Robust sparse PCA based on projection-pursuit

Peter Filzmoser

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

Graz, Austria

September 8, 2011

Vienna University of Technology

Joint work with . . .

Christophe Croux, K.U. Leuven, Belgium

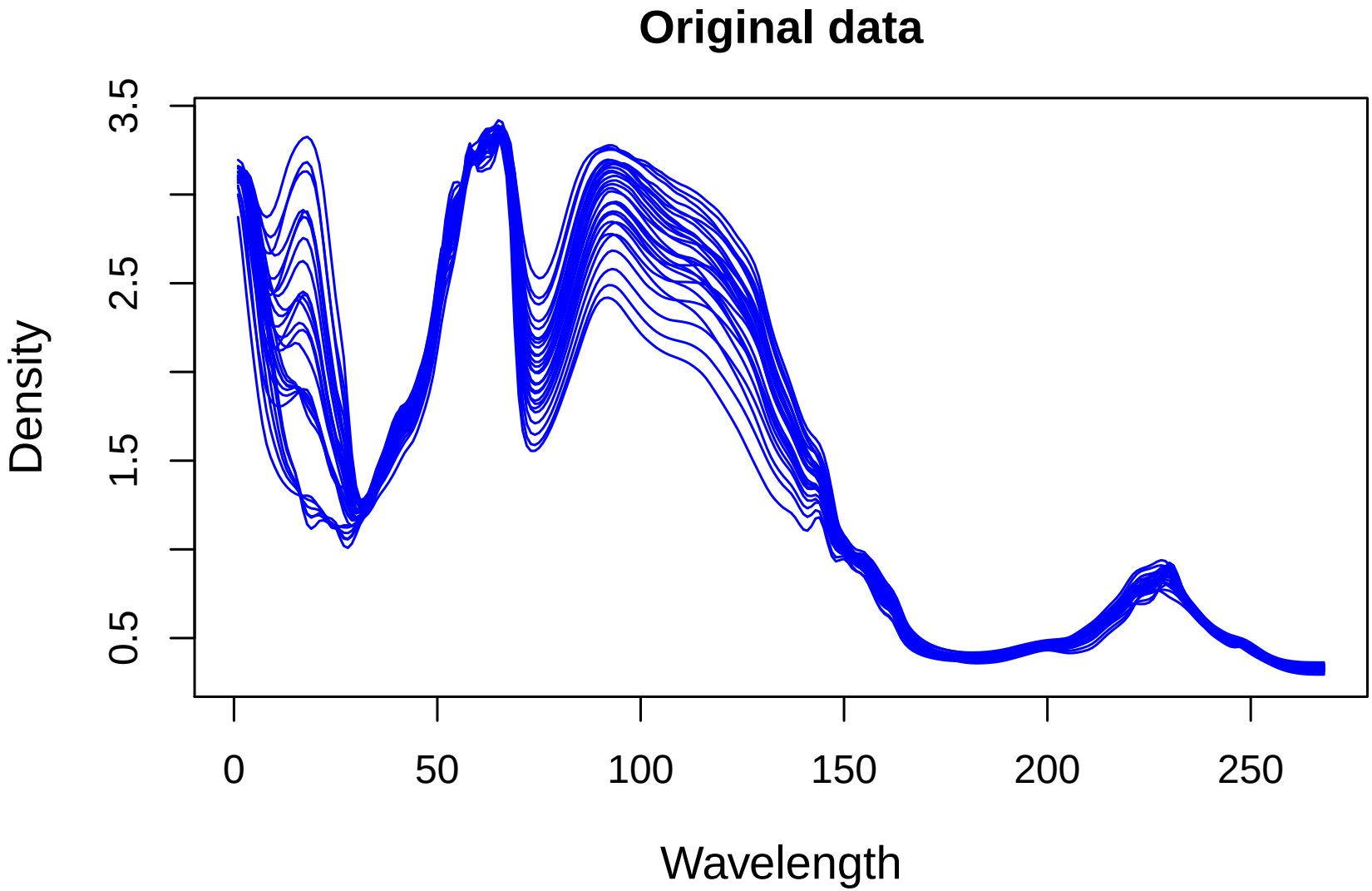
Heinrich Fritz, Vienna University of Technology

Outline

- **Projection-pursuit PCA**
- **PCA and sparsity**
- **Algorithm**
- **Examples**
- **Summary**

Motivating example

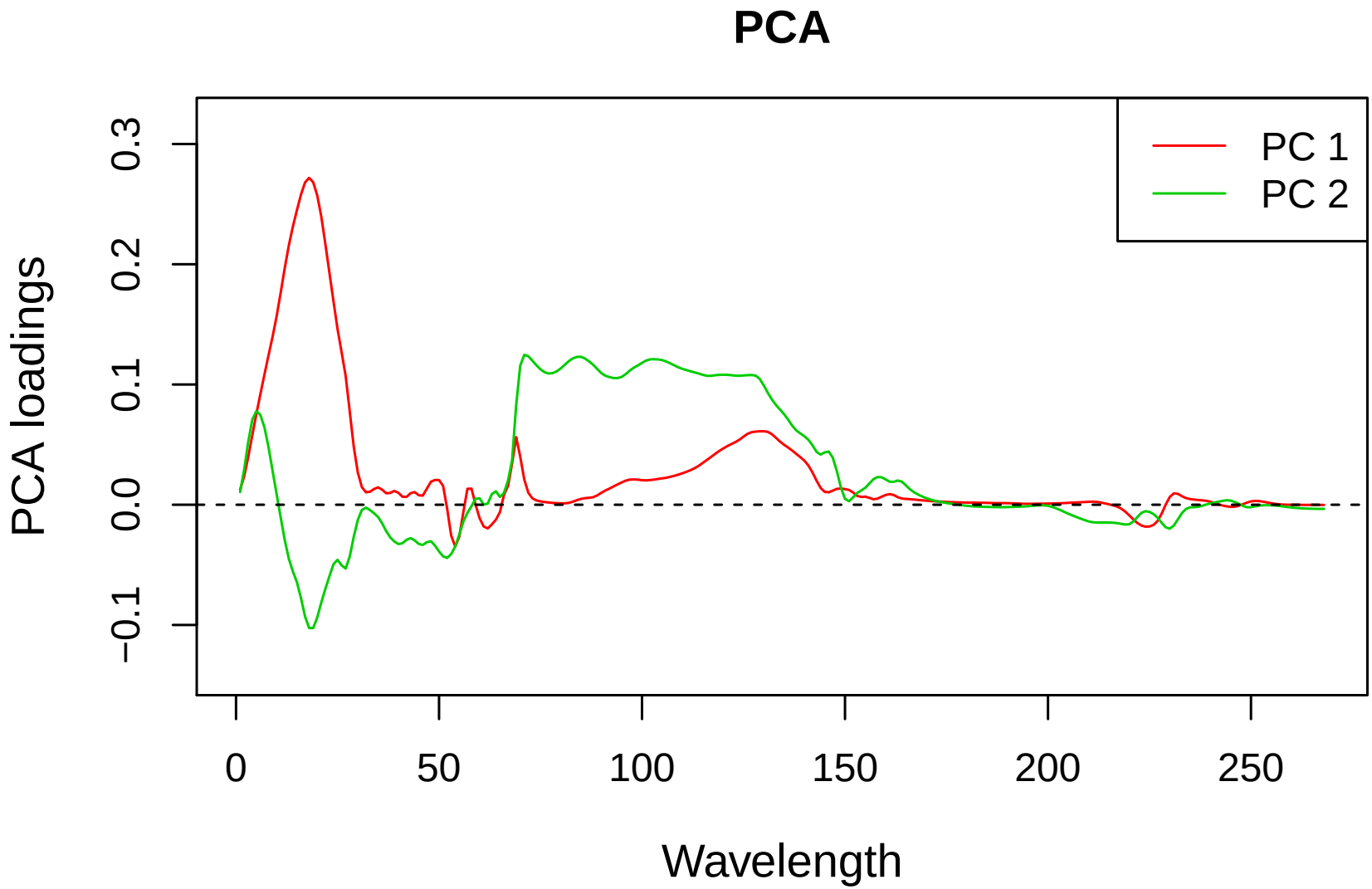
Data set “yarn”: density of 28 pet yarns, measured at 268 wavelength



Peter.Filzmoser@tuwien.ac.at

Motivating example

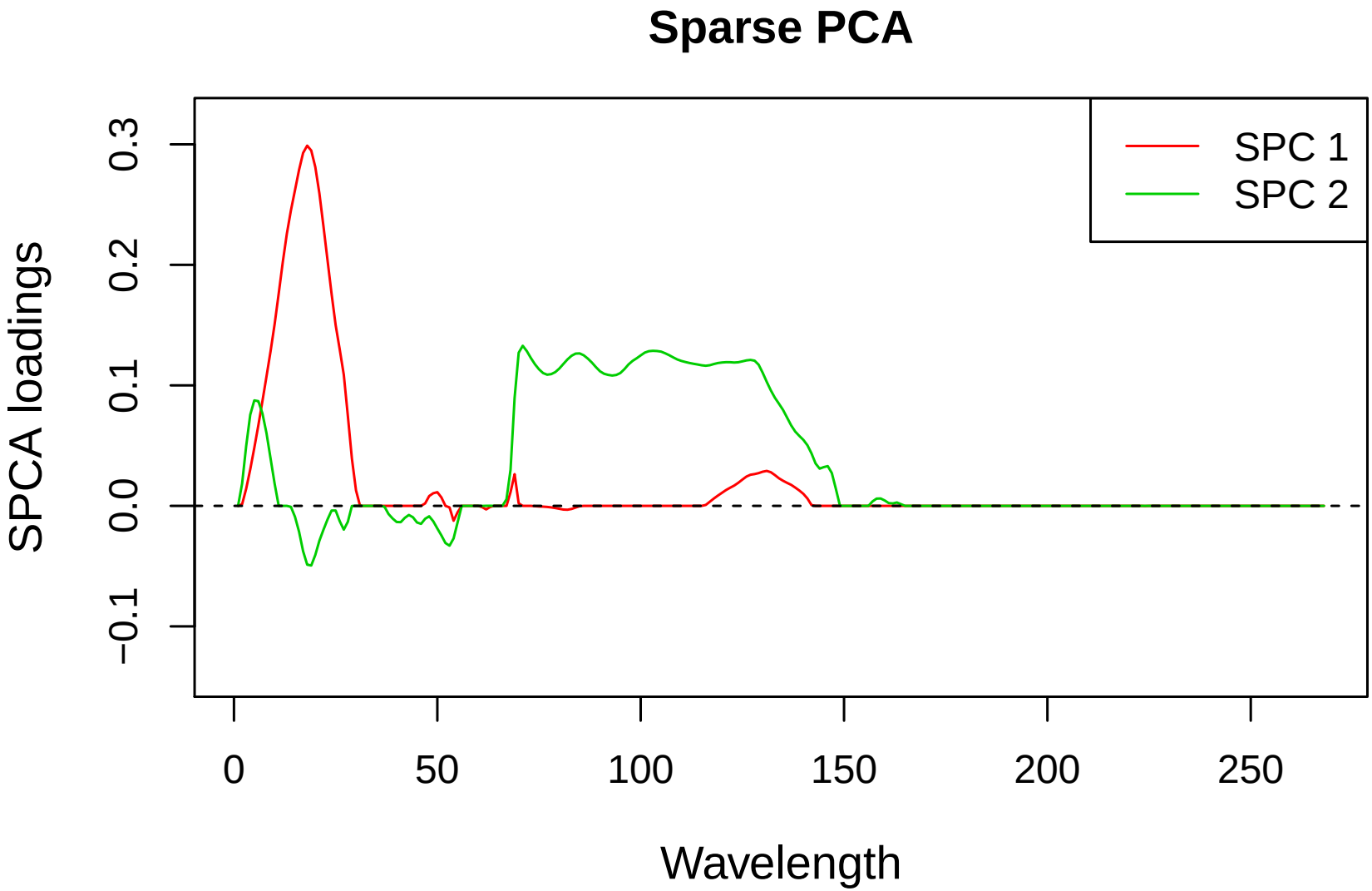
First two PCA loadings:



Peter.Filzmoser@tuwien.ac.at

Motivating example

First two Sparse PCA loadings:



Peter.Filzmoser@tuwien.ac.at

Projection-pursuit PCA

Given: n observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in I\!R^p$, collected in the rows of $\boldsymbol{X}$.
The first PCA direction is given by

$$\boldsymbol{a}_1 = \underset{\|\boldsymbol{a}\|=1}{\text{argmax}}\, V(\boldsymbol{a}^t\boldsymbol{x}_1, \ldots, \boldsymbol{a}^t\boldsymbol{x}_n).$$

V is a variance measure:

- **classical case:** V is the empirical variance (Var)
  $\Longrightarrow \boldsymbol{a}_1$ corresponds to the first eigenvector of the sample covariance matrix.

- **robust case:** V is squared MAD or squared Q_n estimator (Rousseeuw and Croux, 1993)
  $\Longrightarrow \boldsymbol{a}_1$ is the direction of the first **robust** PC.

Peter.Filzmoser@tuwien.ac.at

Projection-pursuit PCA

Suppose the first $j-1$ PCA directions have already been found ($j > 1$).
The jth direction ($j \leq p$) is defined as:

$$\boldsymbol{a}_j = \underset{\|\boldsymbol{a}\|=1, \boldsymbol{a}\perp\boldsymbol{a}_1,\ldots,\boldsymbol{a}\perp\boldsymbol{a}_{j-1}}{\text{argmax}} V(\boldsymbol{a}^t\boldsymbol{x}_1,\ldots,\boldsymbol{a}^t\boldsymbol{x}_n)$$

Loadings matrix for the first k PCs: $\boldsymbol{A}_k = (\boldsymbol{a}_1,\ldots,\boldsymbol{a}_k)$,

Scores matrix of the first k PCs: $\boldsymbol{Z}_k = \boldsymbol{X}\boldsymbol{A}_k$ with elements $z_{ij} = \boldsymbol{a}_j^t\boldsymbol{x}_i$

Peter.Filzmoser@tuwien.ac.at

PCA and sparsity

. . . add an L_1 penalty in the objective function (see Tibshirani, 1996):

SCoTLASS criterion (Jolliffe et al., 2003):
jth sparse PCA direction ($1 \leq j \leq p$):

$$\max_{\|\boldsymbol{a}\|=1, \boldsymbol{a} \perp \boldsymbol{a}_1, \ldots, \boldsymbol{a} \perp \boldsymbol{a}_{j-1}} \boldsymbol{a}^t \hat{\boldsymbol{\Sigma}} \boldsymbol{a}, \quad \text{subject to } \|\boldsymbol{a}\|_1 \leq t,$$

with L_1 norm $\|\boldsymbol{a}\|_1 = \sum_{j=1}^{p} |\boldsymbol{a}_j|$, and $\hat{\boldsymbol{\Sigma}}$ the empirical covariance matrix.

Equivalent formulation:

$$\max_{\|\boldsymbol{a}\|=1, \boldsymbol{a} \perp \boldsymbol{a}_1, \ldots, \boldsymbol{a} \perp \boldsymbol{a}_{j-1}} \boldsymbol{a}^t \hat{\boldsymbol{\Sigma}} \boldsymbol{a} - \lambda_1 \|\boldsymbol{a}\|_1,$$

with the **tuning parameter** λ_1.

- L_1 penalty forces some of the loadings to become exactly zero;
- $\hat{\boldsymbol{\Sigma}}$ necessary as input $\longrightarrow$ robust estimation??? $p > n$???

Peter.Filzmoser@tuwien.ac.at

Projection-pursuit PCA and sparsity

First sparse PCA direction:

$$\tilde{\boldsymbol{a}}_1 = \underset{\|\boldsymbol{a}\|=1}{\text{argmax}}\, V(\boldsymbol{a}^t\boldsymbol{x}_1, \ldots, \boldsymbol{a}^t\boldsymbol{x}_n) - \lambda_1\|\boldsymbol{a}\|_1.$$

- Setting $\lambda_1 = 0 \longrightarrow$ unconstrained first PCA direction $\boldsymbol{a}_1$;
- increasing $\lambda_1 \longrightarrow$ sparsity gains importance compared to (robust) variance maximization.

Peter.Filzmoser@tuwien.ac.at

Projection-pursuit PCA and sparsity

jth sparse PCA direction ($1 < j \leq p$):

$$\tilde{\boldsymbol{a}}_j = \underset{\|\boldsymbol{a}\|=1, \boldsymbol{a} \perp \tilde{\boldsymbol{a}}_1, \ldots, \boldsymbol{a} \perp \tilde{\boldsymbol{a}}_{j-1}}{\operatorname{argmax}} V(\boldsymbol{a}^t \boldsymbol{x}_1, \ldots, \boldsymbol{a}^t \boldsymbol{x}_n) - \lambda_j \|\boldsymbol{a}\|_1$$

λ_j is a **tuning parameter**, possibly different from λ_1.

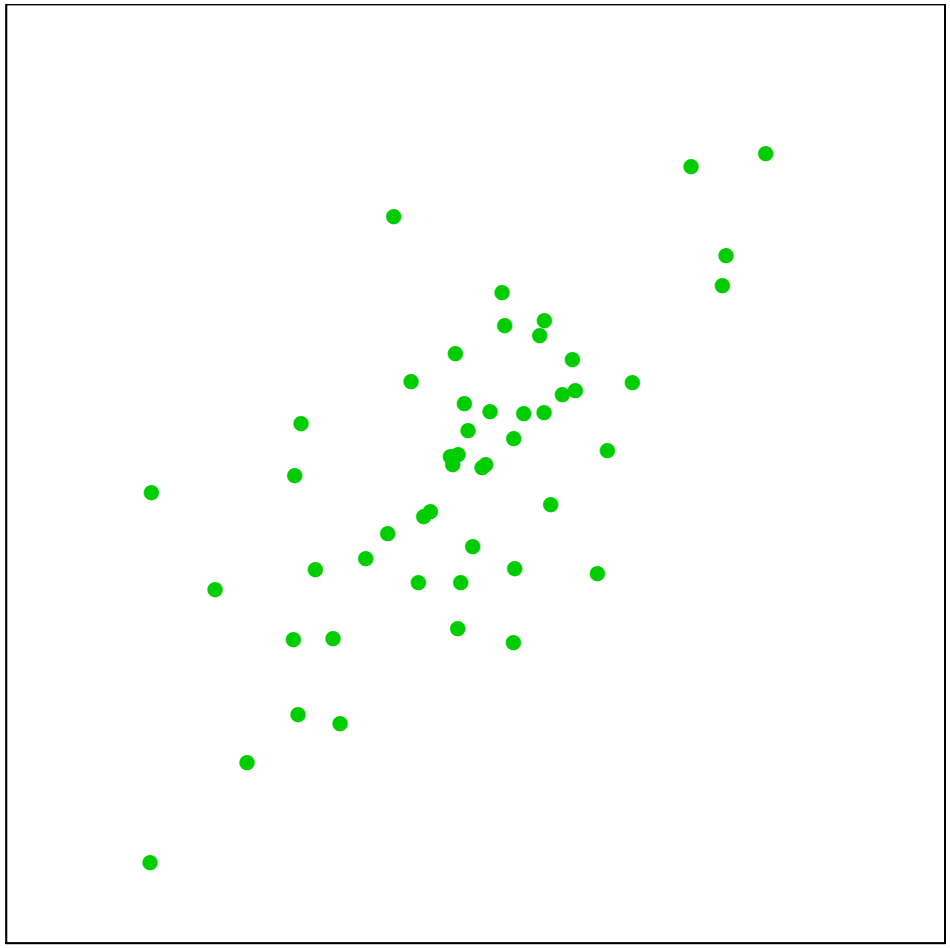
If $V = \text{Var} \longrightarrow$ SCoTLASS and projection-pursuit definitions are the same.

Projection-pursuit approach: PCs are computed sequentially; stop at a desired number $k < p$.

Peter.Filzmoser@tuwien.ac.at

Projection-pursuit algorithm

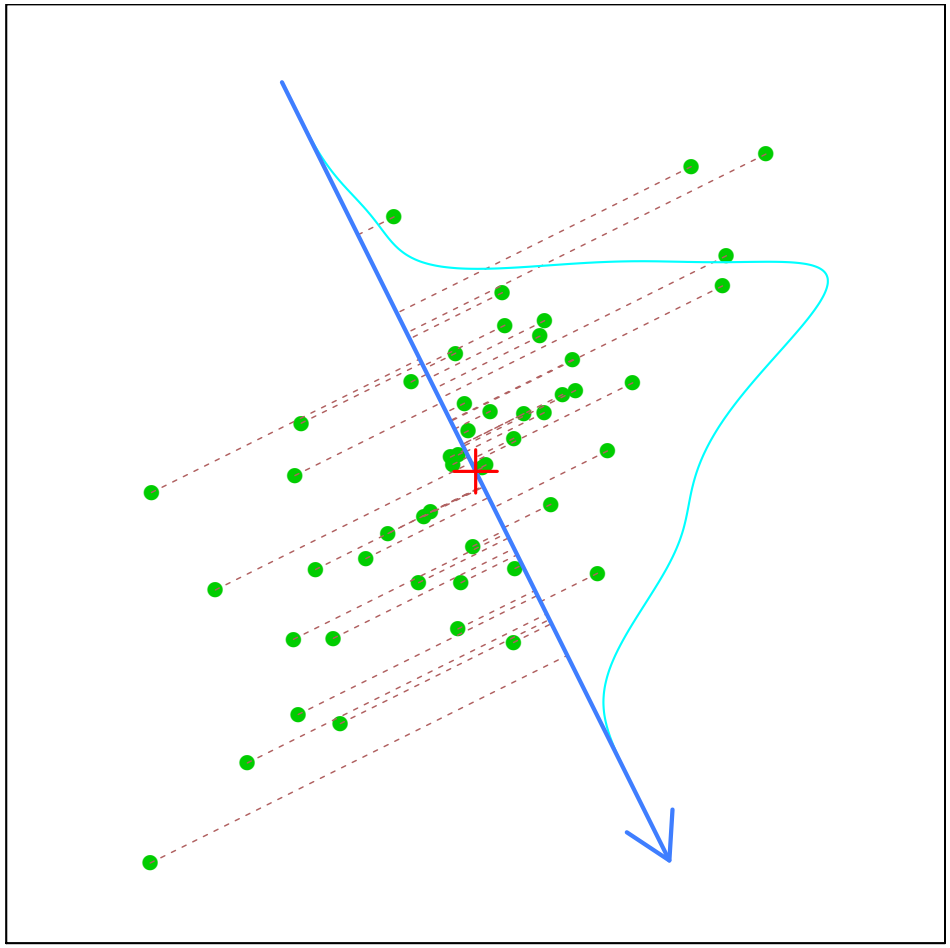
- project data on a direction



Peter.Filzmoser@tuwien.ac.at

Projection-pursuit algorithm

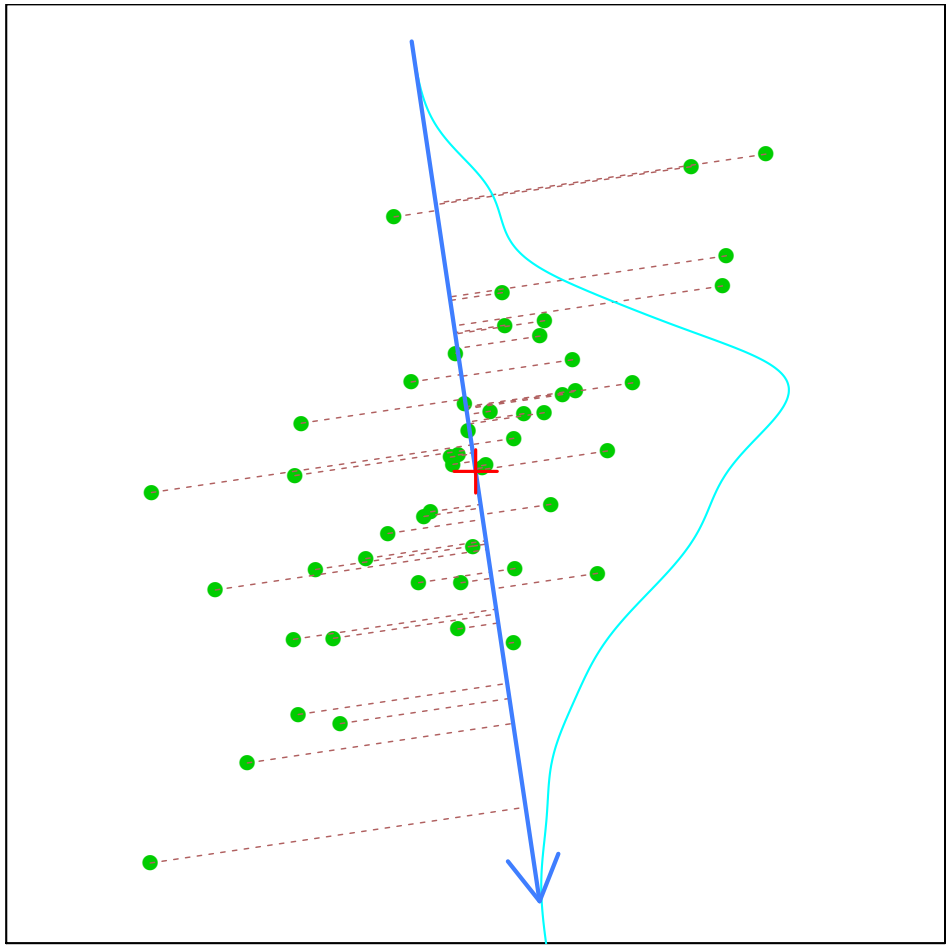
- project data on a direction
- compute (robust) variance



Peter.Filzmoser@tuwien.ac.at

Projection-pursuit algorithm

- project data on a direction
- compute (robust) variance
- repeat for “many” directions



Peter.Filzmoser@tuwien.ac.at

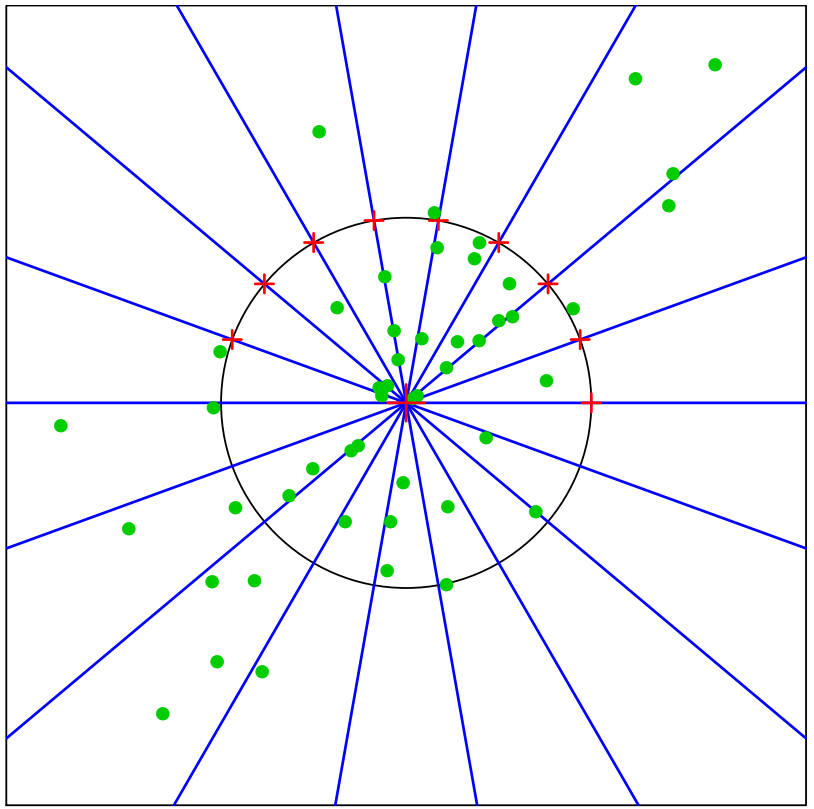
Grid algorithm

Optimization is done on a regular grid in the plane:

- select N_g regular grid points on the unit circle
- consider N_g candidate directions of each grid point through the center
- maximize the variance of the projected data, measured by a measure V

$\Longrightarrow$ direction which gives maximum (robust) variance

(Croux, Filzmoser, Oliveira, 2007)



Peter.Filzmoser@tuwien.ac.at

Grid algorithm for sparse PCA

Assume that the first $j-1$ sparse PCA directions $\tilde{\boldsymbol{a}}_{j-1}$ are already obtained and collected in the columns of $\tilde{\boldsymbol{A}}_{j-1}$, with $1 \leq j \leq k-1$.

Compute $\tilde{\boldsymbol{a}}_j$:

- $\tilde{\boldsymbol{A}}_{j-1}^{\perp}$ contains orthonormal basis of the subspace orthogonal to $\tilde{\boldsymbol{A}}_{j-1}$.
- Project data into this space by $\boldsymbol{x}_i^{(j-1)} = (\tilde{\boldsymbol{A}}_{j-1}^{\perp})^t \boldsymbol{x}_i$, for $i = 1, \ldots, n$.
- Use objective function

$$f(\boldsymbol{a}) = V(\boldsymbol{a}^t \boldsymbol{x}_1^{(j-1)}, \ldots, \boldsymbol{a}^t \boldsymbol{x}_n^{(j-1)}) - \lambda_j \|\tilde{\boldsymbol{A}}_{j-1}^{\perp} \boldsymbol{a}\|_1$$

  with $\|\boldsymbol{a}\| = 1$.
- Cycle through all **pairs of variables** using the **Grid algorithm**; update optimal direction iteratively.

Peter.Filzmoser@tuwien.ac.at

Selection of the tuning parameter

The tuning parameter λ_j regulates the degree of sparseness.

We want to have similar degree of sparseness in the different PCs

$$\Longrightarrow \text{ take } \lambda_j := \lambda\, \mathcal{V}(\boldsymbol{X}^{(j)}),$$

where $\boldsymbol{X}^{(j)} = \boldsymbol{X}\tilde{\boldsymbol{A}}_{j-1}^{\perp}$, and $\mathcal{V}$ denotes the total robust variance, defined for any matrix $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p)$ as

$$\mathcal{V}(\boldsymbol{Y}) = \sum_{i=1}^{p} V(\boldsymbol{y}_i).$$

Peter.Filzmoser@tuwien.ac.at

Selection of the tuning parameter λ

BIC type criterion:

$$\text{BIC}(\lambda) = \frac{\widetilde{\text{RV}}}{\text{RV}} + \text{df}(\lambda)\frac{\log(n)}{n},$$

with

- $\widetilde{\text{RV}} = \mathcal{V}(\boldsymbol{X} - \boldsymbol{X}\tilde{\boldsymbol{A}}_k\tilde{\boldsymbol{A}}_k^t)$ ... total robust variance of residuals from sparse PCA
- $\text{RV} = \mathcal{V}(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}_k\boldsymbol{A}_k^t)$ ... total robust variance of residuals from unconstrained PCA
- $\text{df}(\lambda)$ ... number of non-zero loadings when using λ.

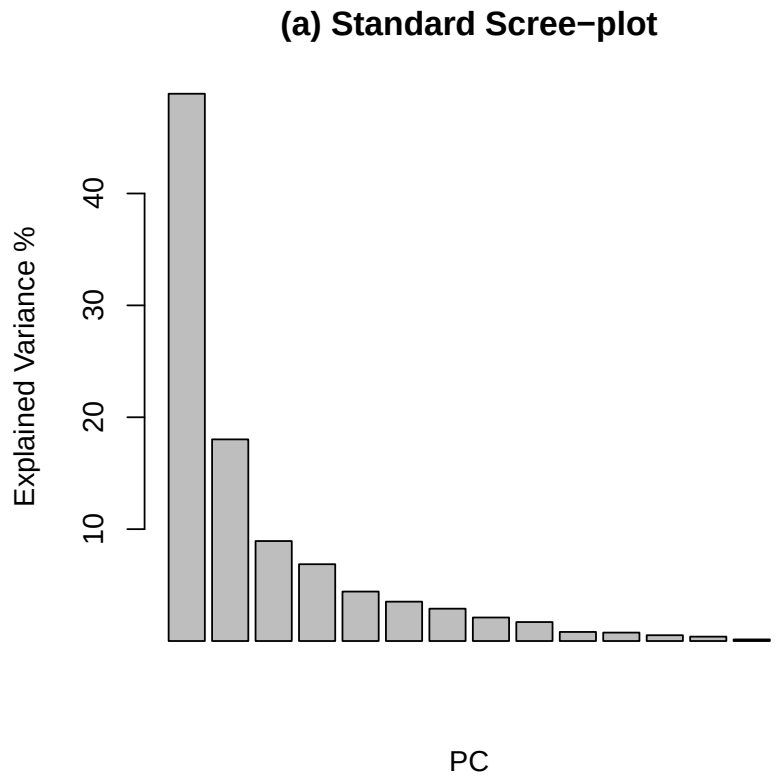
Select λ by minimizing $\text{BIC}(\lambda)$ over a grid $[0, \lambda_{max}]$, where λ_{max} results in full sparseness.

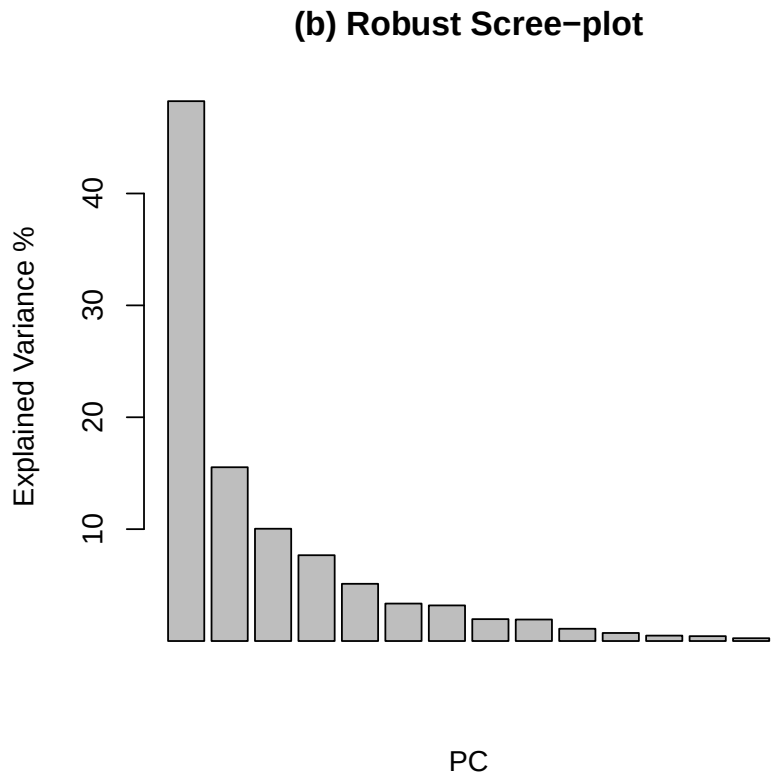
Example car data set

Car data (Kibler et al., 1989): 205 different car models; 26 variables containing technical and insurance-related data

$\longrightarrow$ only continuous variables, no missings: data set of size 195×14

Compare classical ($V = \text{Var}$) and robust ($V = Q_n^2$) approach.

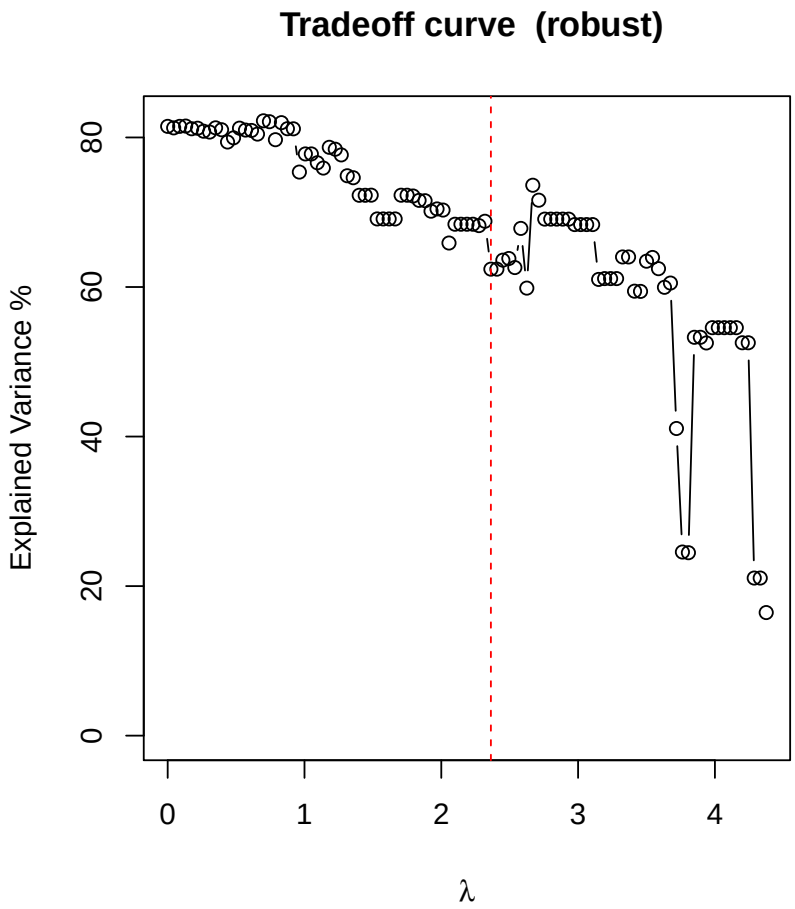




Peter.Filzmoser@tuwien.ac.at

Example car data set

"Tradeoff curve" for sparse robust PCA:



BIC criterion $\Longrightarrow \lambda = 2.36$

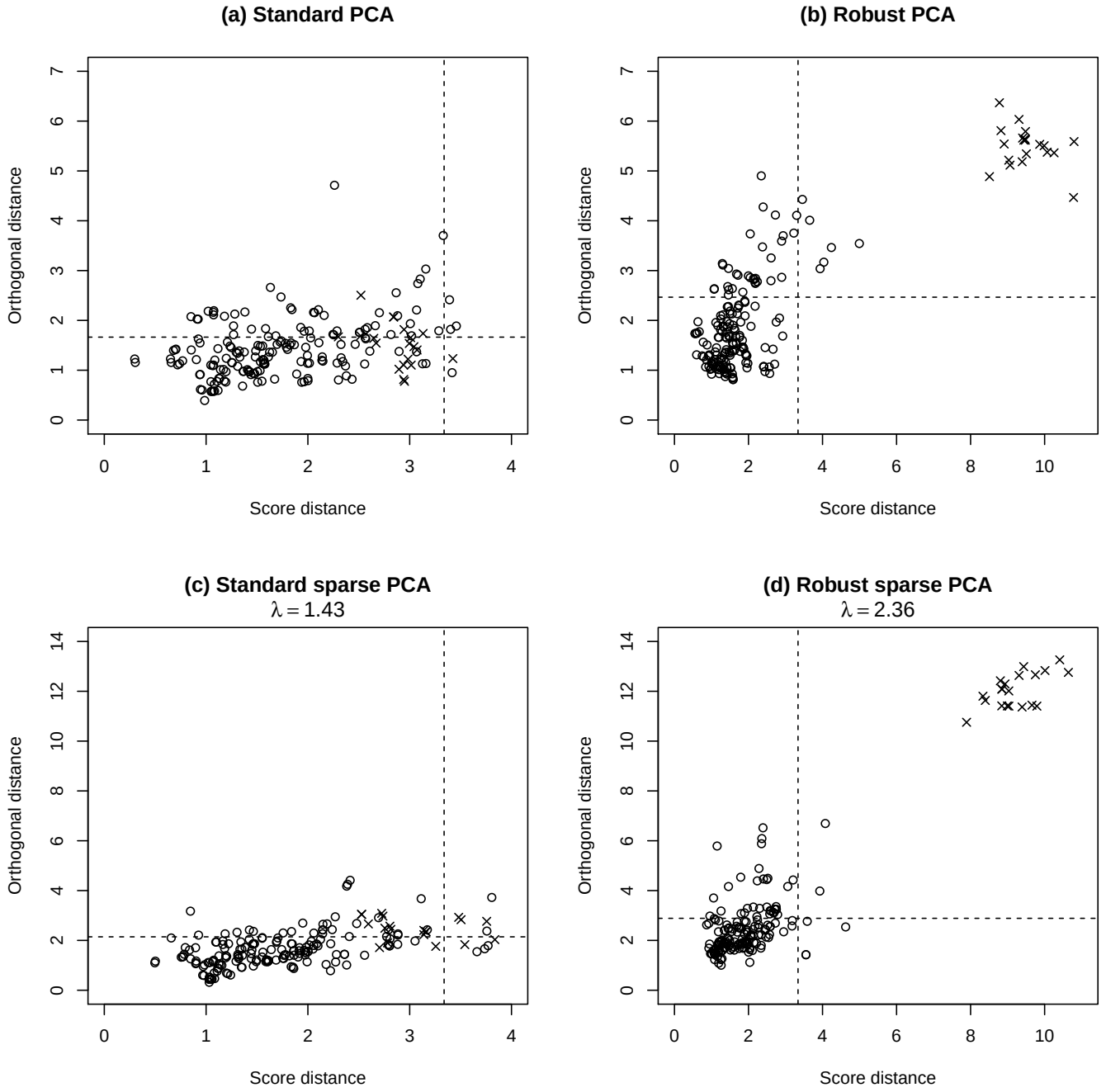
Peter.Filzmoser@tuwien.ac.at

Example car data set

	Robust PCA ($\lambda = 0$)				Robust sparse PCA ($\lambda = 2.36$)			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
symboling	-0.03	-0.04	0.03	-0.17	0	0	0	0
wheel-base	0.24	0.25	0.08	0.16	0	0.50	0	0
length	0.29	0.18	-0.05	0.04	0.24	0	0.85	0
width	0.26	0.16	0.14	0.03	0.21	0	0	0
height	0.08	0.39	-0.26	0.32	0	0.87	0	0
curb-weight	0.24	0.13	0.12	0.00	0.32	0	0	0
bore	0.24	0.16	-0.25	0.04	0.21	0	0.03	0
stroke	0.00	-0.24	0.29	-0.58	0	0	0	0
compression-ratio	-0.47	0.61	0.49	-0.11	-0.45	0	0.53	0
horsepower	0.36	-0.01	0.16	-0.20	0.43	0	0	0
peak-rpm	0.08	-0.38	0.60	0.64	0	0	0	1.00
city-mpg	-0.31	0.04	-0.02	0.14	-0.30	0	0	0
highway-mpg	-0.33	0.07	-0.04	0.14	-0.35	0	0	0
price	0.33	0.31	0.34	-0.12	0.40	0	0.06	0
EV %	49.20	15.54	10.12	5.97	45.73	8.32	6.03	4.16
Cumulative EV %	49.20	64.74	74.85	80.82	45.73	54.05	60.08	64.24

Peter.Filzmoser@tuwien.ac.at

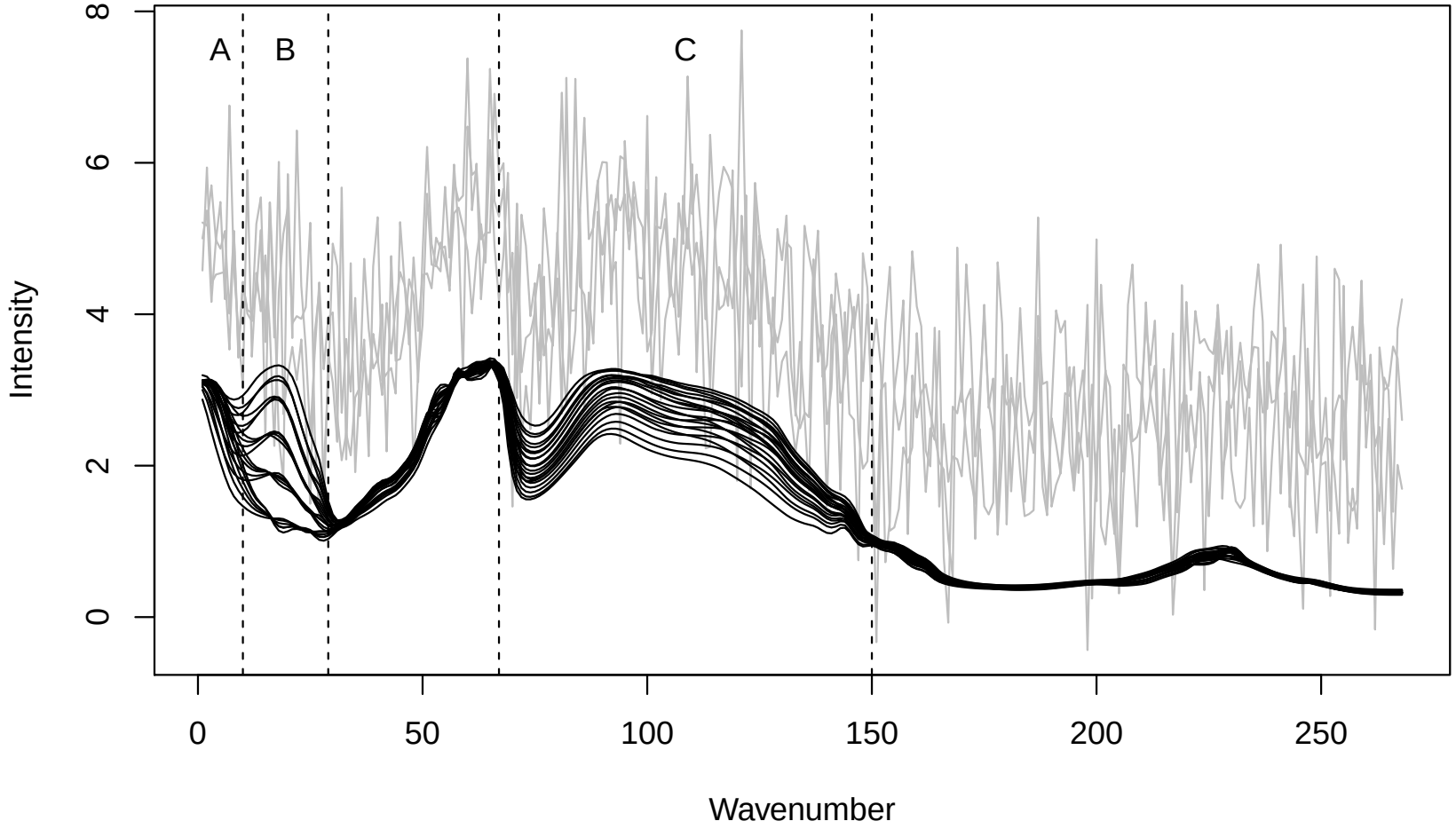
Example car data set



Peter.Filzmoser@tuwien.ac.at

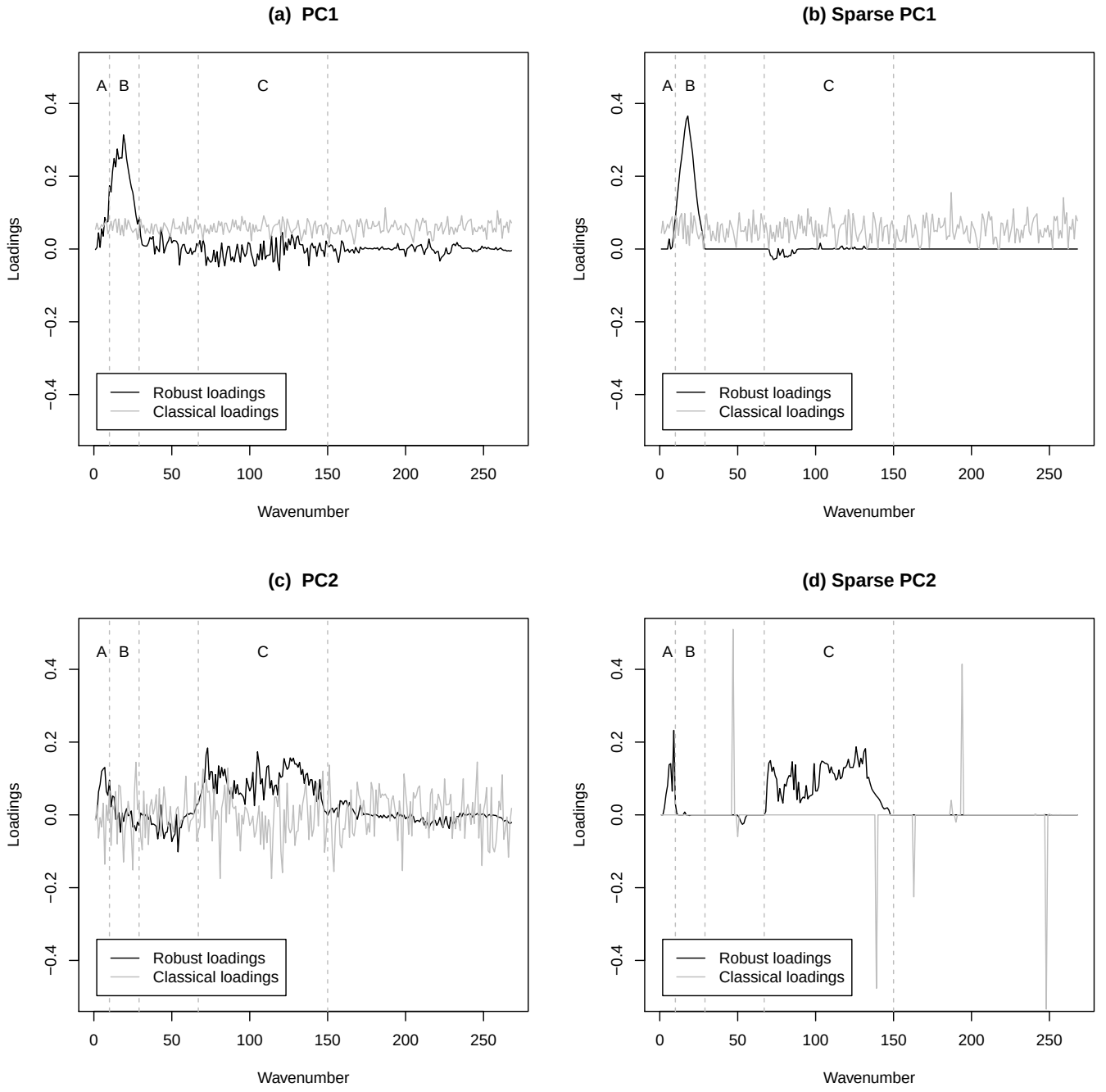
Example yarn data set

Yarn data, with 3 outlying spectra added



Peter.Filzmoser@tuwien.ac.at

Example yarn data set



Peter.Filzmoser@tuwien.ac.at

R code

Implementation in R package pcaPP:

- Unconstrained PCA with Grid algorithm:

```
PCAgrid(x, k = 2, method = c ("mad", "sd", "qn"),
maxiter = 10, splitcircle = 25, scores = TRUE, zero.tol = 1e-16,
center = l1median, scale, trace = 0, store.call = TRUE, control, ...)
```

- Sparse PCA with Grid algorithm:

```
sPCAgrid(x, k = 2, method = c ("mad", "sd", "qn"), lambda = 1,
maxiter = 10, splitcircle = 25, scores = TRUE, zero.tol = 1e-16,
center = l1median, scale, trace = 0, store.call = TRUE, control, ...)
```

Summary

Sparse robust PCA ...

- compromise between maximizing robust variance and simplifying interpretability
- determination of PCA directions is not affected by outliers

Projection-pursuit approach ...

- robust estimation also possible for $p > n$
- components are extracted sequentially (stop after k)
- robust variance estimator and sparsity criterion can easily be changed

Peter.Filzmoser@tuwien.ac.at

Some references

Croux, C., Filzmoser, P., and Fritz, H. (2011). Robust sparse principal component analysis. Technical Report, SM-2011-2, Vienna University of Technology, Austria; (submitted for publication).

Croux, C., Filzmoser, P., and Oliveira, M. (2007). Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, **87**, 218–225.

Guo, J., James, G., Levina, E., Michailidis, G., and Zhu, J. (2011). Principal component analysis with sparse fused loadings. *Journal of Computational and Graphical Statistics*. To appear.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, **12**, 531–¡96¿547.

Rousseeuw, P. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, **88**(424), 1273–1283.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**(2), 265–286.

Peter.Filzmoser@tuwien.ac.at