Dimension reduction and nonparametric regression: A robust combination

Claudia Becker

Department of Statistics
University of Dortmund
Germany
cbecker@statistik.uni-dortmund.de

1. Introduction
2. The challenging task
3. Robust procedures for dimension adjustment
4. Example

1. Introduction

Example

$$\begin{pmatrix} y_1 \\ \vdots \\ y_{59} \end{pmatrix}, \quad \begin{pmatrix} x_{1,1} \\ \vdots \\ x_{59,1} \end{pmatrix}, \ldots, \begin{pmatrix} x_{1,12} \\ \vdots \\ x_{59,12} \end{pmatrix}$$

mortality in the US (↑ y)

air pollution and socio-economic conditions (↑ x … ↑)

$x_{1,i}$ = population density,
$x_{2,i}$ = SO_2 air pollution, etc.

$i=1,\ldots,59 \triangleq$ metropolitan areas in the US

⇒ **n = 59** vectors of data, of **d = 12** components each

Task: relate y and x by function f

2. The challenging task

Assume

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} \xrightarrow{\ g\ } Y$$

where

- g unknown, no further assumptions
- sample $(y_i,\mathbf{x}_i)$ of size n given
- d "large"

Task: estimate g nonparametrically

⇒ nonparametric regression methods

But: problem in higher dimensions

Probability for an observation of **X** to lie in a ball of radius r

$P(\|\mathbf{X}\| \leq 0.5)$	dimension d: 1	5	10	20
$\mathbf{X} \sim U_d(\mathcal{U}_1(\mathbf{0}))$	0.5	0.0313	0.001	$< 10^{-6}$
$\mathbf{X} \sim N_d(0,1)$	0.3839	0.1175	$< 10^{-6}$	≈ 0

Classical techniques fail for 'large' dimension d due to the

curse of dimensionality

Some approaches to solve the problem

- Projection Pursuit Regression

  (e.g. Friedman and Stuetzle 1981)

- Regression Trees

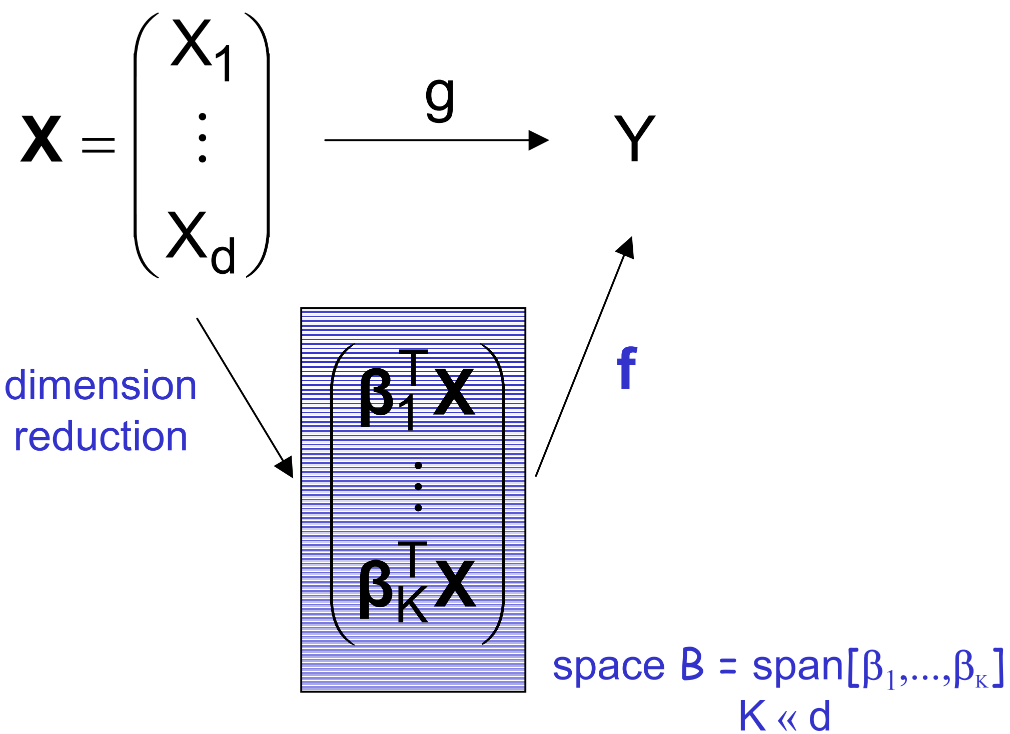
  (e.g. MART; Friedman 2000)

- Combine dimension reduction and nonparametric function estimation

dimension adjustment methods

3. Robust procedures for dimension adjustment

Model for dimension reduction
(Li 1991)



Dimension adjustment method

- estimate K
- estimate $\mathcal{B}$
- project X into $\mathcal{B}$
- estimate f

Why "robust"?

Air pollution and mortality example:

(Becker and Gather 1999)

9 observations identified

as outliers

- estimating the dimension K

  e.g. $\hat{K}=0$ if true $K=1$

  (Gather et al. 2001)

- estimating reduced space $\mathcal{B}$

  e.g. orthogonal to true $\mathcal{B}$

  (Gather et al. 2001)

- estimating f in reduced space

Dimension reduction: Sliced inverse regression (SIR)

(Li 1991)

Under certain conditions:

$$\mathbf{\Sigma}^{-1/2}(E(\mathbf{X} \mid Y) - E(\mathbf{X}))$$

lies (a.s.) in the linear subspace spanned by the directions

$$\mathbf{\Sigma}^{1/2}\boldsymbol{\beta}_1, \ldots, \mathbf{\Sigma}^{1/2}\boldsymbol{\beta}_K$$

⇨ estimate inverse regression curve roughly

⇨ determine space in which it is mainly spread out

estimator for β

SIR based on

- sample mean and covariance
- classical principal component analysis

→ not robust

Robust alternative: DAME

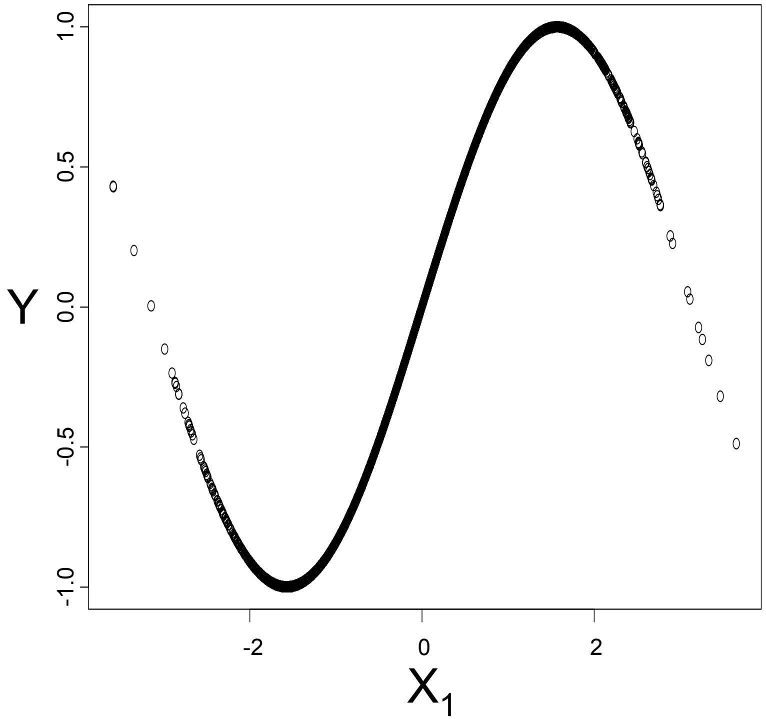
same basics as SIR, but based on

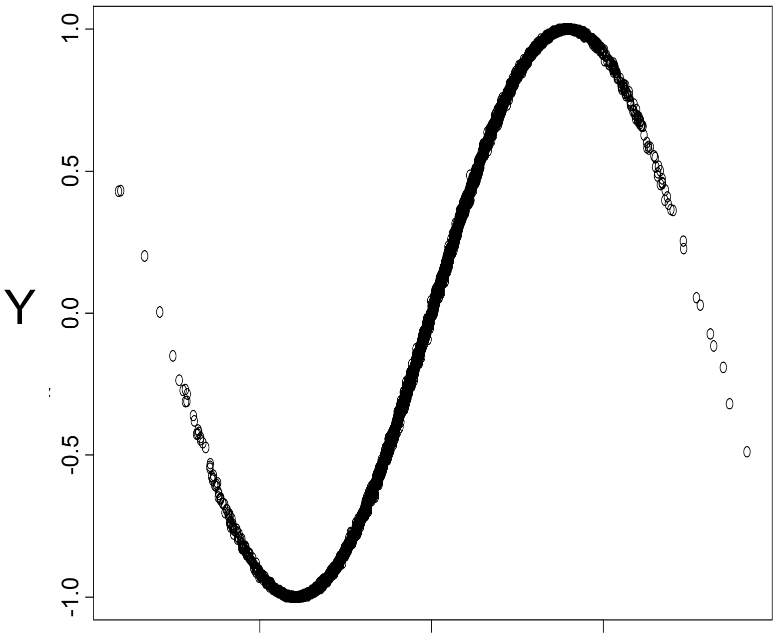
- robust location and covariance estimates
- robust PCA

→ less influenced by outliers
(Gather et al. 2001)

Example

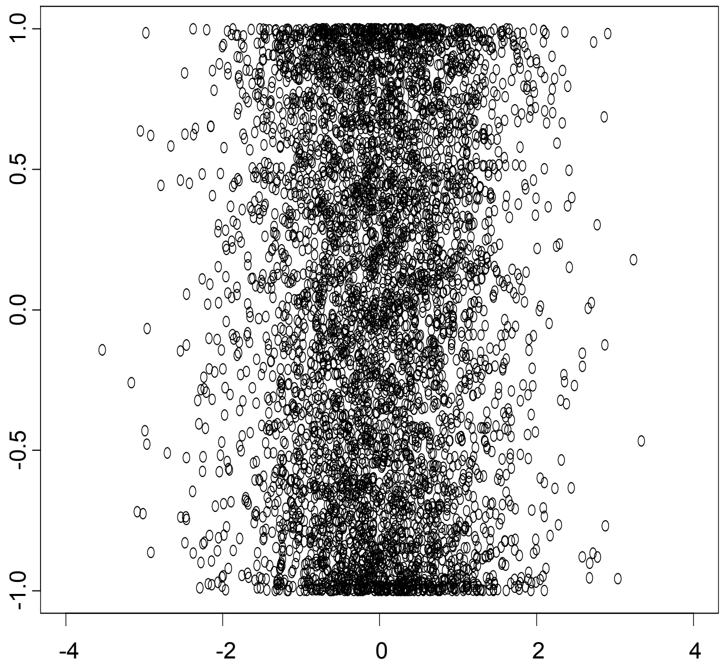
$d = 10,\ K = 1,\ Y = g(X_1,\ldots,X_{10}) = \sin(X_1)$



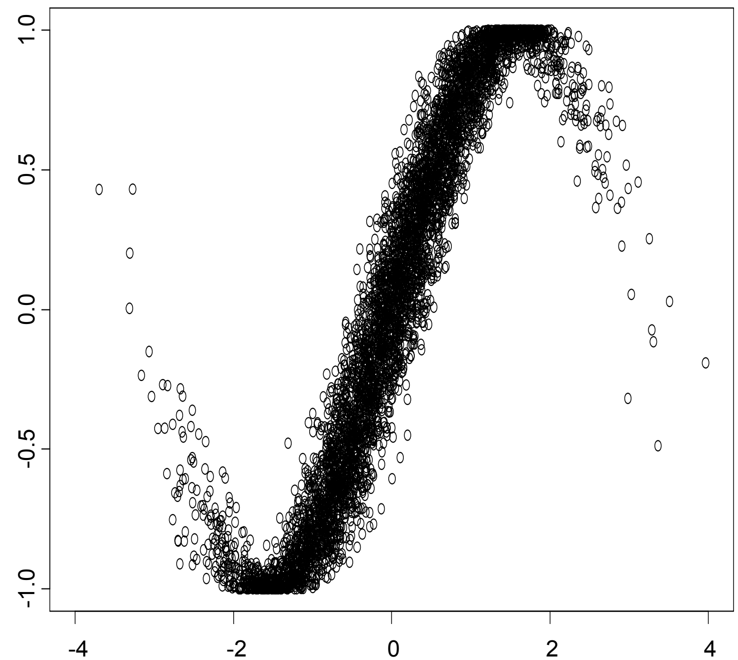


SIR: projection into $\mathcal{B}$

same data, but with one extreme outlier in X_1 direction



SIR: projection into $\mathcal{B}$



DAME: projection into $\mathcal{B}$

Robust nonparametric regression
(Davies and Kovac 2001)

Estimating f controlling the number of local extremes:

- $\hat{f}$ with k local extremes
- residuals $y - \hat{f}$ "look like white noise"
- take $\hat{f}$ with smallest k

➡ Run method

(run lenght of residual signs short enough)

➡ Taut-strings robustified

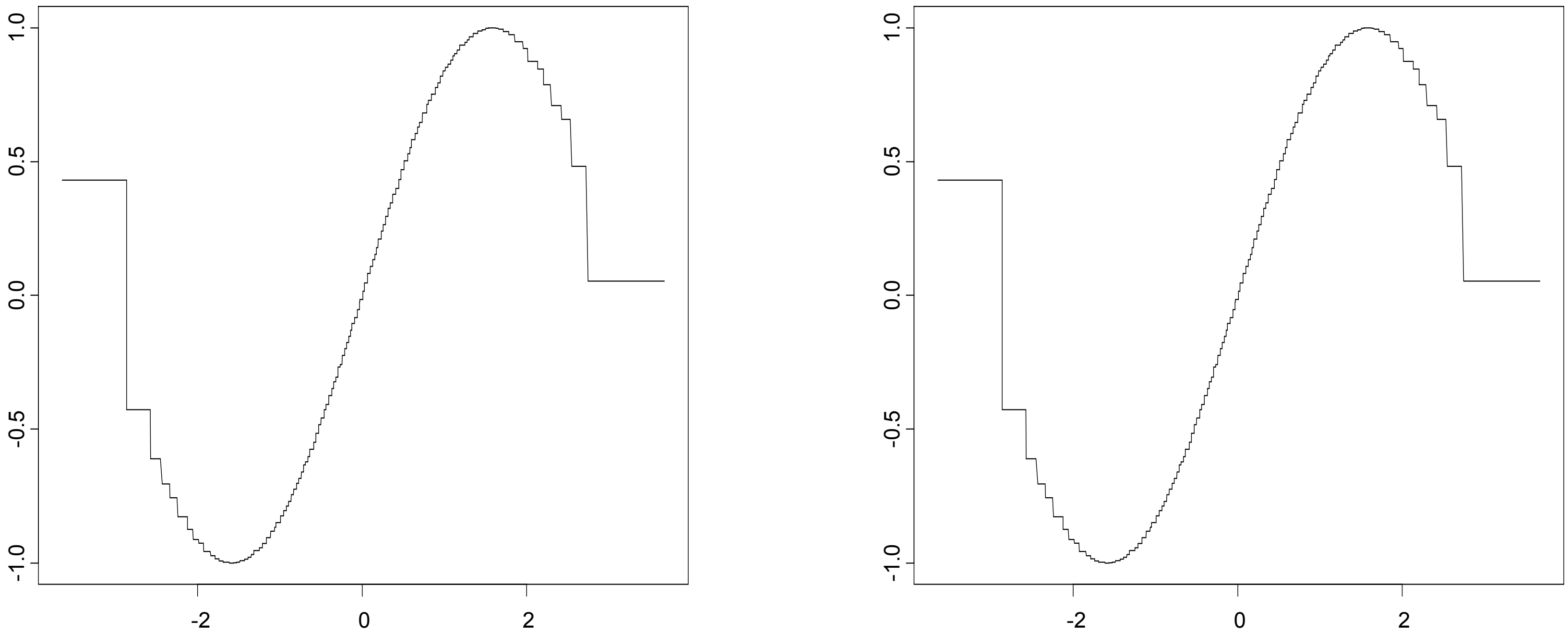
(absolute multiresolution coefficients small enough)

Both yield step function ⇨ smoothing as final step

Example

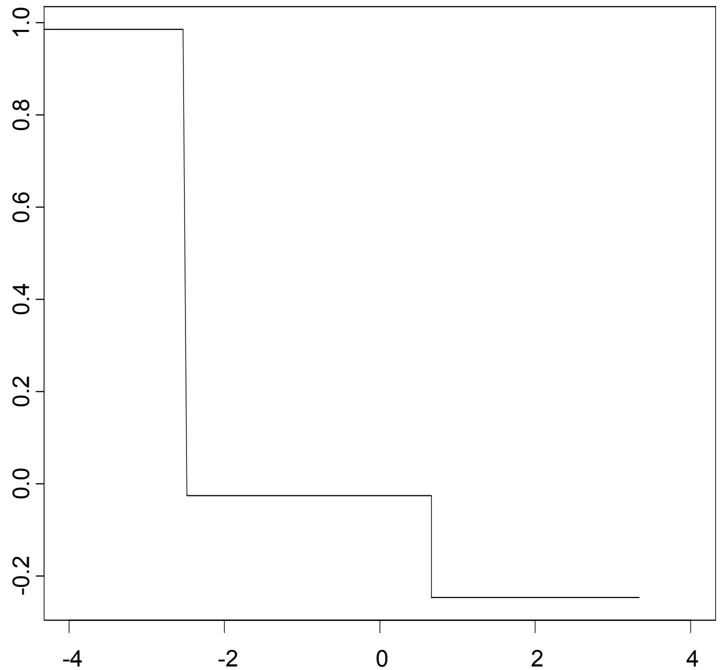
d = 10, K = 1, $Y = g(X_1,\ldots,X_{10}) = \sin(X_1)$

undisturbed data, projections by SIR and DAME

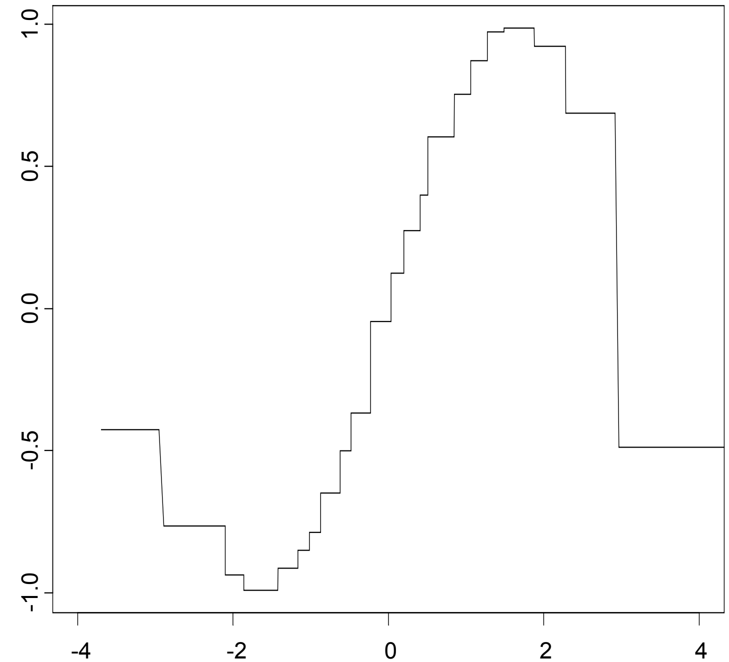


SIR DAME

same data, again with outlier in X_1 direction

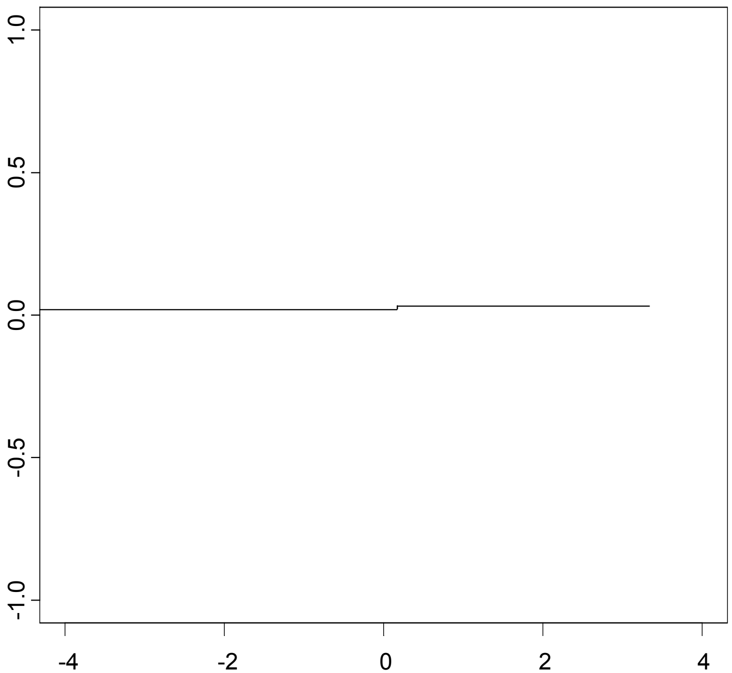


SIR

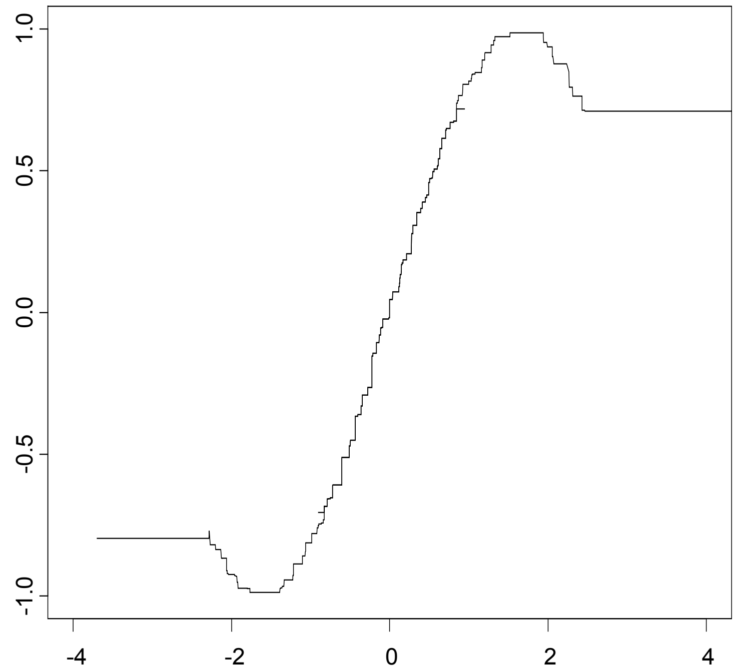


DAME

same data, again with outlier in X_1 direction



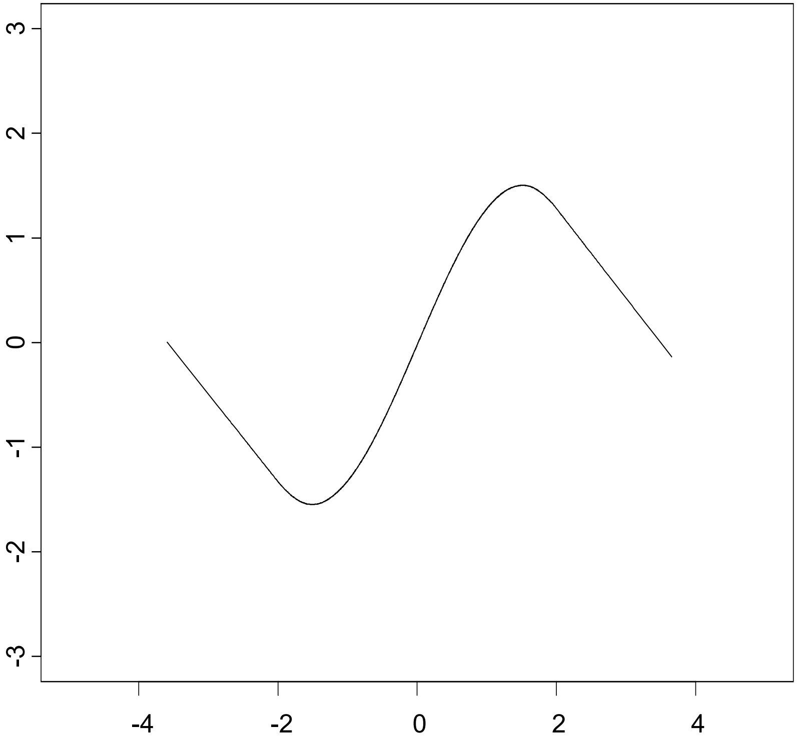
SIR



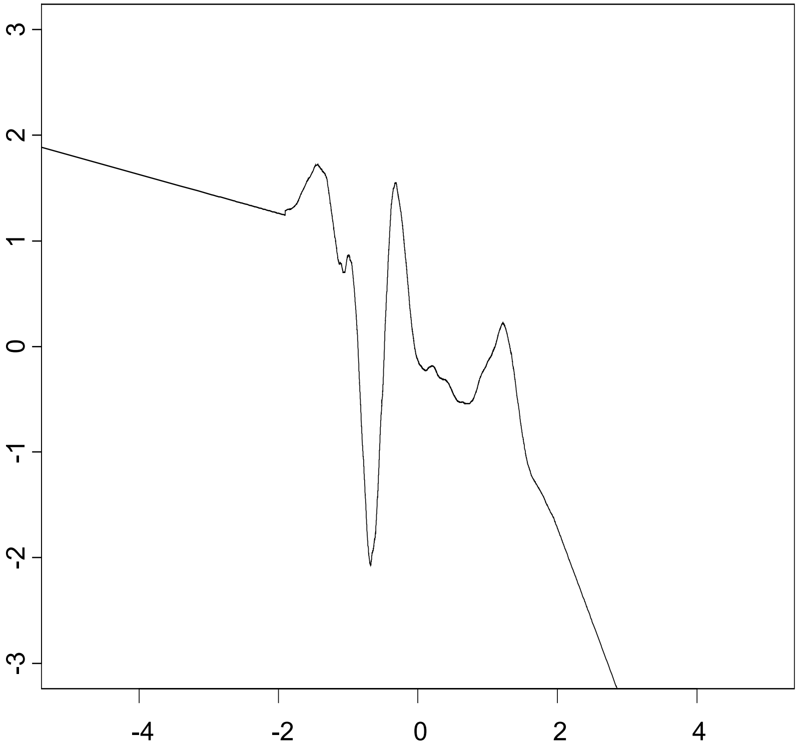
DAME

Comparison with results of PPR

undisturbed data



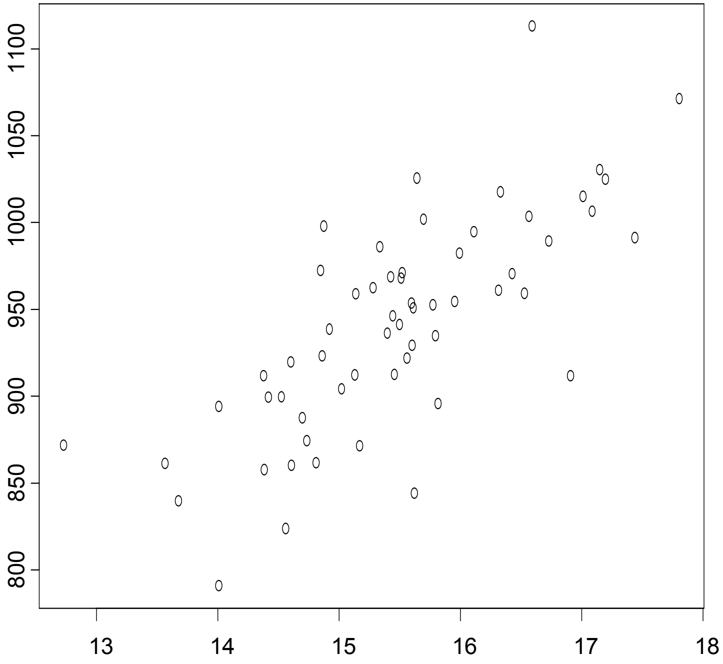
contaminated data



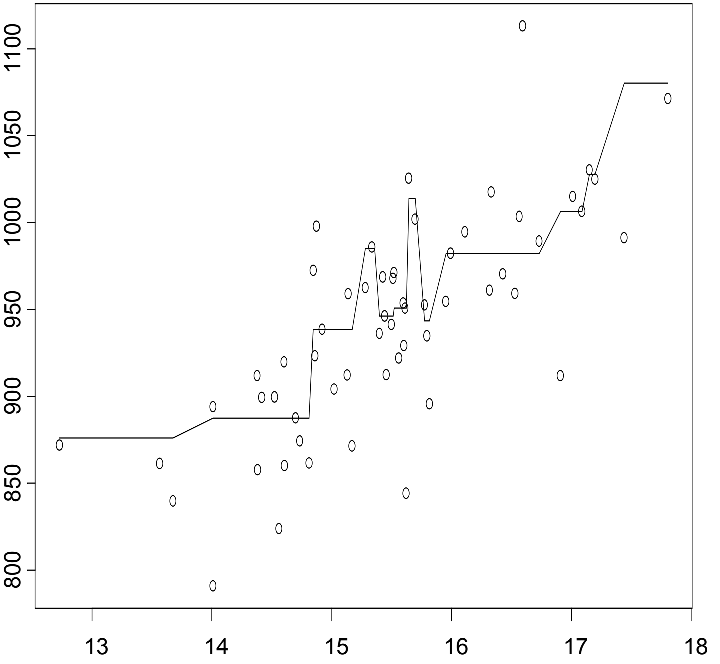
4. Example

Air pollution and mortality data

SIR

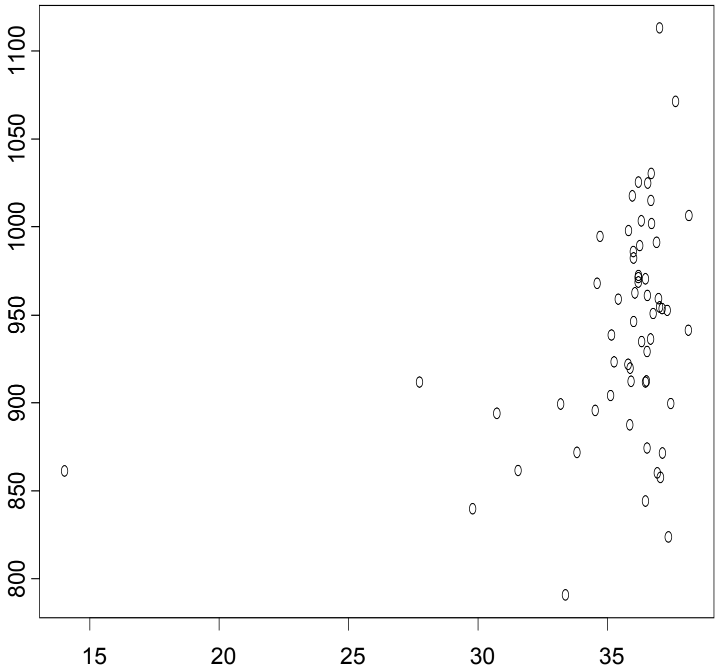


estimation of f

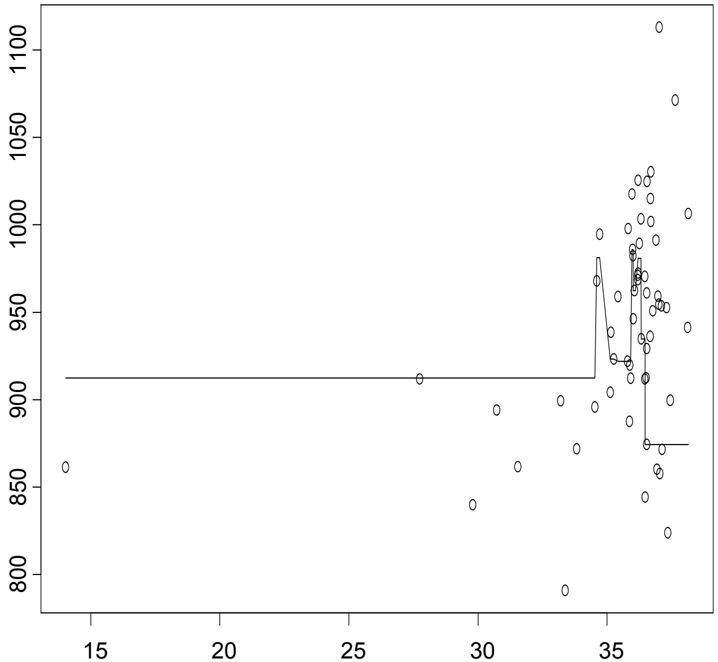


Air pollution and mortality data

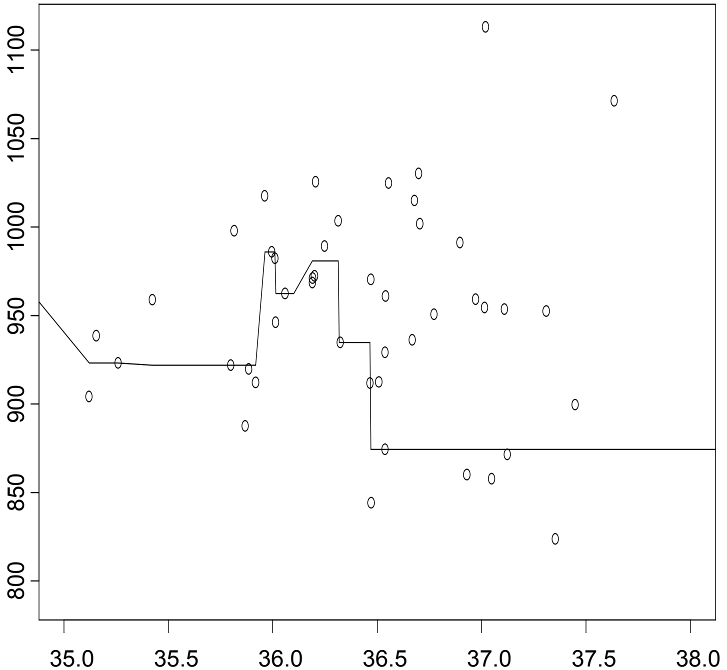
DAME

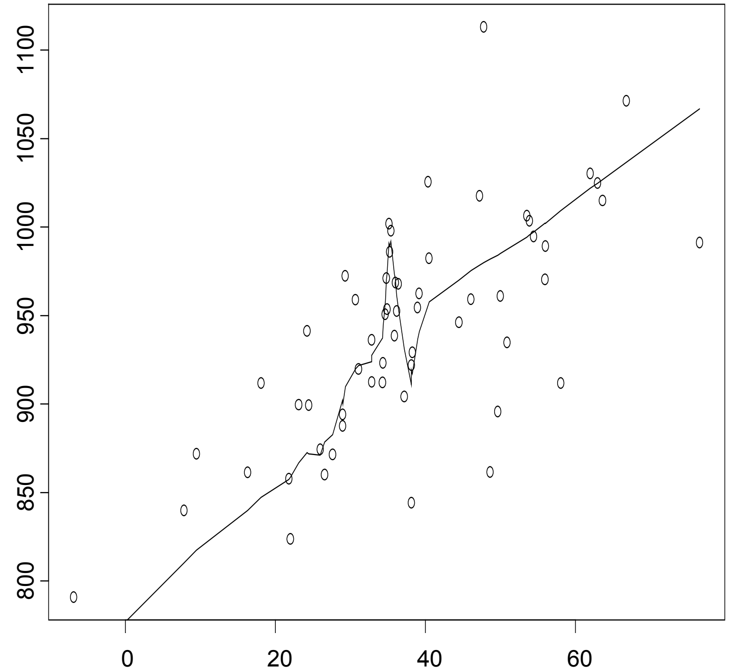


estimation of f



Comparison with results of PPR





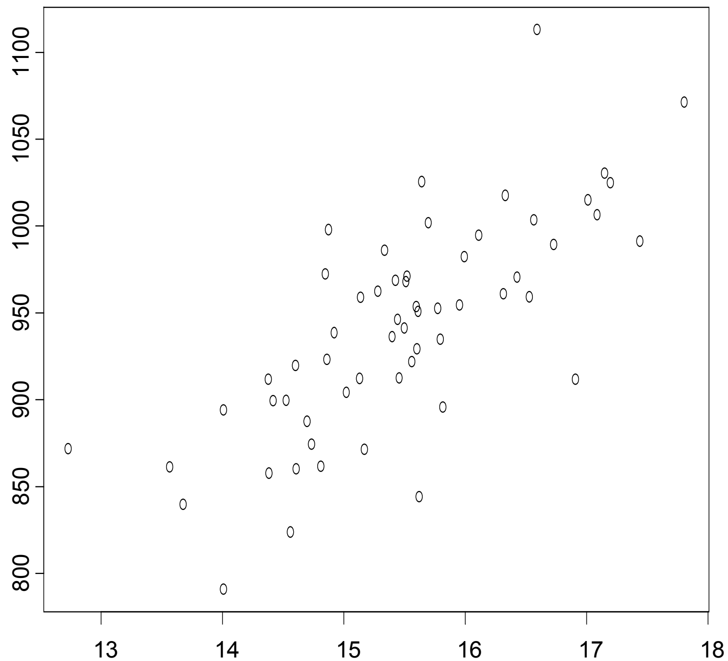
robust dimension adjustment

projection pursuit regression

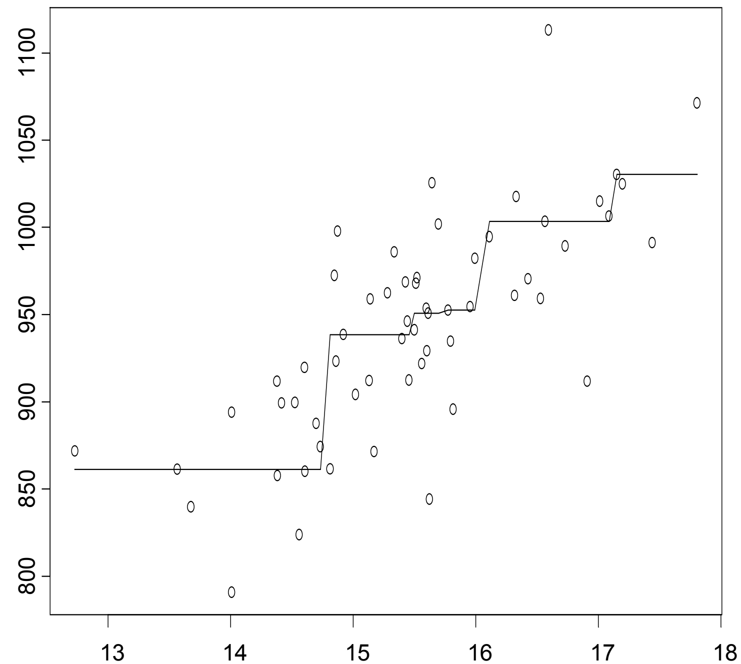
4. Example

Air pollution and mortality data

SIR

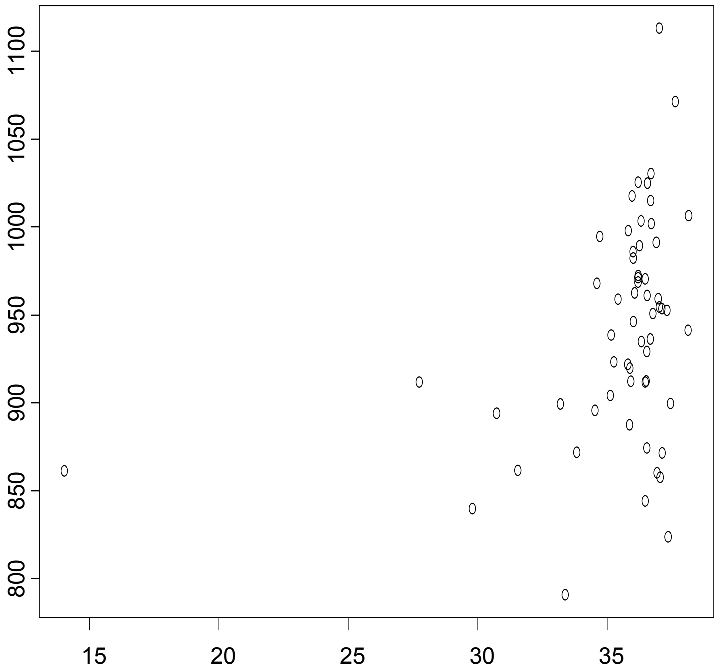


estimation of f

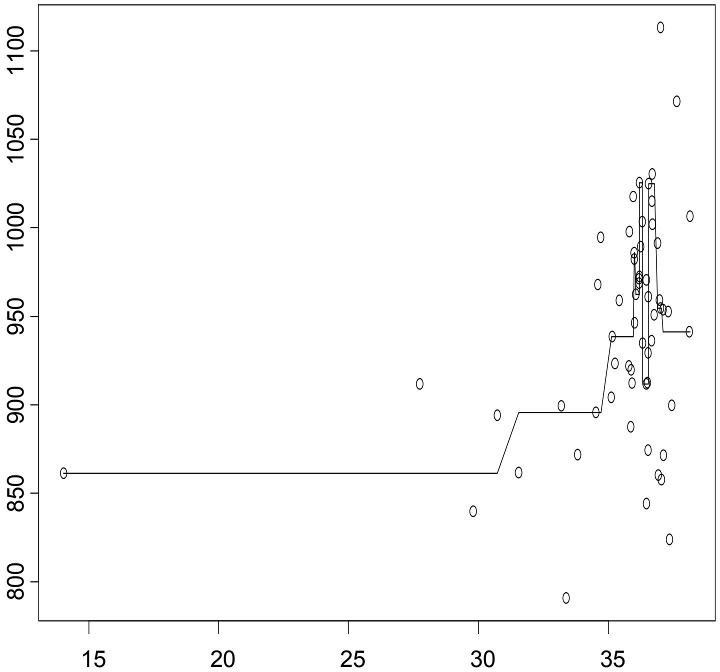


Air pollution and mortality data

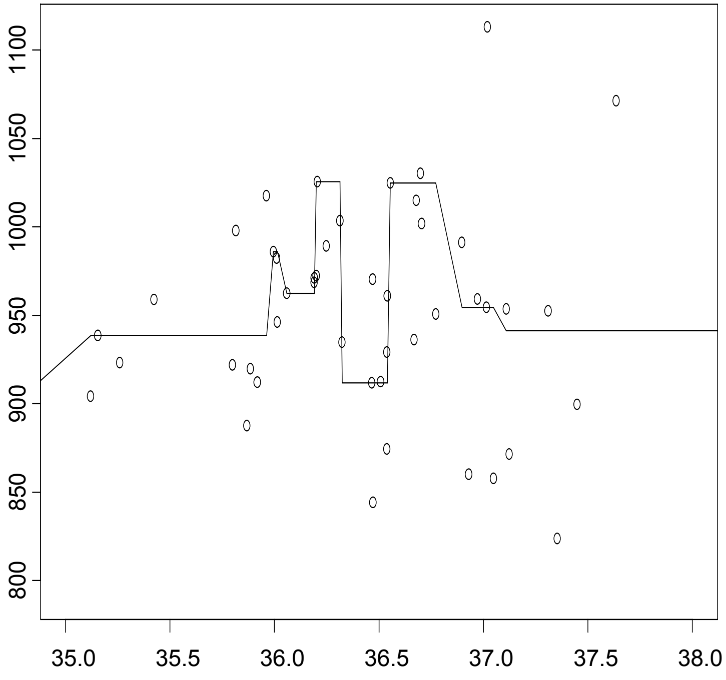
DAME



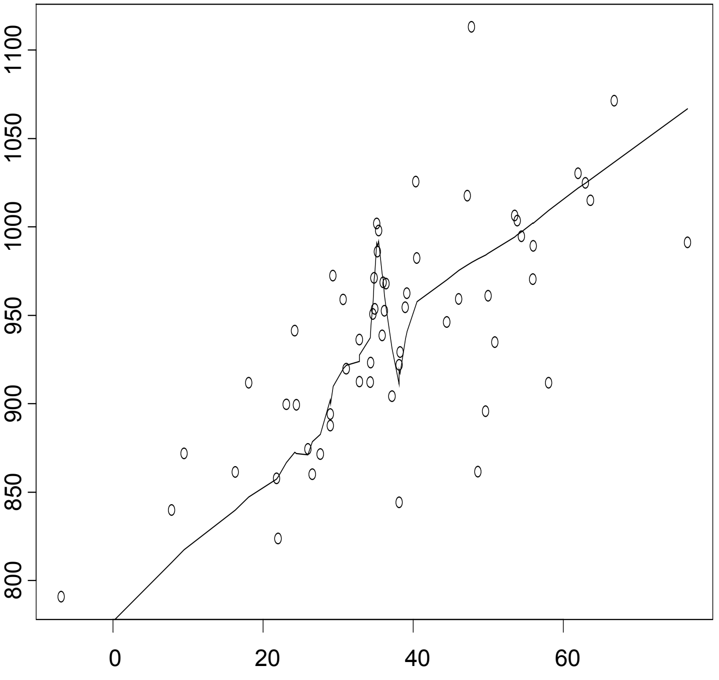
estimation of f



Comparison with results of PPR



robust dimension adjustment



projection pursuit regression

Conclusion and further work

- Dimension adjustment methods useful for high-dimensional data
- Robustness necessary in both, dimension reduction and function estimation
- Problem of outliers in X
  ⇨ outlier identification + rejection in dimension reduction possible
- Final step: smoothing
  (work in progress, Majidi 2001)
- Possible alternative: robust PPR