M7777 Applied Functional Data Analysis
12. Sparse FDA

Jan Koláček (kolacek@math.muni.cz)

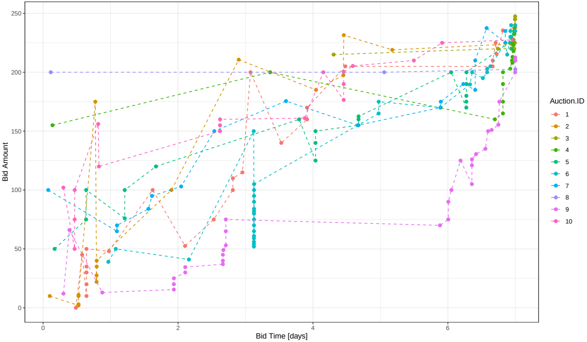
Dept. of Mathematics and Statistics, Faculty of Science, Masaryk University, Brno

Sparse FDA

Ebay Auctions

Jank and Shmueli, 2007

- 7-Day auctions for new Palm M515 PDAs
- 149 Auctions, collected May-June 2003

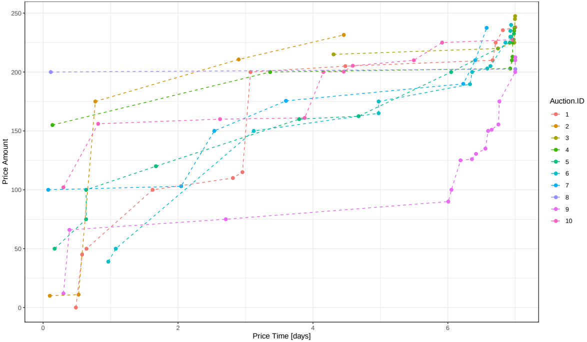


Sample of 10 auctions – bid histories.

Sparse FDA

Auction Price

- Only increases if bid is greater than current price



Sample of 10 auctions – price histories.

Sparse FDA

We will consider a model

$$Y_{ij} = \underbrace{\mu(t_{ij}) + \varepsilon_i(t_{ij})}_{X_i(t_{ij})} + \delta_{ij},$$

for $1 \leq i \leq n$, $1 \leq j \leq n_i$, with assumptions

$\mu(t)$ ... the mean function (required to be smooth)

$\varepsilon_i(t)$ ... subject specific error functions, induce correlation between observations on the same subject, let's denote $c(s,t) = \text{Cov}(X(s), X(t)) = \text{Cov}(\varepsilon(s), \varepsilon(t))$

δ_{ij} ... errors explaining measurement noise, iid across both i and j, let's denote $\text{Var}(\delta_{ij}) = \sigma^2(t_{ij})$.

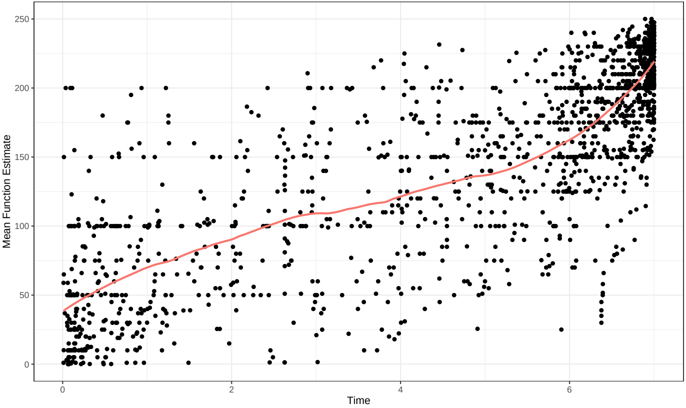
It means, that we observe a process $Y(t)$ in n samples $X_i(t)$, the i-th sample is observed in times $t_1, \dots, t_{n_i}$ with setting

$$\text{Cov}(Y(s), Y(t)) = c(s,t) + \sigma^2(s) I_{s=t}.$$

Sparse FDA

The Main Idea

1. Let us consider all measurements Y_{ij}, $1 \leq i \leq n$, $1 \leq j \leq n_i$
2. Get an estimate $\hat{\mu}(t)$ of the mean function $\mu(t)$ (nonparametric, e.g. local linear kernel smoother, spline smoothing etc.)



Sparse FDA

3. Let us consider a set of time points pairs

$$\mathbf{T} = \{(t_{ij_1}, t_{ij_2}) : 1 \leq i \leq n, 1 \leq j_1 \leq n_i, 1 \leq j_2 \leq n_i, j_1 \neq j_2\}$$

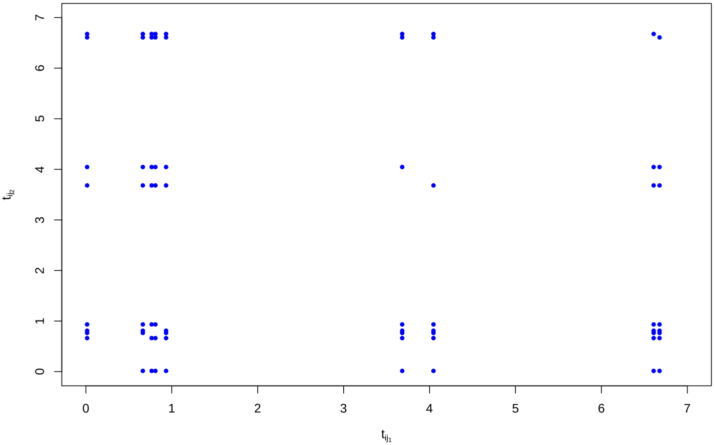
with its values

$$Z(t_{ij_1}, t_{ij_2}) = (Y_{ij_1} - \hat{\mu}(t_{ij_1}))(Y_{ij_2} - \hat{\mu}(t_{ij_2})),\ (t_{ij_1}, t_{ij_2}) \in \mathbf{T}$$

and get the covariance surface estimate $\hat{c}(s, t)$ (bivariate local linear etc.).

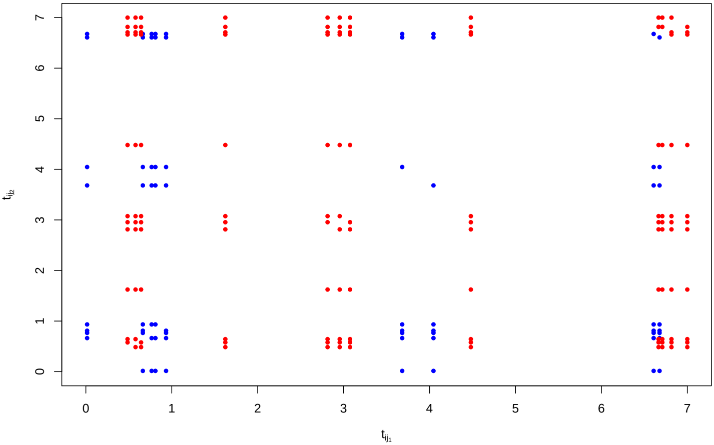
Sparse FDA

Samples: 1

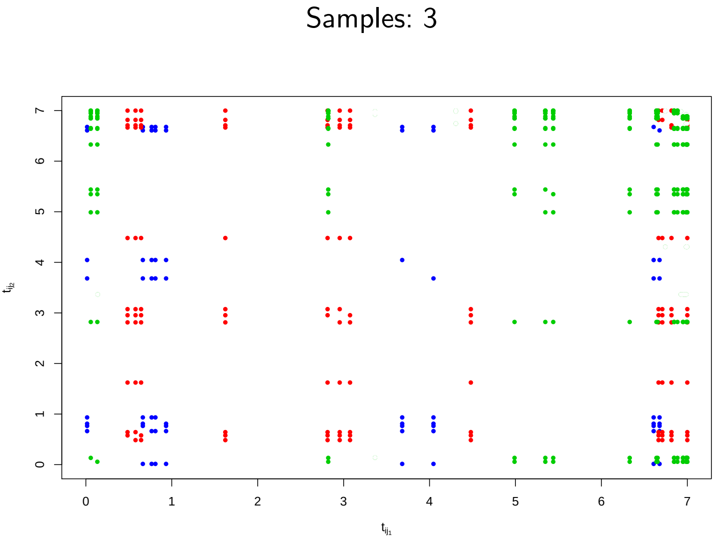


Sparse FDA

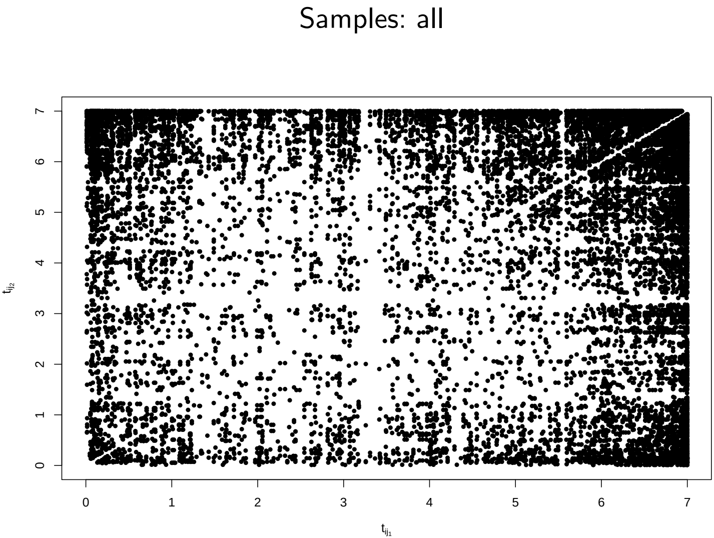
Samples: 2



Sparse FDA

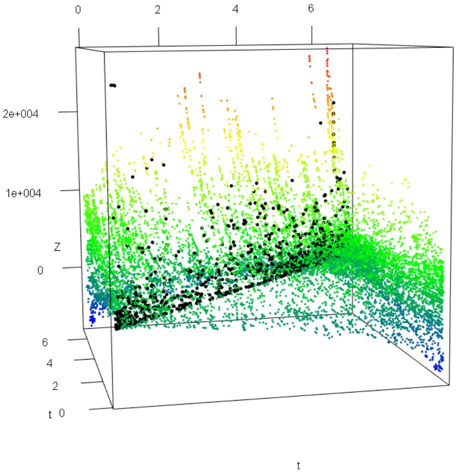


Sparse FDA



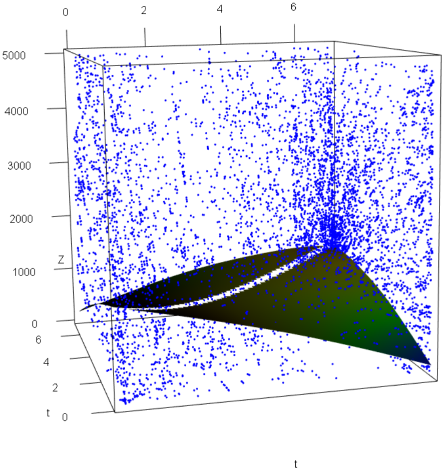
Sparse FDA

Auction Price



Raw Covariance Plot, black – diagonal terms

Auction Price



Covariance Estimate $\hat{c}(s, t)$, diagonal exluded

Sparse FDA

4. Take diagonal terms only

$$\mathbf{T}_{diag} = \{(t_{ij}, t_{ij}) : 1 \leq i \leq n, 1 \leq j \leq n_i\} \text{ and its } Z(t_{ij}, t_{ij})$$

and by a univariate smoother get $\tilde{c}(t,t)$. Thus, an estimate of $\sigma^2(t)$

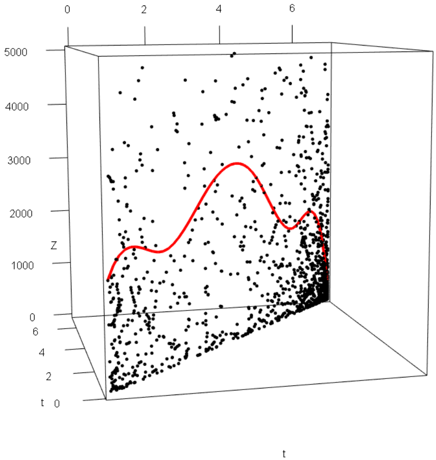
$$\hat{\sigma}^2(t) = \tilde{c}(t,t) - \hat{c}(t,t).$$

5. The estimate of $\mathrm{Cov}(Y(s), Y(t))$ takes the form

$$\hat{\sigma}(s,t) = \hat{c}(s,t) + \hat{\sigma}^2(t)$$

Sparse FDA

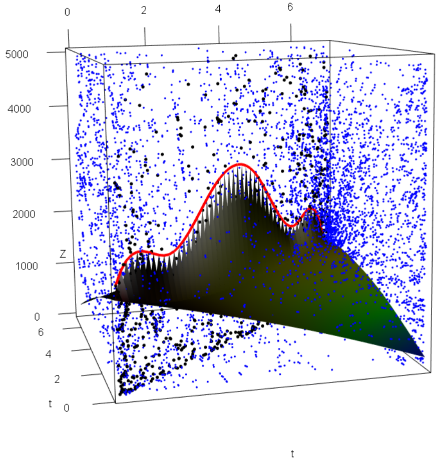
Auction Price



Variance Estimate $\hat{\sigma}^2$

Sparse FDA

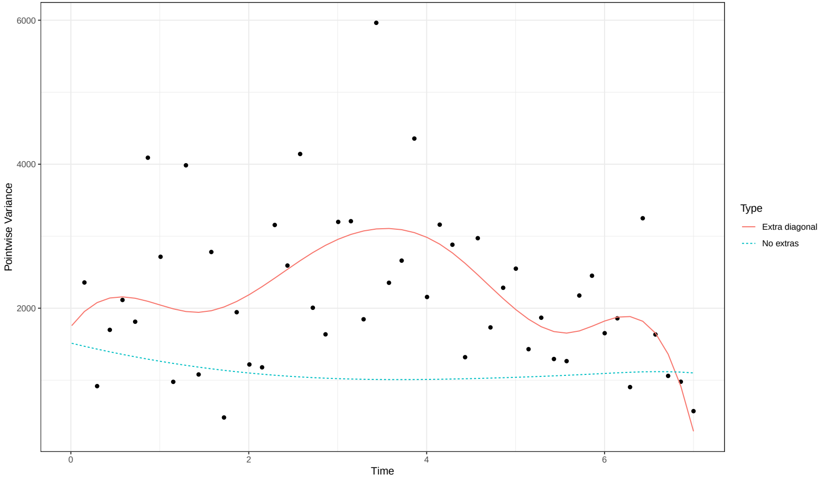
Auction Price



Covariance Estimate $\hat{\sigma}(s,t) = \hat{c}(s,t) + \hat{\sigma}^2(t)$

Sparse FDA

Auction Price



Comparison of Variance Estimates

Sparse FDA

6. Let's consider the estimate of $\hat{\sigma}(s,t)$ and its Karhunen – Loève decomposition for functions

$$\hat{\sigma}(s,t) = \sum_{j=1}^{\infty} \lambda_j \xi_j(s) \xi_j(t) \quad \Rightarrow \text{ obtain } \hat{\xi}_j(t), \hat{\lambda}_j,\ j = 1, \ldots, K.$$

7. Estimate principal scores $c_{ij} = \int \xi_j(t)[Y_i(t) - \mu(t)]dt$ through the conditional expectation

$$\hat{c}_{ij} = \mathrm{E}[c_{ij}|\mathbf{Y}_i] = \hat{\lambda}_j \hat{\boldsymbol{\xi}}_j^T \hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)$$

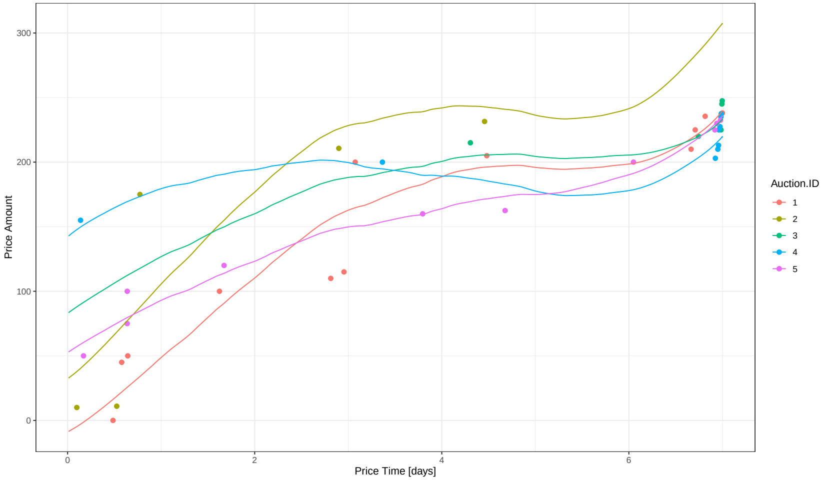
Yao et al. (2005)

8. Finally, reconstruct the whole curves

$$\widehat{Y}_i(t) = \hat{\mu}(t) + \sum_{j=1}^{K} \hat{c}_{ij} \hat{\xi}_j(t).$$

Filled Data

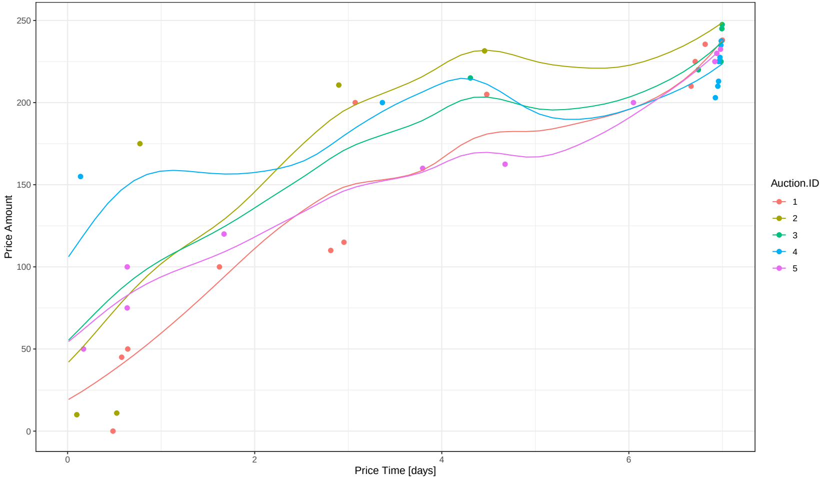
Auction Price – proposed method



Auction Prices Estimates

Filled Data

Auction Price – FDAPACE



Auction Prices Estimates

Kernel Smoothing

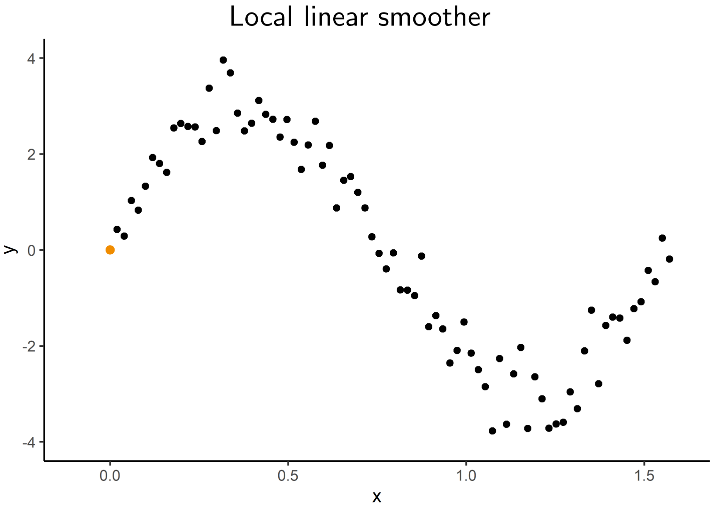
Mean function estimate

Local linear smoother with global bandwidth

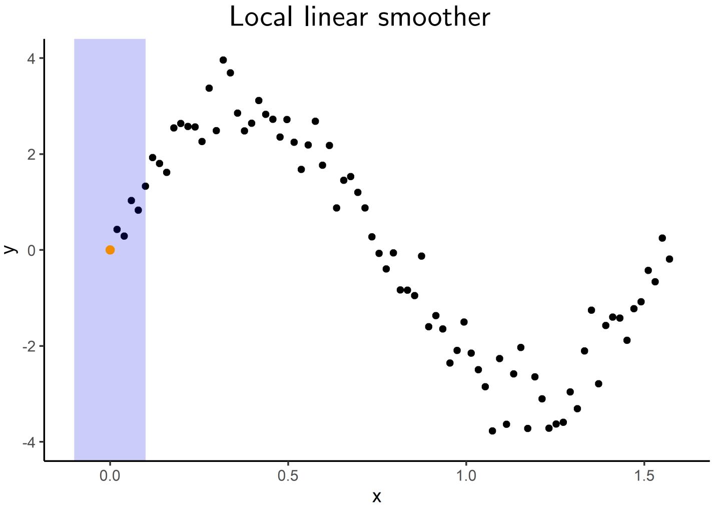
$$\sum_{i=1}^{n} \sum_{j=1}^{N_i} \left[K\left(\frac{T_{ij} - t}{h} \right) Y_{ij} - \beta_0 - \beta_1 (t - T_{ij}) \right]^2 \rightarrow \min$$

- $K(x)$ ... **kernel** function (a symmetric density)
- h ... **global** bandwidth
- $\hat{\mu}(t) = \hat{\beta}_0(t)$

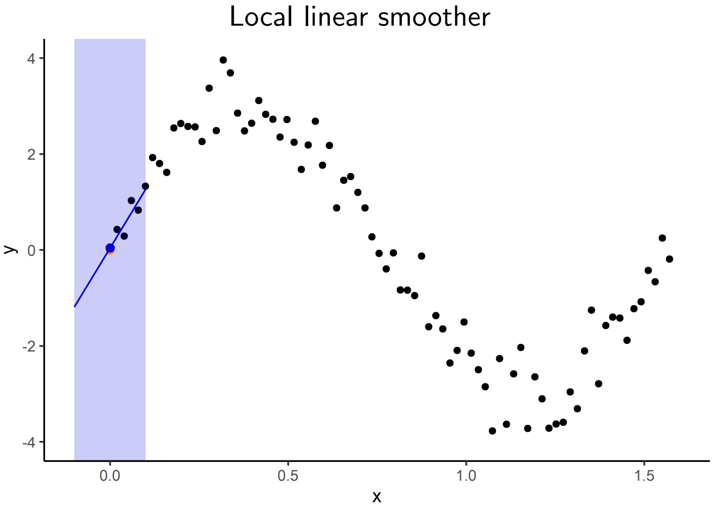
Kernel Smoothing



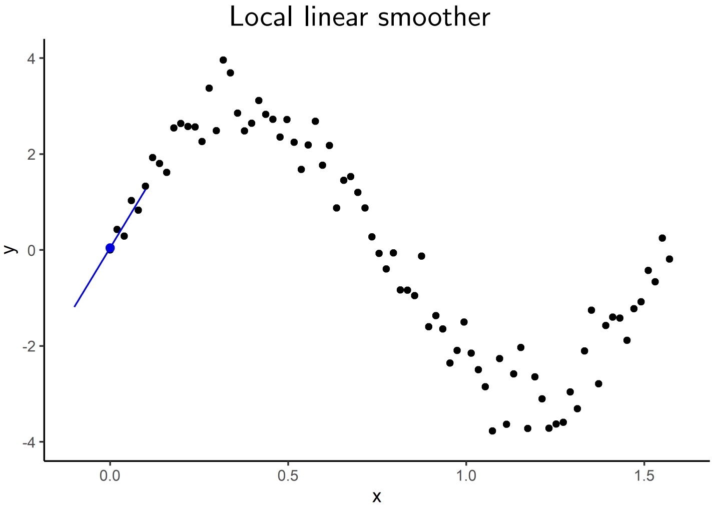
Kernel Smoothing



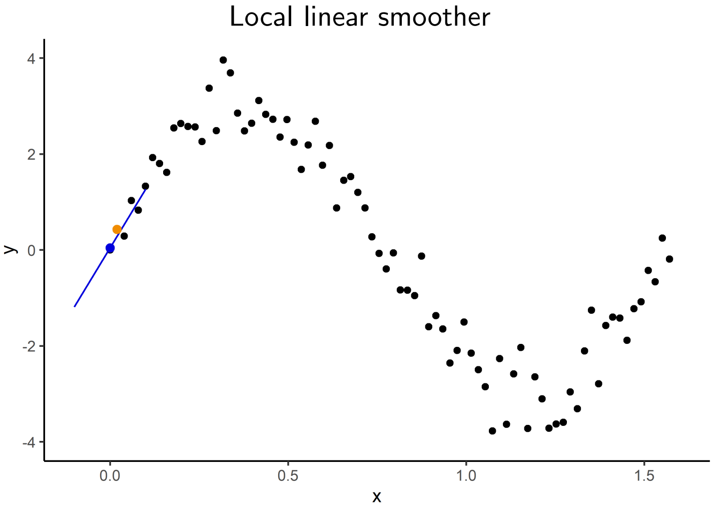
Kernel Smoothing



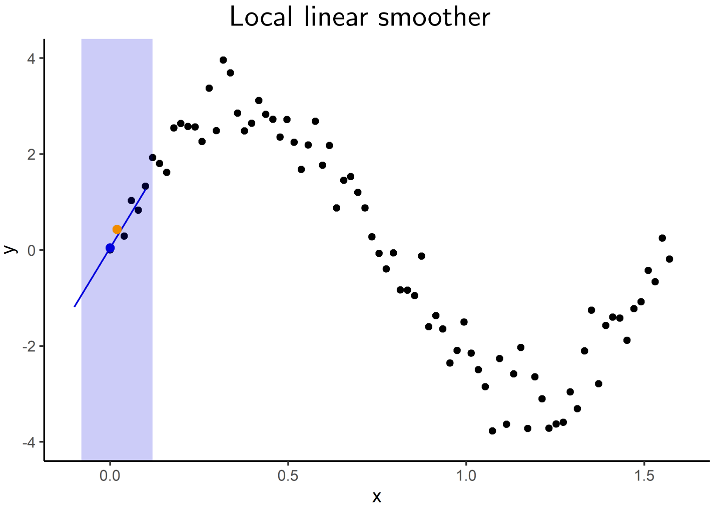
Kernel Smoothing



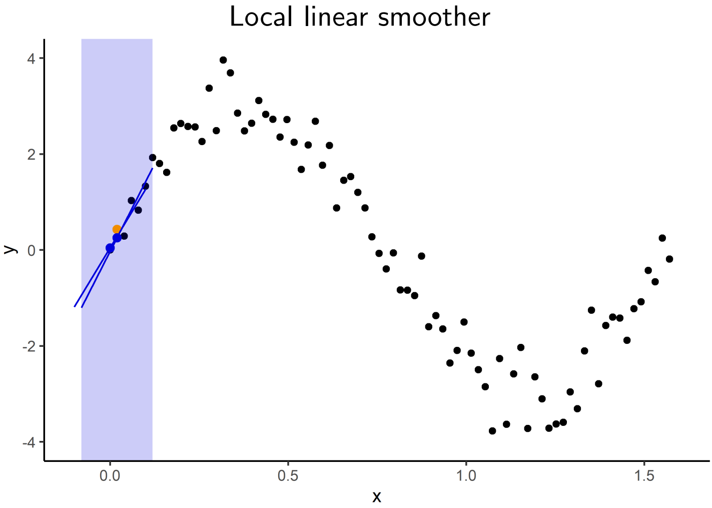
Kernel Smoothing



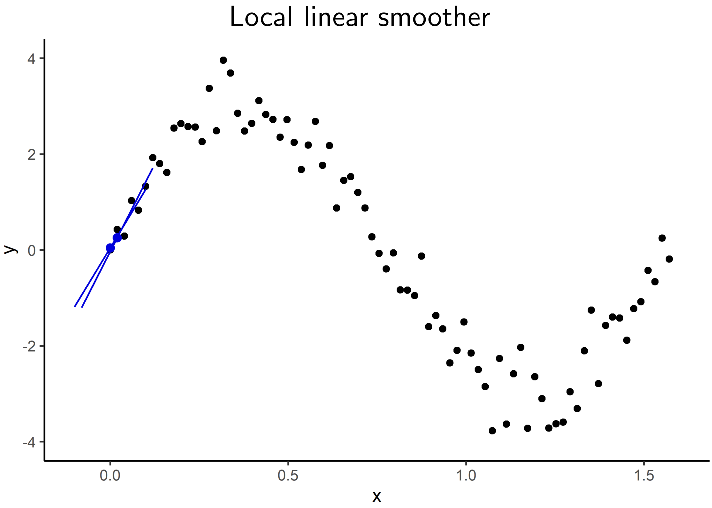
Kernel Smoothing



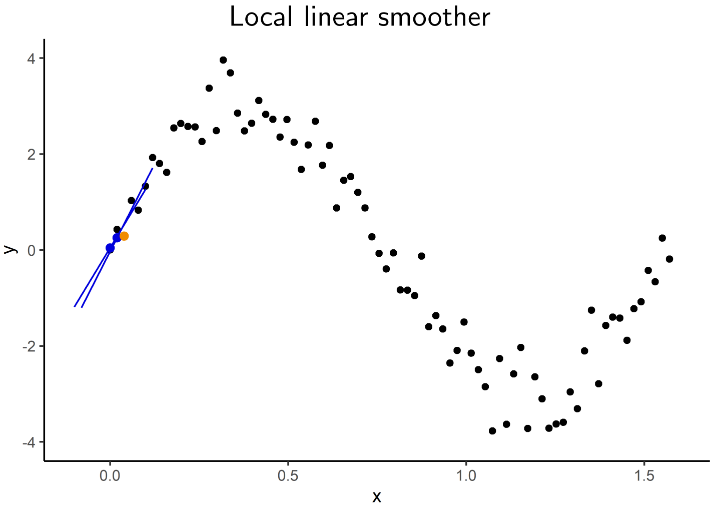
Kernel Smoothing



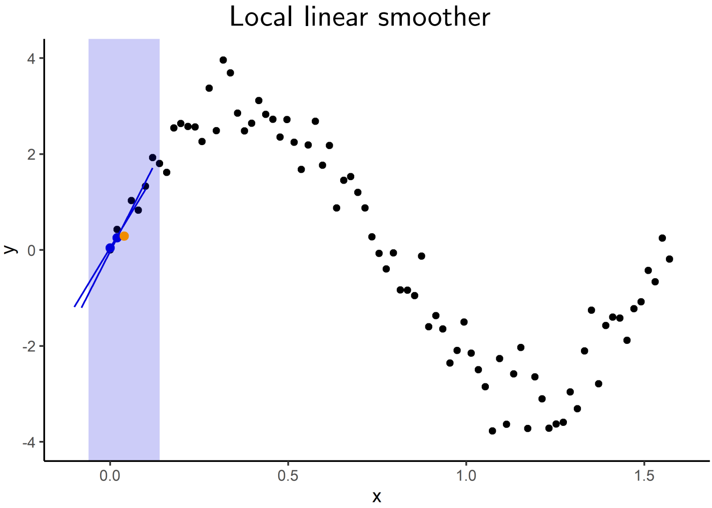
Kernel Smoothing



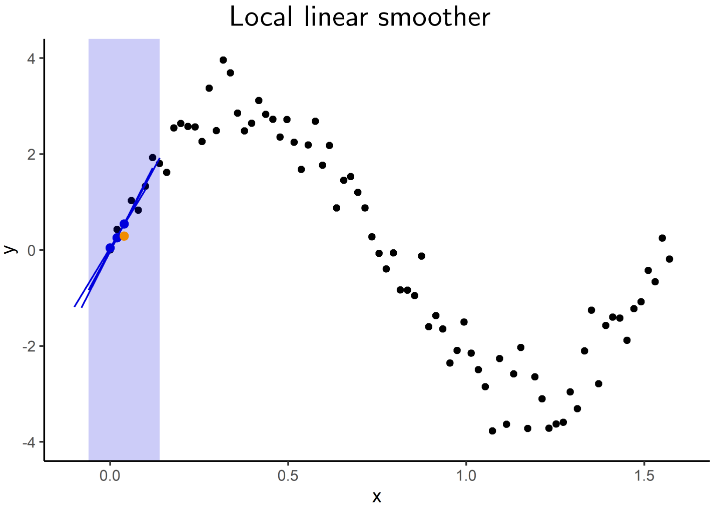
Kernel Smoothing



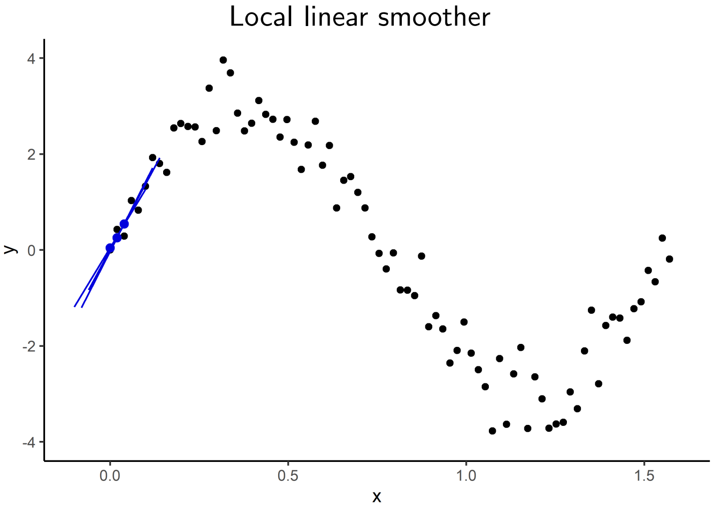
Kernel Smoothing



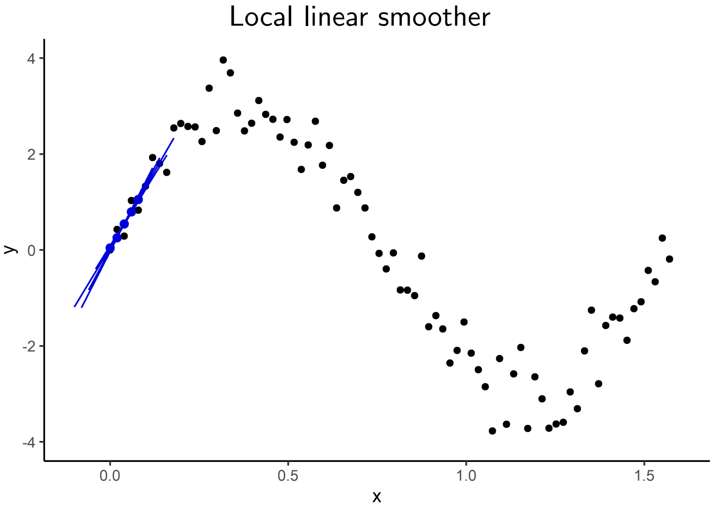
Kernel Smoothing



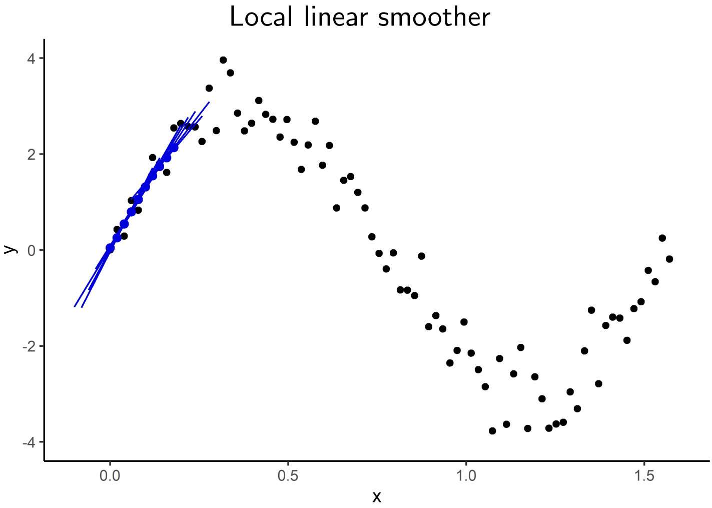
Kernel Smoothing



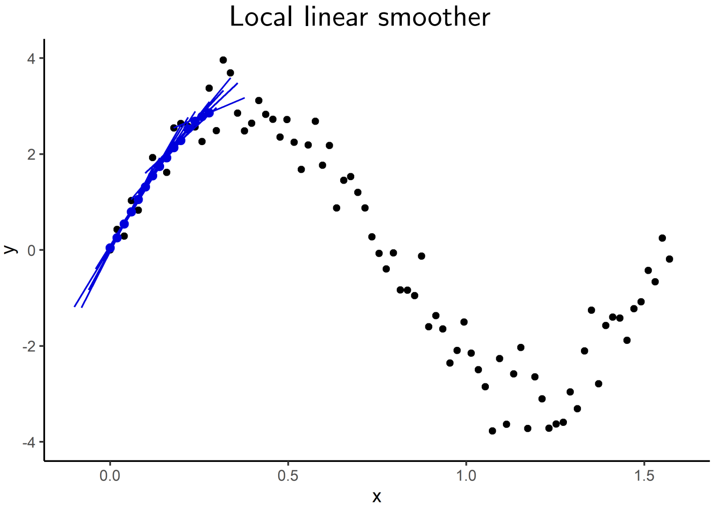
Kernel Smoothing



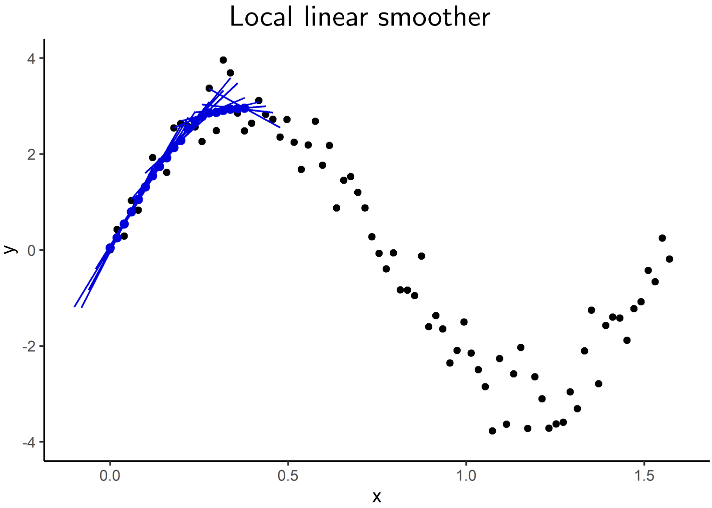
Kernel Smoothing



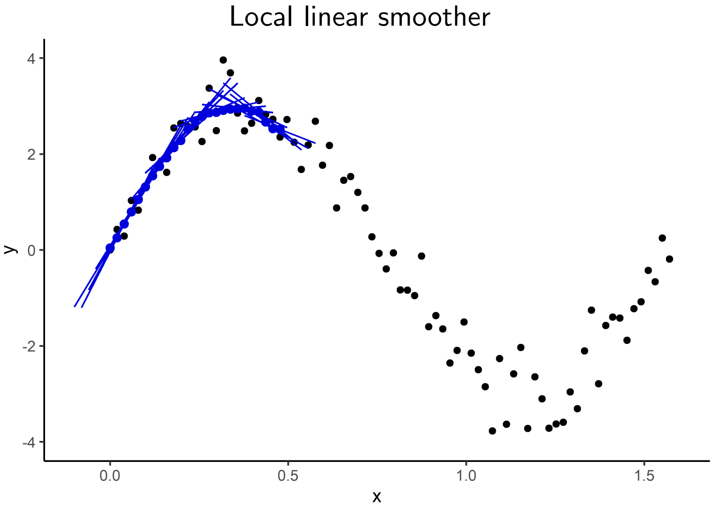
Kernel Smoothing



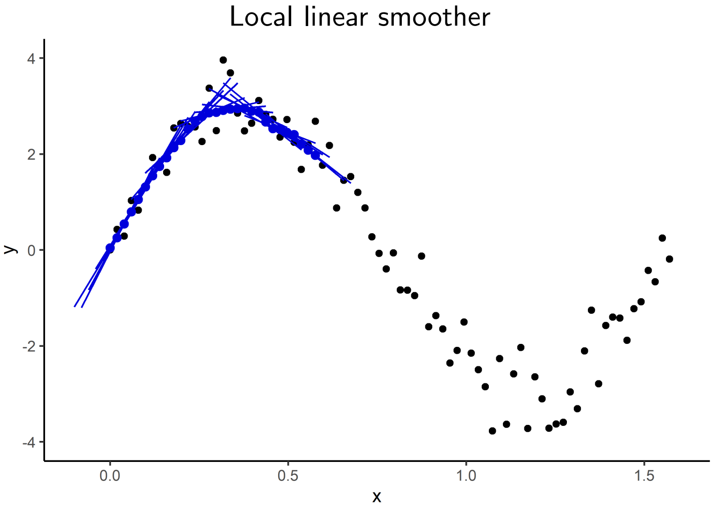
Kernel Smoothing



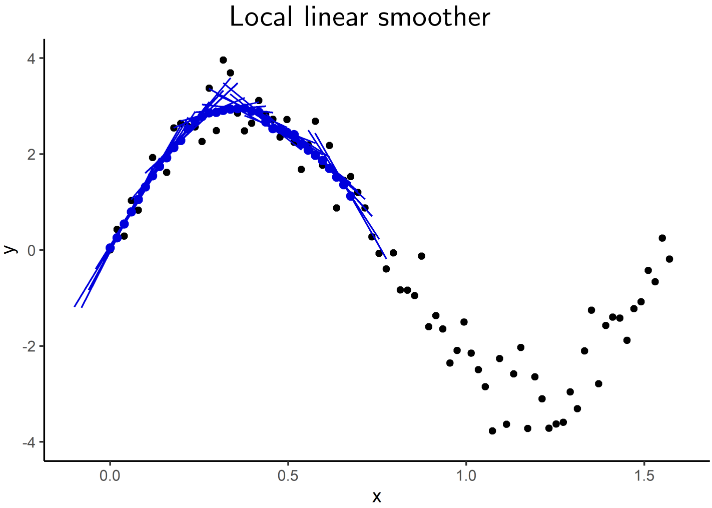
Kernel Smoothing



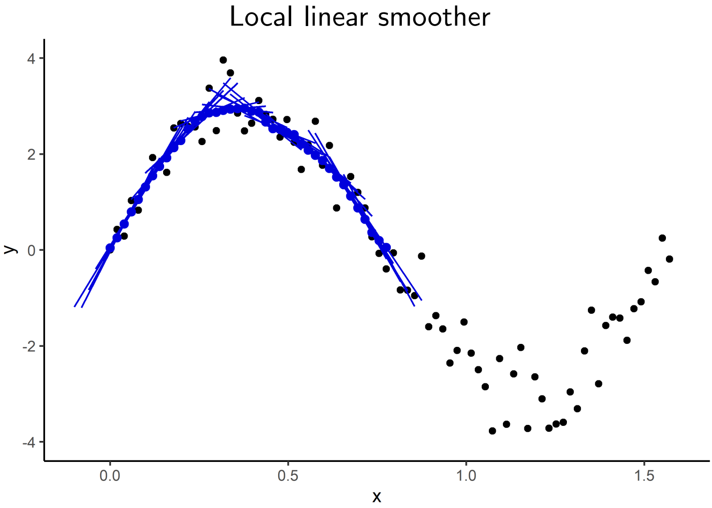
Kernel Smoothing



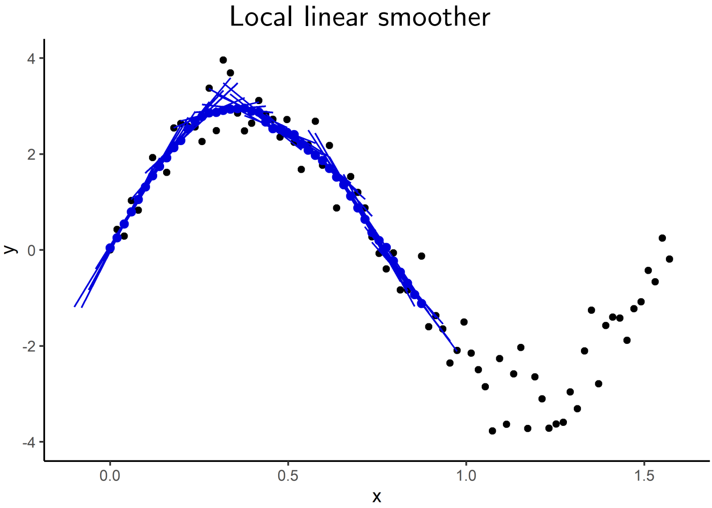
Kernel Smoothing



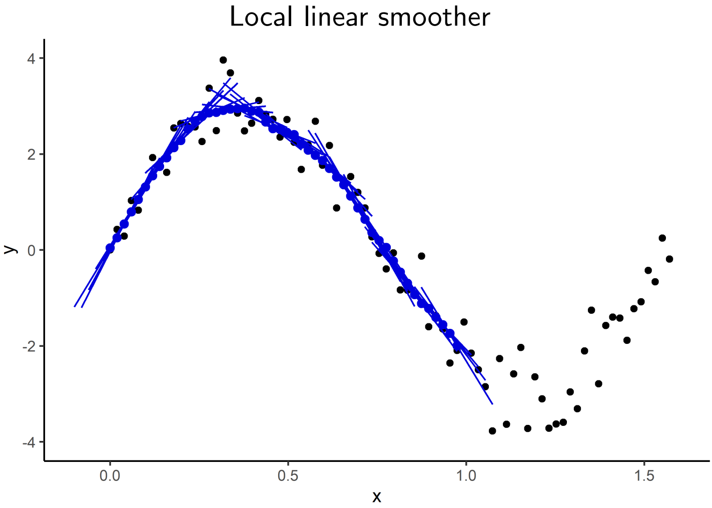
Kernel Smoothing



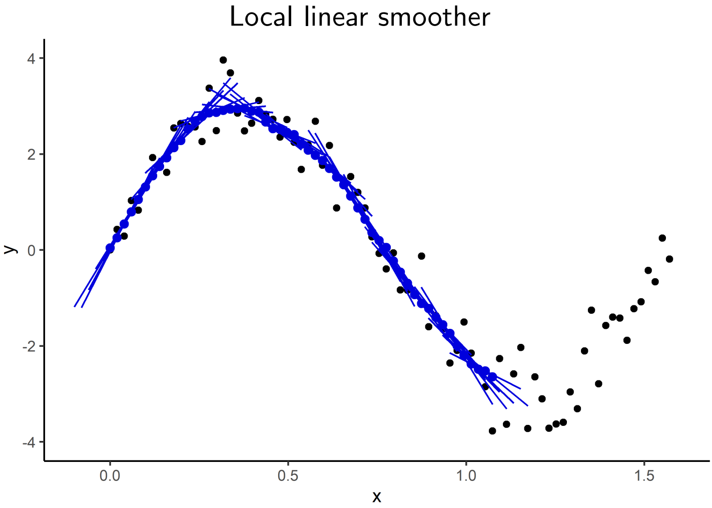
Kernel Smoothing



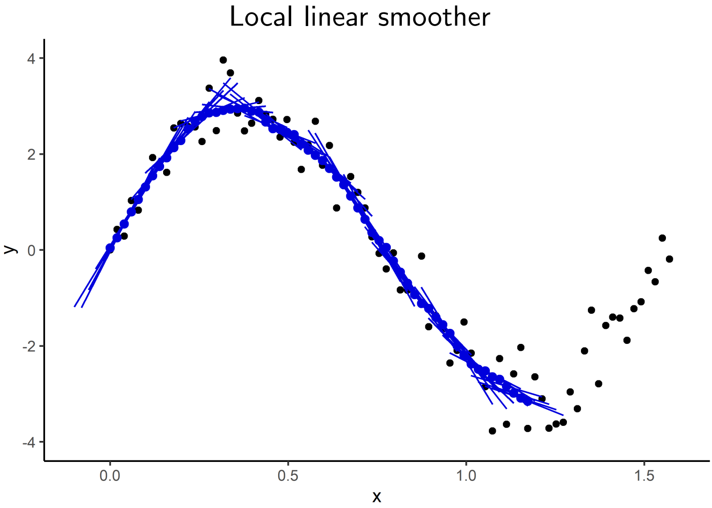
Kernel Smoothing



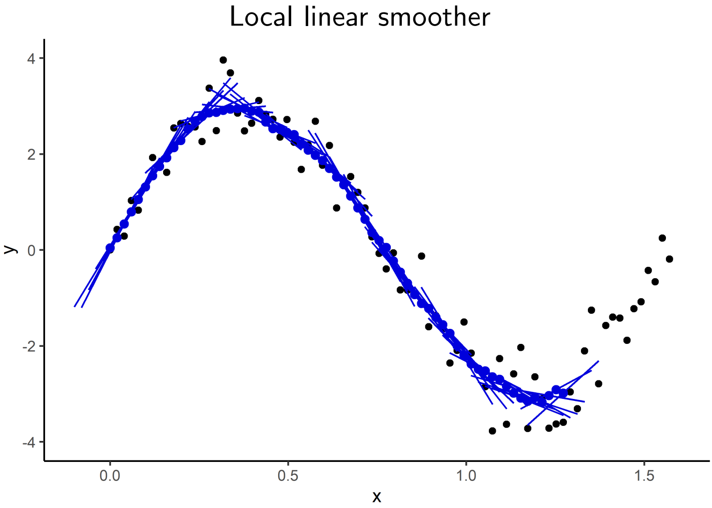
Kernel Smoothing



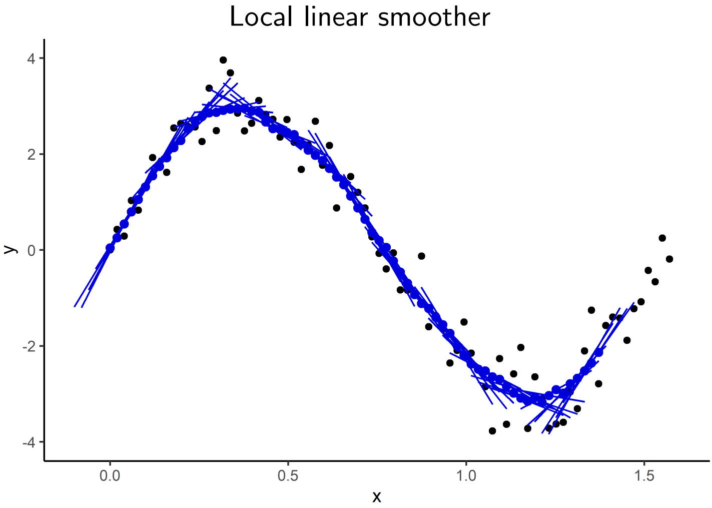
Kernel Smoothing



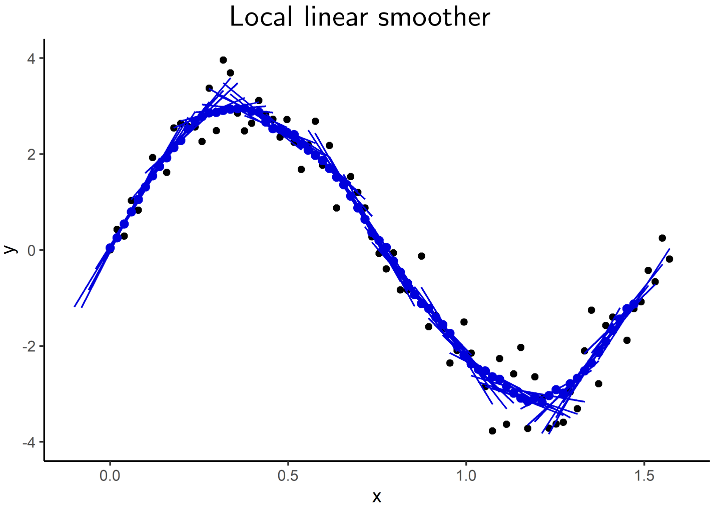
Kernel Smoothing



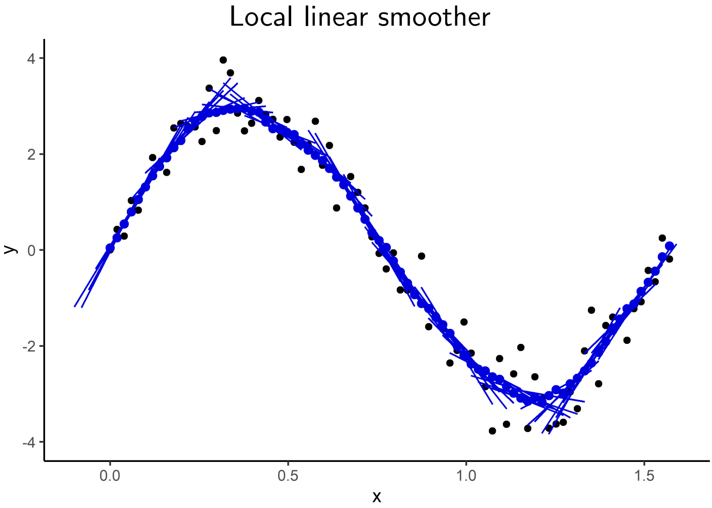
Kernel Smoothing



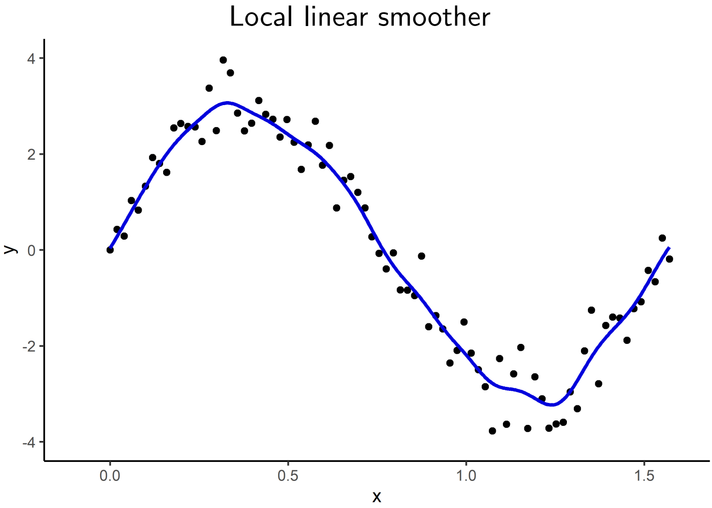
Kernel Smoothing



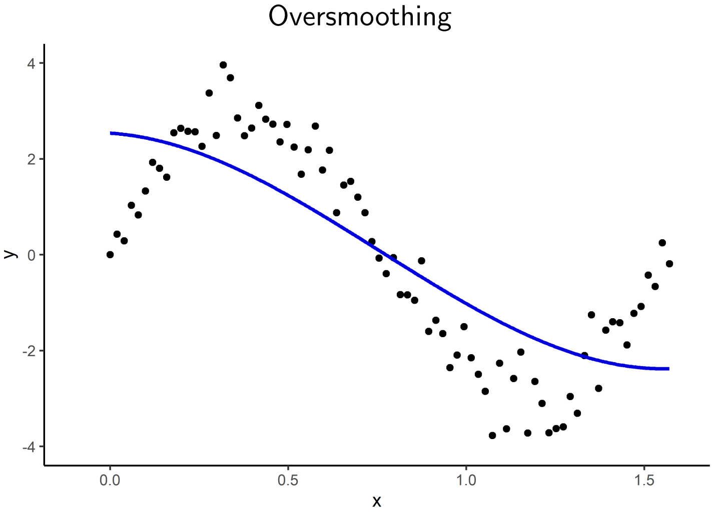
Kernel Smoothing



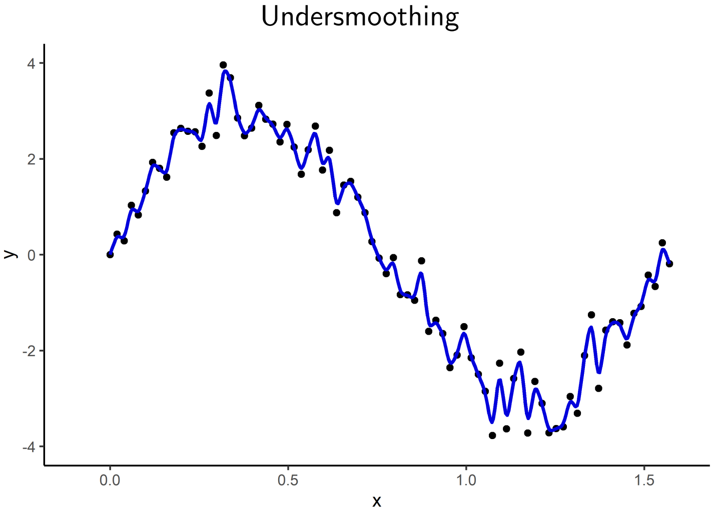
Kernel Smoothing



Kernel Smoothing



Kernel Smoothing



Kernel Smoothing

Mean function estimate

Local linear smoother with **global** bandwidth

$$\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left[K\left(\frac{T_{ij}-t}{h}\right)Y_{ij}-\beta_0-\beta_1(x-T_{ij})\right]^2\to\min$$

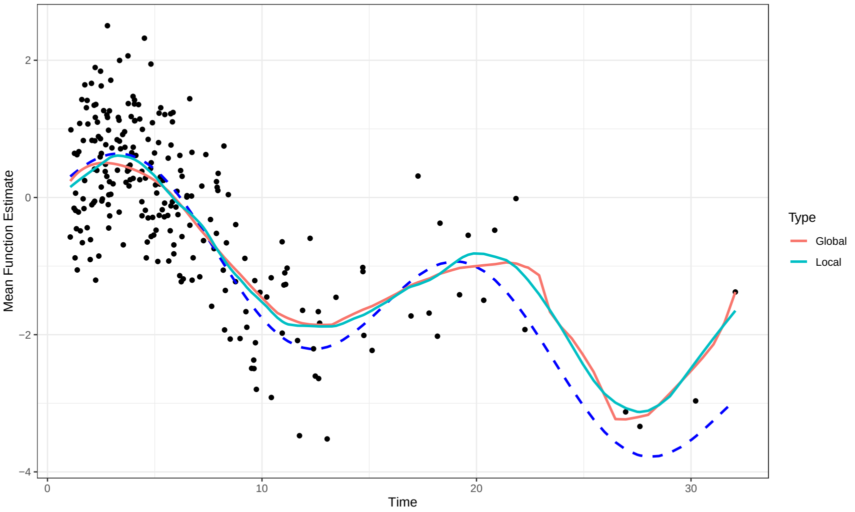
Local linear smoother with **local** bandwidth (Fan & Gijbels (1992))

$$\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left[\alpha(T_{ij})K\left(\frac{T_{ij}-t}{h}\alpha(T_{ij})\right)Y_{ij}-\beta_0-\beta_1(t-T_{ij})\right]^2\to\min$$

optimal $\alpha(\cdot)\sim f^{1/5}(\cdot)$

Kernel Smoothing

Simulation



Kernel Smoothing

Covariance function estimate

Local linear smoother with global bandwidth

$$\sum_{i=1}^{n}\sum_{\substack{j_1=1\\ j_1\neq j_2}}^{N_i}\sum_{j_2=1}^{N_i}\Bigg[K\left(\frac{T_{ij_1}-s}{h},\frac{T_{ij_2}-t}{h}\right)Z(T_{ij_1},T_{ij_2})$$
$$-\beta_0-\beta_{11}(s-T_{ij_1})-\beta_{12}(t-T_{ij_2})\Bigg]^2\to\min$$

- $\hat{c}(s,t)=\hat{\beta}_0(s,t)$
- **goal**: adapt the local bandwidth method

Kernel Smoothing

Covariance function estimate

Local linear smoother with local bandwidth

$$\sum_{i=1}^{n} \sum_{\substack{j_1=1 \\ j_1 \neq j_2}}^{N_i} \sum_{j_2=1}^{N_i} \Big[\alpha(T_{ij_1})\alpha(T_{ij_2}) K\left(\frac{T_{ij_1}-s}{h}\alpha(T_{ij_1}), \frac{T_{ij_2}-t}{h}\alpha(T_{ij_2}) \right) Z(T_{ij_1}, T_{ij_2})$$

$$-\beta_0 - \beta_{11}(s - T_{ij_1}) - \beta_{12}(t - T_{ij_2})\Big]^2 \to \min$$

Known issues:

- symmetry of $\hat{c}(s,t)$ (OK for symmetric kernels)
- positive definiteness of $\hat{c}(s,t)$ (particularly depends on h)
- optimal $\alpha(\cdot)$
- optimal h

Problems to solve

1. Motor Oil Data
   The dataset contains amount of Fe particles depending on operating time and a number of oil changes. Data were collected 2006 – 2016 from 29 heavy-duty army vehicles.
   - Load the variable `df.motor` from the `motoroil.RData` file and plot it (see Figure 1).
   - Use functions from the file `functionsM7777.R` to fill the data (see Figure 2).
   - Try to neglect the number of oil changes and put all groups together (see Figure 3).
   - Fill the data using one of mentioned methods (see Figure 4).
   - Do the same with using the `FPCA` package and compare results (see Figures 5, 6).
   - (optional) Is the number of oil changes negligible? Conduct the fANOVA analysis. Is it correct to do it?

Problems to solve

Motor Oil Data

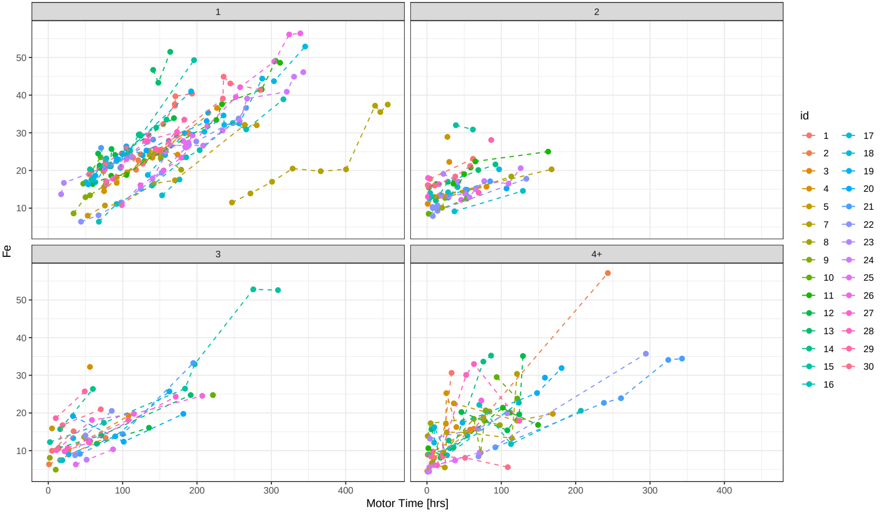


Figure 1.

Motor Oil Data – filled

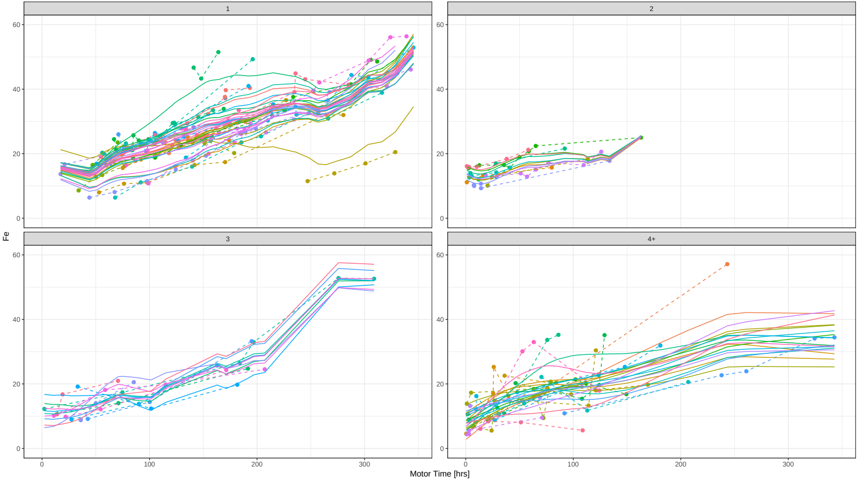


Figure 2.

Problems to solve

Motor Oil Data

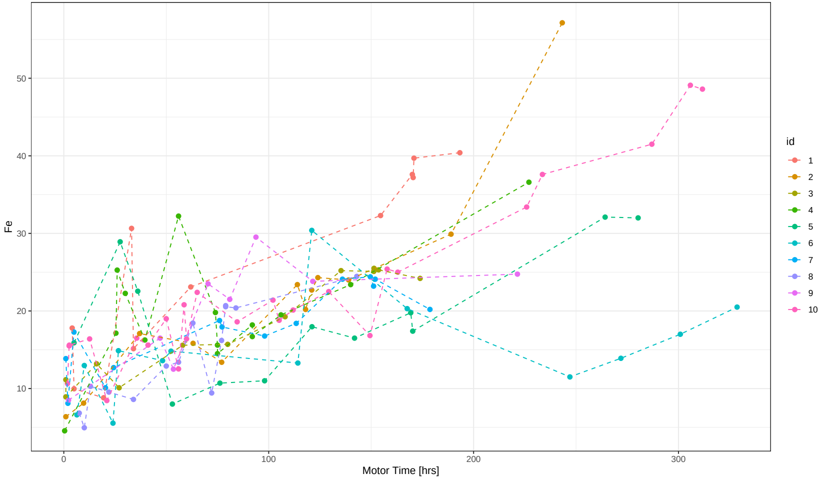


Figure 3.

Motor Oil Data – filled

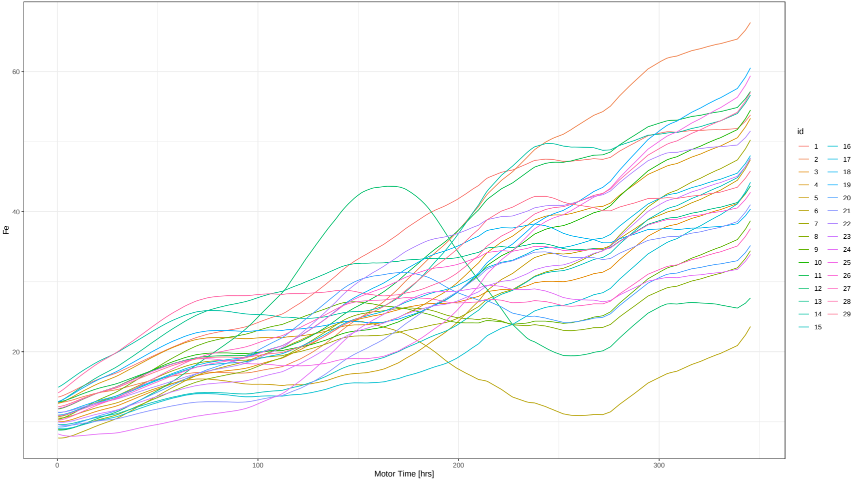


Figure 4.

Problems to solve

functionsM777.R

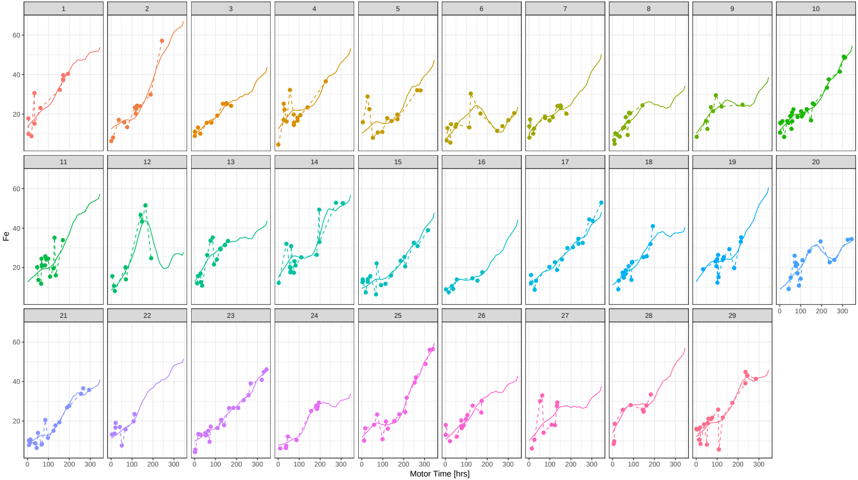


Figure 5.

Problems to solve

FDAPACE

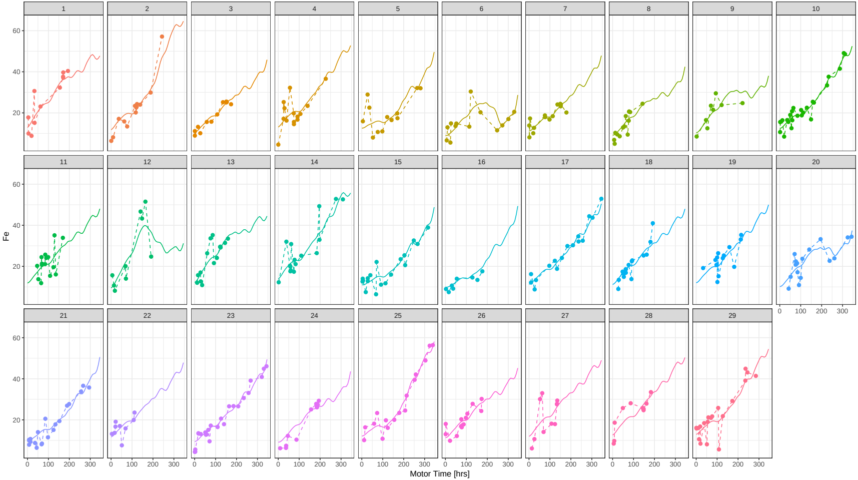


Figure 6.