPCA (left) and SRPCA (right) for varying values of n and p (fixed λ). The ROSPCA plot displays both the sparse (dashed line) and total (solid line) computation times.

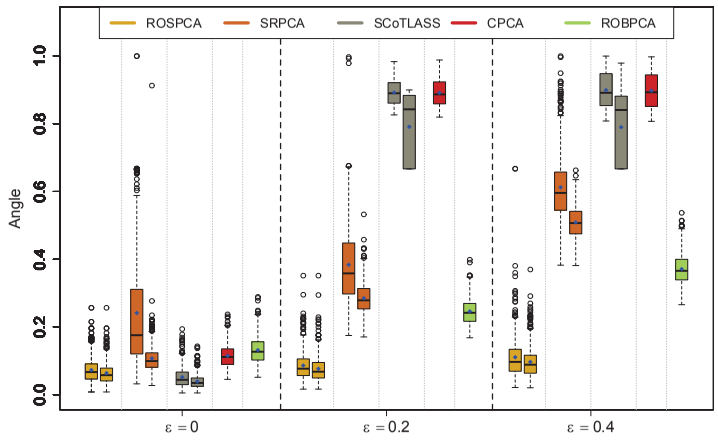
Simulations: Settings

To illustrate the performance of SCoTLASS, SRPCA and ROSPCA, we consider the following simulation:

- $n \in \{50, 100, 500\}$.
- $k = 2$.
- $p = 10$: PC 1 loads on variables 1-4. PC 2 loads on variables 5-8. Variables 9 and 10 are noise.
  $p = 500$: PC 1 loads on variables 1-20. PC 2 loads on variables 21-40. Other variables are noise.
- $\boldsymbol{X} = \boldsymbol{X}_u + \boldsymbol{X}_{noise}$ with $\boldsymbol{X}_u \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{X}_{noise} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$.
- The outliers $\sim N_p(\boldsymbol{\mu}_{out}, 20\boldsymbol{I}_p)$.
- The proportion of outliers $\varepsilon = 0, 0.2, 0.4$.

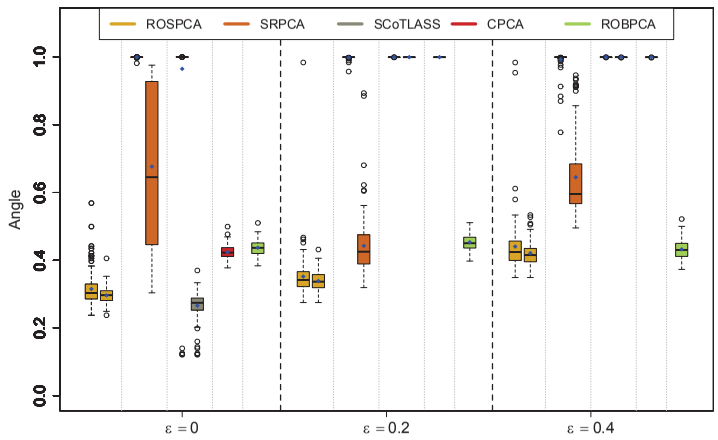
Simulations: Robustness

Angle (over 500 runs) between the true subspace and the estimated onefor $n = 100$, $p = 10$. Left boxplot is based on the selected λ using BIC, right boxplot is the optimal λ over a grid of λ-values.



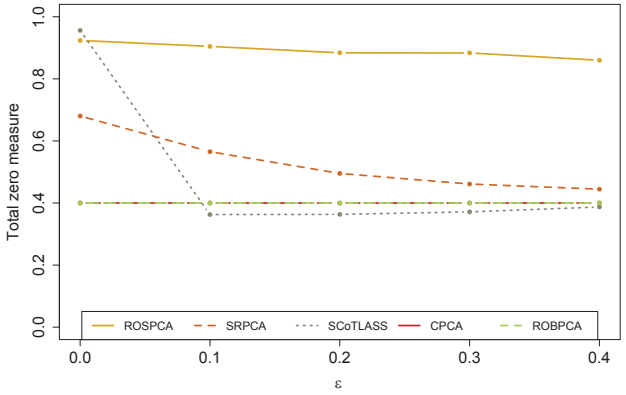
Simulations: Robustness

Angle (over 100 runs) values $n = 100$, $p = 500$:



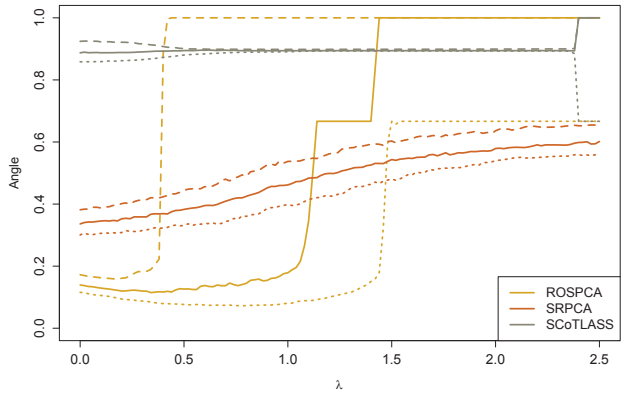
Simulations: Sparsity

Zero measure: for each element of $\boldsymbol{P}$, it is equal to 1 if the estimated and true value are both zero or both non-zero, and 0 otherwise.
Total zero measure: the average zero measure over all elements of $\boldsymbol{P}$ and all simulations.



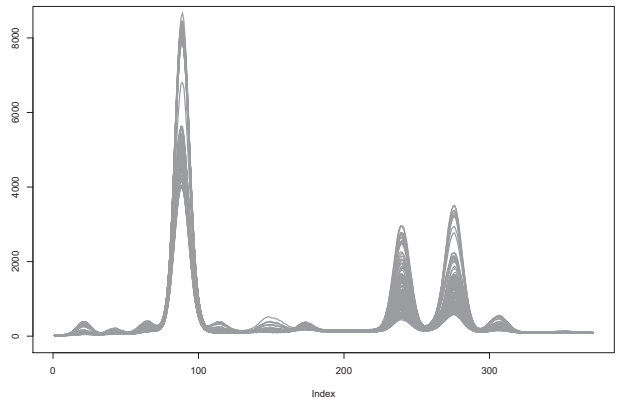
Simulations: λ selection

Quantile plot of angle values over 500 simulations at $n = 100$ and $\varepsilon = 0.2$ as a function of λ.



Example: glass data

EPXMA spectra of 180 collected glass samples.

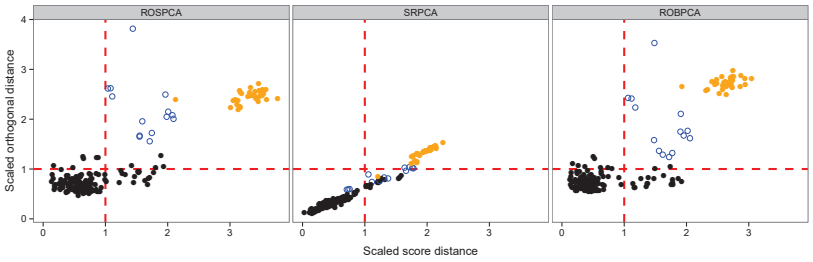


Example: glass data

- A sparse method is interesting since the data appears to have a sparse structure.
- Two known groups of outliers: observations with high calcium concentrations (blue), and observations that were measured after the spectrometer was cleaned (orange).
- The selected $k = 4$.
- The goal: achieve comparable outlier detection results while obtaining sparse loadings that reflect the atomic structure of the glass samples.

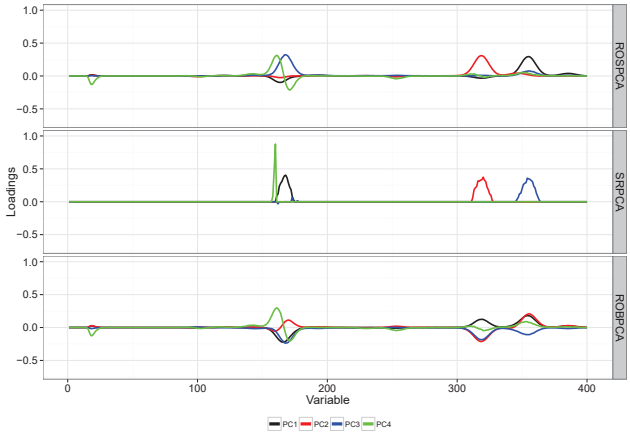
Glass data: outlier detection

Scaled outlier maps of ROSPCA ($\lambda = 0.96$, 146s), SRPCA ($\lambda = 72.7$, 419s) and ROBPCA.



Glass data: sparsity

Loadings of ROSPCA, SRPCA, and ROBPCA



Glass data: sparsity

Number of non-zero loadings (larger than 10^{-5}) for each method per PC. The bottom row is the number of variables that have zero loadings (smaller than 10^{-5}) on all 4 PCs.

	ROSPCA	SRPCA	ROBPCA
PC1	359	14	733
PC2	272	17	735
PC3	491	34	737
PC4	408	4	736
No. of excluded variables	200	696	13

Travel data

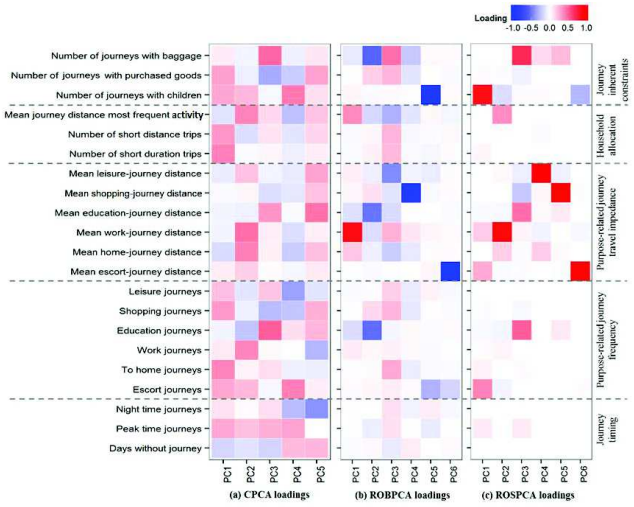
Goal: to study travel behavior determinants based on a multiday travel survey conducted in the region of Ghent, Belgium

Data: 717 individuals recorded all travel activities that were carried out over a 7-day course, observed in a period between September and December 2008.

Variables: number of peak-time journeys, number of work journeys, number of education journeys, mean home-journey distance, mean shopping-journey distance, number of journeys with children, number of journeys with baggage,...

Analysis: Plevka et al. 2016

Travel data: loadings



References

- Croux, C., Filzmoser, P., Fritz, H. (2013). Robust sparse Principal Component Analysis, *Technometrics,* 55, 202–214.
- Hubert, M., Reynkens, T., Schmitt, E., Verdonck, T. (2016). Sparse PCA for high-dimensional data with outliers, *Technometrics*, 58, 424–434.
- Jolliffe, I.T., Trendafilov, N.T., Uddin, M.A. (2003). Modified principal component technique based on the LASSO, *Journal of Computational and Graphical Statistics,* 12, 531–547.
- Plevka, V., Segaert, P., Tampère, C., Hubert, M. (2016). Analysis of travel activity determinants using robust statistics. *Transportation*, 43, 979–996.

Want some more robustness?

Workshop to celebrate Peter Rousseeuw on the occasion of his 60th birthday.

May 31 - June 2, 2017

Leuven, Belgium

All information will become available at:

wis.kuleuven.be/stat/robust/PR60