Exponential Kernel Smoothing

Project Report

Course:
Nonparametric Statistics
Winter Semester 2021/22

Professor:
Vladimir Spokoiny

Teaching Assistant:
Egor Gladin

Author:
Thibaud Joel Hadamczik

Contents

1 Setting

We want to assess the performance of kernel smoothing with an exponential kernel. Therefore we consider a regression model $Y_i = f(X_i) + \varepsilon_i$ for an unknown univariate function on $[0, 1]$ with homogeneous Gaussian noise $\varepsilon \sim N(0, \sigma^2 I_n)$. For our simulations we will consider $\sigma = 1$. We will apply a data-driven bandwidth choice by unbiased risk estimation and leave-one-out cross validation and compare the obtained bandwidths $\hat{h}_{UR}$ and $\hat{h}_{LO}$ to the oracle bandwidth h^*.

We consider the Fourier basis $\{\psi_j(x)\}_{j=1}^{\infty}$ and generate the true function by $f(x) = c_1\psi_1(x)+\cdots+c_m\psi_m(x)$, where the $(c_j)_{1\leq j\leq m}$ are chosen randomly: With γ_j i.i.d. standard normal,

$$c_j = \begin{cases} \gamma_j & ,1 \leq j \leq 10, \\ \gamma_j/(j-10)^2 & ,11 \leq j \leq m. \end{cases}$$

Furthermore we consider deterministic equidistant design points $(x_i)_{1\leq i\leq n}$ on $[0, 1]$. For our setting we will set $n = 100$ and $m = 30$. Kernel regression with an exponential kernel is a linear estimation method and the response estimate for bandwidth h, data $\mathbf{Y}$ and our kernel function $K(x) = \exp(-|x|)$ is obtained by

$$\tilde{f}_h = \mathcal{K}_h \mathbf{Y},$$

where

$$\begin{aligned} \mathcal{K}_h &= (k_{ij,h})_{1\leq i,j\leq n}, \\ k_{ij,h} &= \frac{w_{j,h}(X_i)}{N_h(X_i)}, \\ N_h(X_i) &= \sum_{j=1}^{n} w_{j,h}(X_i), \\ w_{j,h}(X_i) &= K(h^{-1}(X_j - X_i)), \end{aligned}$$

which is equivalent to

$$\tilde{f}_h(X_i) = \frac{\sum_{j=1}^{n} Y_j w_{j,h}(X_i)}{\sum_{j=1}^{n} w_{j,h}(X_i)} = \sum_{j=1}^{n} Y_j k_{ij,h}.$$

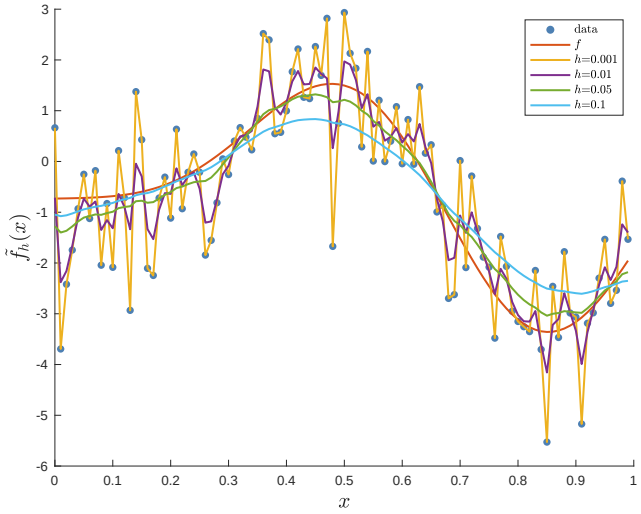


Figure 1: Observed data with obtained response estimate $\tilde{f}_h$ for different bandwidths h.

In Figure 1 the obtained response $\tilde{f}_h$ is illustrated for a sample of data points and different bandwidths h. We can see that for bandwidths of 0.001 and 0.01 we are in the oversmoothing regime, while for $h = 0.1$ we are clearly undersmoothing. We will therefore focus our analysis on the bandwidths $h \in H := \left\{\frac{i}{500}, i = 1, \ldots, 50\right\}$, i.e. consider 50 equidistant bandwidths in the interval $[0.002, 0.1]$.

2 Oracle estimate

To obtain the oracle bandwidth h^*, which minimizes the associated risk,

$$h^* = \text{argmin}_{h \in H} \mathcal{R}_h,$$

where

$$\mathcal{R}_h = \|f - \mathcal{K}_h f\|^2 + \sigma^2 \operatorname{tr}(\mathcal{K}_h \mathcal{K}_h^T),$$

we conduct a Monte Carlo simulation drawing 1000 realizations at the design points and then computing the response estimate $\tilde{f}_h$ for each realization for the different bandwidths.

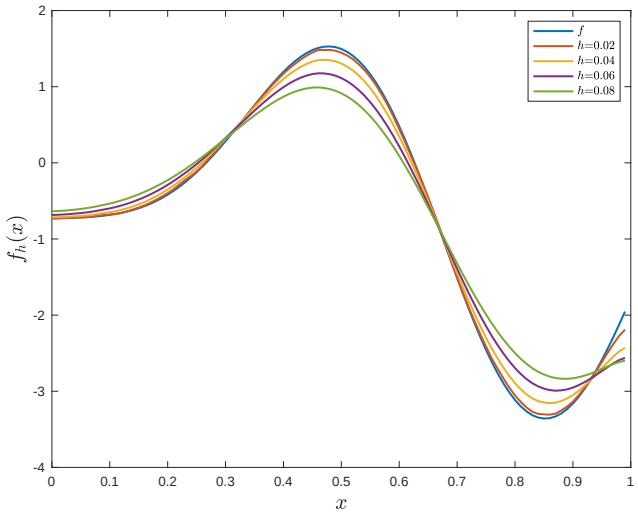


Figure 2: Expected response for different bandwidths and true function f.

In Figure 2 the true function f is plotted against the expected response $f_h = \mathbb{E}\tilde{f}_h$ for a subset of the analyzed bandwidths h and as expected, the smaller h, the closer f_h is to the true function f.

Next, to identify the oracle bandwidth choice h^*, we calculate the squared bias $\|f_h - f\|^2$, the variance $\sigma^2 \operatorname{tr}(\mathcal{K}_h \mathcal{K}_h^T)$ and the risk $\mathcal{R}_h$ for the different bandwidths. The results are shown in Figure 3. The risk minimizing bandwidth is obtained as $h^* = 0.042$ with an associated risk of $\mathcal{R}_{h^*} = 8.66$.

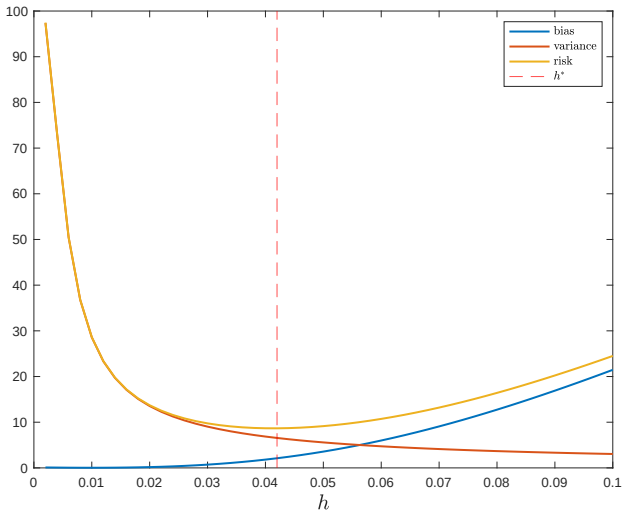


Figure 3: Squared bias, variance and risk in dependence of the bandwidth h.

3 Data-driven choice of the bandwidths

To analyze the data-driven bandwidth selection by unbiased risk estimation and leave-one-out cross validation, we applied these two estimation methods for the Monte Carlo realizations.

For unbiased risk estimation, the data-driven bandwidth was obtained by

$$\hat{h}_{UB} = \text{argmin}_{h \in H} \, \tilde{\mathcal{R}}_h = \text{argmin}_{h \in H} \, \|\mathcal{K}_h \mathbf{Y} - \mathbf{Y}\|^2 + 2\sigma^2 \, \text{tr}(\mathcal{K}_h),$$

where $\tilde{\mathcal{R}}_h$ denotes the unbiased risk estimator, while for leave-one-out cross validation it was obtained by

$$\hat{h}_{LO} = \text{argmin}_{h \in H} \, \|\tilde{f}_h^- - \mathbf{Y}\|^2,$$

where

$$\tilde{f}_h^-(X_i) = \frac{1}{N_h^{-i}} \sum_{j \neq i} Y_j w_{j,h}(X_i), \qquad N_h^{-i} = \sum_{j \neq i} w_{j,h}(X_i)$$

estimates the response for a design point X_i by considering the data for all design points except X_i. In Figure 4 the obtained response estimates by unbiased risk estimation $\tilde{f}_{\hat{h}_{UR}}$, leave-one-out cross validation $\tilde{f}_{\hat{h}_{LU}}$ and the oracle estimate $\tilde{f}_{h^*}$ are plotted against the true function for one exemplary data sample.

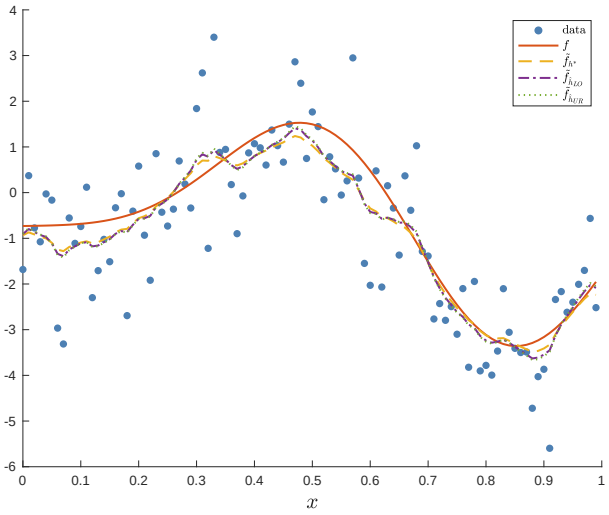
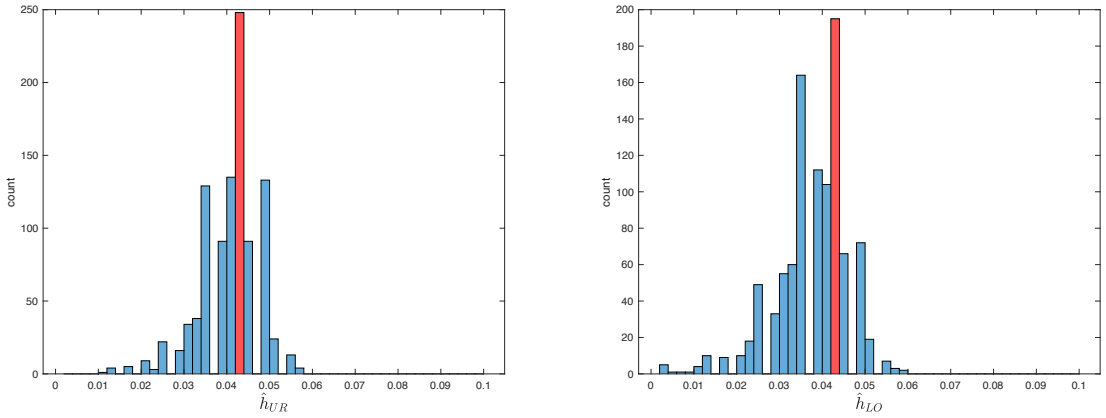


Figure 4: Oracle estimator, estimates obtained by unbiased risk estimation and leave-one-out cross validation and true function.

The histograms of the selected bandwidths by unbiased risk estimation $\hat{h}_{UR}$ and leave one out cross validation $\hat{h}_{LO}$ are shown in Figure 5 with the red bar indicating the oracle bandwidth h^*.

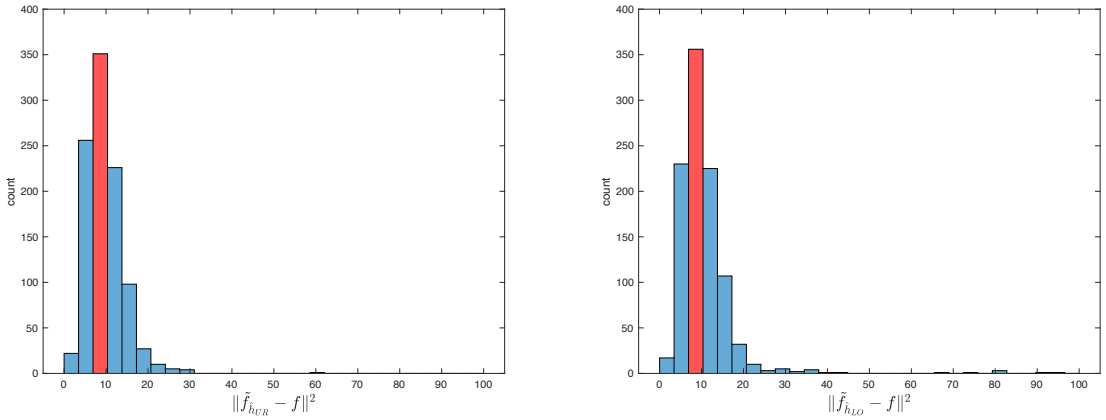


(a) Selected bandwidths by unbiased risk estimation.

(b) Selected bandwidths by leave-one-out cross validation.

Figure 5: Histograms of the obtained bandwidths for unbiased risk estimation and leave-one-out cross validation.

We also computed the resulting loss for the two methods for each Monte Carlo realization. The histograms of the squared losses for both methods are shown in Figure 6 with the red bar indicating the bin which contains the risk of the oracle estimator f_{h^*}. It is remarkable that both methods do a solid job at approximating the true risk.



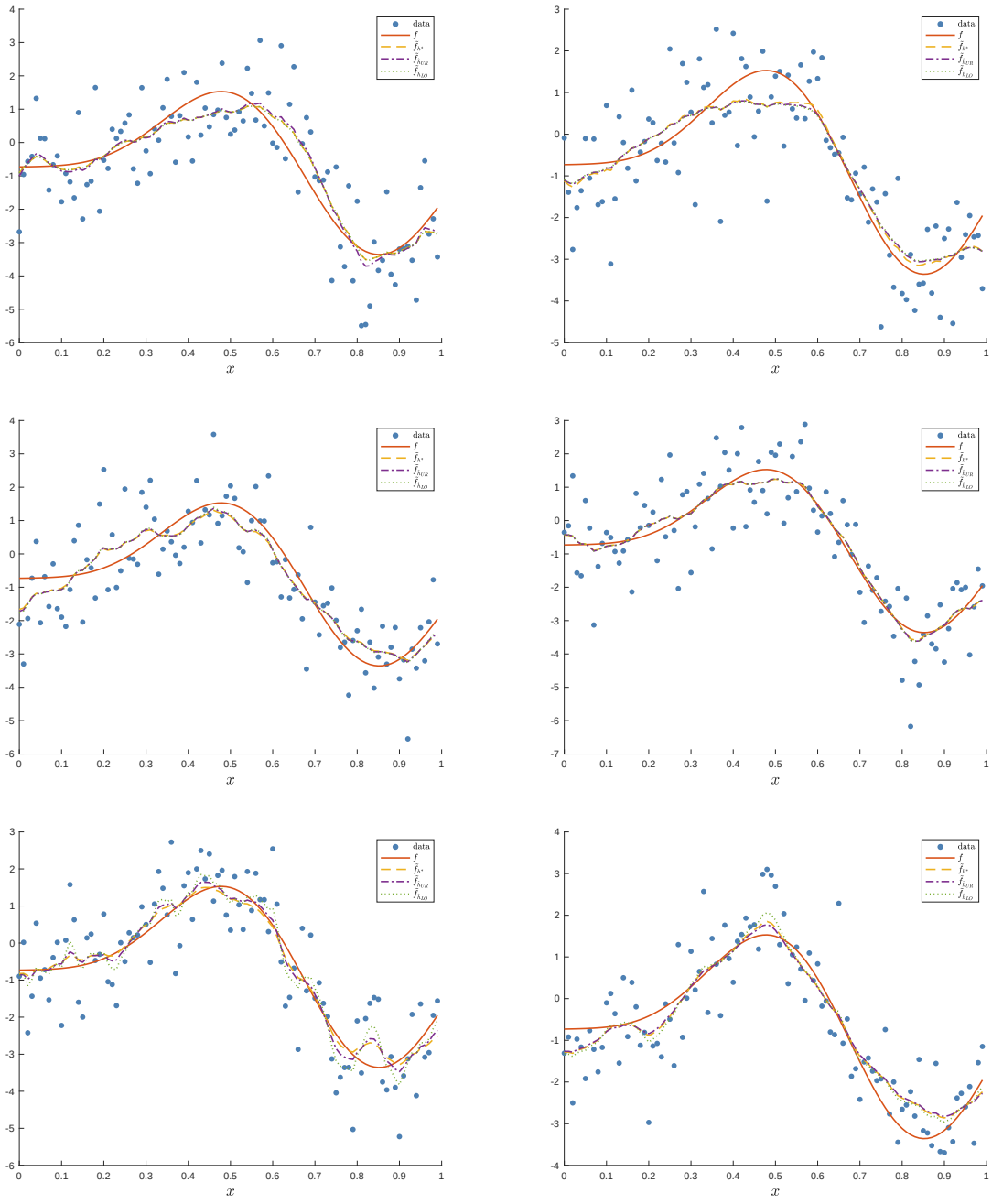
(a) Unbiased risk estimation.

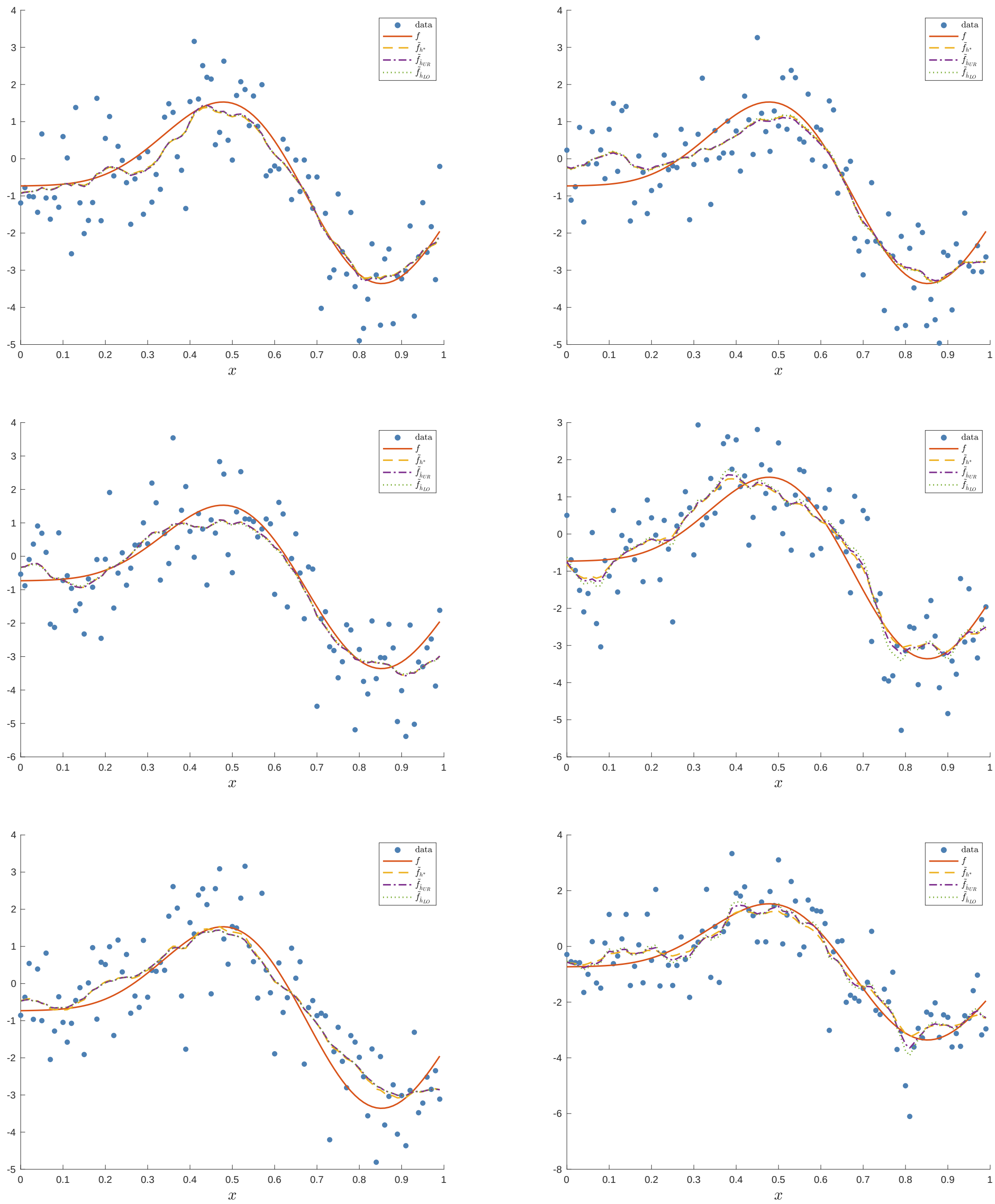
(b) Leave-one-out cross validation

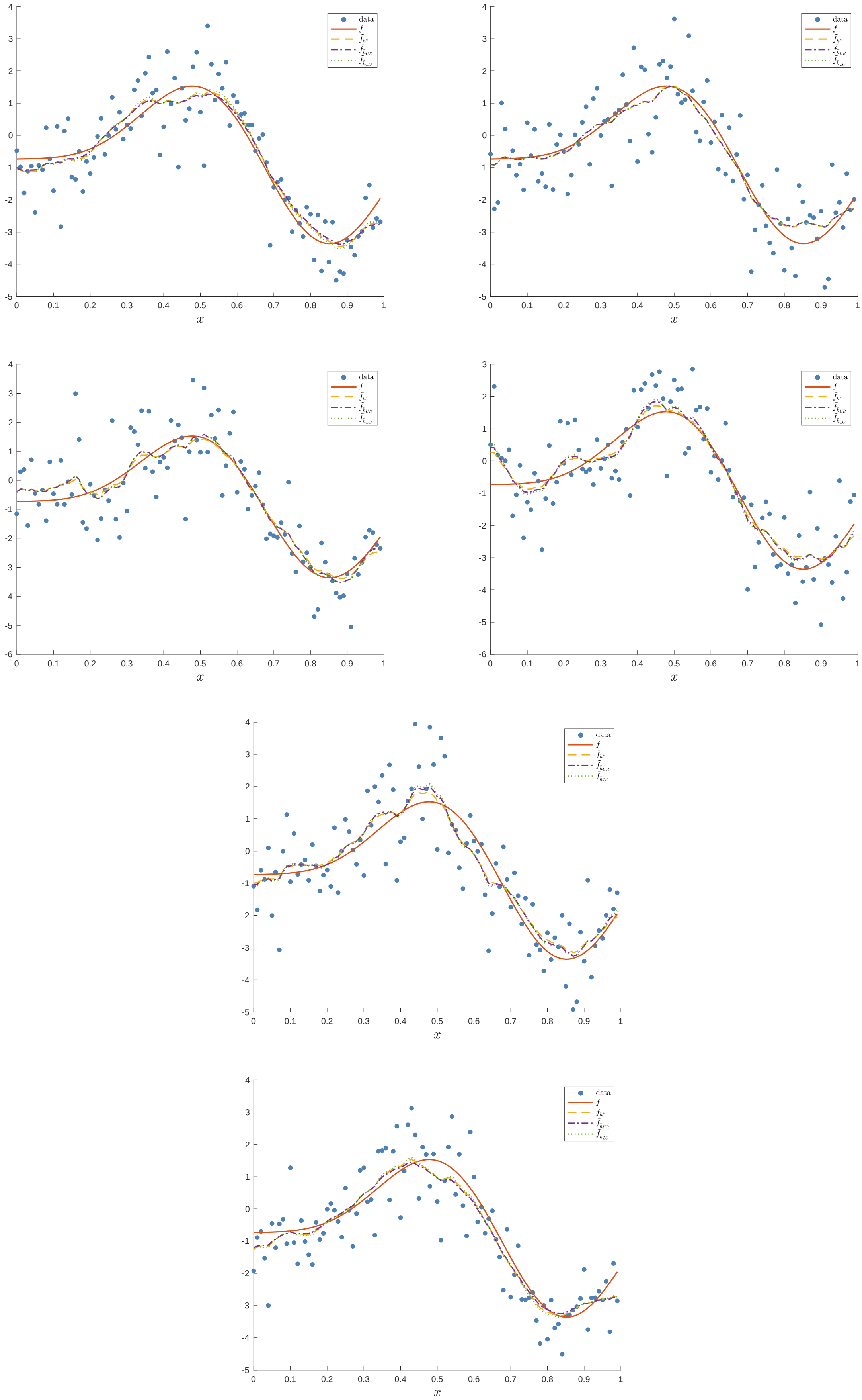
Figure 6: Histograms of the squared losses for unbiased risk estimation and leave-one-out cross validation.

To illustrate the variability of the different estimators and how they depend on the concrete data, we considered a subset of 18 Monte Carlo realizations and plotted the oracle estimate $\tilde{f}_{h^*}$ and the estimates obtained by unbiased risk estimation $\tilde{f}_{\hat{h}_{UR}}$ and leave-one-out cross validation $\tilde{f}_{\hat{h}_{LO}}$ for these data samples. The results are shown in Figure 7.

Figure 7: 18 Monte Carlo realizations of the data and the response estimates by unbiased risk estimation $\tilde{f}_{\hat{h}_{UR}}$ and leave-one-out cross validation $\tilde{f}_{\hat{h}_{LO}}$ plotted against the true function f and the oracle estimate $\tilde{f}_{h^*}$.




data
f
$\hat{f}_{h^*}$
$\hat{f}_{h_{UR}}$
$\hat{f}_{h_{LO}}$
x


data
f
$\hat{f}_{h^*}$
$\hat{f}_{h_{UR}}$
$\hat{f}_{h_{LO}}$
x