Weierstrass Institute for
Applied Analysis and Stochastics

Leibniz Association

Bootstrap tuning in model choice problem

Vladimir Spokoiny,
(with Niklas Willrich)
WIAS, HU Berlin, MIPT, IITP Moscow

SFB 649 Motzen, 17.07.2015

Outline

Consider a linear model

$$\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon} \in I\!R^n$$

for an unknown parameter vector $\boldsymbol{\theta}^* \in I\!R^p$ and a given $p \times n$ design matrix Ψ.

Suppose that a family of linear smoothers

$$\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$$

is given, where $\mathcal{S}_m$ is for each $m \in \mathcal{M}$ a given $p \times n$ matrix.

We also assume that this family is ordered by the complexity of the method.

The task is to develop a data based model selector $\widehat{m}$ which performs nearly as good as the optimal choice which depends on the model and is not available.

We consider the following linear Gaussian model:

$$Y_i = \Psi_i^\top \boldsymbol{\theta}^* + \varepsilon_i\,, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}\,, \qquad i = 1, \ldots, n. \tag{1}$$

We also write this equation in the vector form

$$\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon} \in I\!R^n$$

where Ψ is $p \times n$ deterministic design matrix and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$.
In what follows, we allow the model (1) to be completely misspecified:

- **True model**: Y_i are independent, the response $\boldsymbol{f}^* = I\!E\boldsymbol{Y} \in I\!R^n$ with entries f_i:

$$Y_i = f_i + \varepsilon_i\,. \tag{2}$$

- The linear parametric assumption $\boldsymbol{f}^* = \Psi^\top \boldsymbol{\theta}^*$ can be violated;
- The noise $\boldsymbol{\varepsilon} = (\varepsilon_i)$ can be heterogeneous and non-Gaussian.

For the linear model (2), define $\boldsymbol{\theta}^* \in I\!R^p$ as the vector providing the best linear fit:

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{\theta}} I\!E\|\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}\|^2 = \left(\Psi\Psi^\top\right)^{-1}\Psi \boldsymbol{f}^*.$$

Below we assume a family $\{\widetilde{\boldsymbol{\theta}}_m\}$ of linear estimators of $\boldsymbol{\theta}^*$ to be given:

$$\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$$

Typical examples include

- **projection estimation** on a m-dimensional subspace;
- **regularized estimation** with a regularization parameter α_m ;
- **penalized estimators** with a quadratic penalty function;
- **kernel estimation** with a bandwidth h_m .

Introduce a weighting $q \times p$-matrix W for some fixed $q \geq 1$ and define quadratic loss and risk with this weighting matrix W:

$$\varrho_m \stackrel{\text{def}}{=} \|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2, \qquad \mathcal{R}_m \stackrel{\text{def}}{=} I\!\!E\|W(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2.$$

Typical examples of W are as follows:

- **Estimation of the whole vector $\boldsymbol{\theta}^*$:**
  Let W be the identity matrix $W = I_p$ with $q = p$. This means that the estimation loss is measured by the usual squared Euclidean norm $\|\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\|^2$.
- **Prediction:** Let W be the square root of the total Fisher information matrix, that is, $W^2 = \mathbb{F} = \sigma^{-2}\Psi\Psi^\top$. Usually referred to as *prediction loss*.
- **Semiparametric estimation:** Let the target of estimation is not the whole vector $\boldsymbol{\theta}^*$ but its subvector $\boldsymbol{\theta}_0^*$ of dimension q. The matrix W can be defined as the projector Π_0 on the $\boldsymbol{\theta}_0^*$ subspace.
- **Linear functional estimation:** The choice of the weighting matrix W can be adjusted to the problem of estimating some functionals of the whole parameter $\boldsymbol{\theta}^*$.

In all cases, the most important feature of the estimators $\widetilde{\boldsymbol{\theta}}_m$ is *linearity*.

In all cases, the most important feature of the estimators $\widetilde{\boldsymbol{\theta}}_m$ is *linearity*. It greatly simplifies the study of their properties including the prominent bias-variance decomposition of the risk of $\widetilde{\boldsymbol{\theta}}_m$. Namely, for the model (2) with $I\!\!E\boldsymbol{\varepsilon} = 0$, it holds

$$\begin{aligned} I\!\!E\widetilde{\boldsymbol{\theta}}_m &= \boldsymbol{\theta}^*_m = \mathcal{S}_m \boldsymbol{f}^*, \\ \mathcal{R}_m &= \|W\left(\boldsymbol{\theta}^*_m - \boldsymbol{\theta}^*\right)\|^2 + \mathrm{tr}\left\{W\mathcal{S}_m \ \mathrm{Var}(\boldsymbol{\varepsilon})\, \mathcal{S}_m^\top W^\top\right\} \\ &= \|W(\mathcal{S}_m - \mathcal{S})\boldsymbol{f}^*\|^2 + \mathrm{tr}\left\{W\mathcal{S}_m \ \mathrm{Var}(\boldsymbol{\varepsilon})\, \mathcal{S}_m^\top W^\top\right\}. \end{aligned} \tag{3}$$

The optimal choice of the parameter m can be defined by risk minimization:

$$m^* \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_{m \in \mathcal{M}} \mathcal{R}_m .$$

The *model selection* problem can be described as the choice of m by data which *mimics the oracle*, that is, we aim at constructing a selector $\widehat{m}$ leading to the adaptive estimate $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$ with the properties similar to the oracle estimate $\widetilde{\boldsymbol{\theta}}_{m^*}$.

- **unbiased risk estimation** [Kneip, 1994];
- **penalized model selection** [Barron et al., 1999], [Massart, 2007]);
- **Lepski's method** [Lepski, 1990], [Lepski, 1991], [Lepski, 1992], [Lepski and Spokoiny, 1997], [Lepski et al., 1997], [Birgé, 2001]
- **risk hull minimization** [Cavalier and Golubev, 2006].
- **Resampling for noise estimation:** For the penalized model selection, [Arlot, 2009] suggested the use of resampling methods for the choice of an optimal penalization, [Arlot and Bach, 2009] used the concept of minimal penalties from [Birgé and Massart, 2007].
- **Aggregation:**
  An alternative approach to adaptive estimation is based on aggregation of different estimates; see [Goldenshluger, 2009] and [Dalalyan and Salmon, 2012];
- **Linear inverse problems:** [Tsybakov, 2000], [Cavalier et al., 2002].

Validity of a bootstrapping procedure for Lepski's method has been studied in [Chernozhukov et al., 2014] with applications to honest adaptive confidence bands.

[Spokoiny and Vial, 2009] offered a propagation approach to calibration of the Lepski's method in the case of the estimation of a one-dimensional quantity of interest.

A similar approach has been applied to local constant density estimation with sup-norm risk in [Gach et al., 2013] and to local quantile estimation in [Spokoiny et al., 2013].

Below we discuss the *ordered case*. The parameter $m \in \mathcal{M}$ is treated as complexity of the method $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$. In some cases the set $\mathcal{M}$ of possible m choices can be countable and/or continuous and even unbounded. For simplicity of presentation, we assume that $\mathcal{M}$ is a finite set of positive numbers, $|\mathcal{M}|$ stands for its cardinality.
Typical examples:

- **the number of terms** in the Fourier expansion;
- **bandwidth** in the kernel smoothing.
- **regularization parameter**;
- **penalty coefficient**.

In general, complexity can be naturally expressed via the variance of the stochastic term of the estimator $\widetilde{\boldsymbol{\theta}}_m$: the larger m, the larger is the variance $\mathrm{Var}(W\widetilde{\boldsymbol{\theta}}_m)$.

In the case of projection estimation with m-dimensional projectors, this variance is linear in m, $\mathrm{Var}(\widetilde{\boldsymbol{\theta}}_m) = \sigma^2 m$.

In general, dependence of the variance term on m may be more complicated but the monotonicity of $\mathrm{Var}(W\widetilde{\boldsymbol{\theta}}_m)$ has to be preserved.

Outline

Consider a family $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$ and $\widetilde{\boldsymbol{\phi}}_m = \mathcal{K}_m \boldsymbol{Y}$ with $\mathcal{K}_m = W\mathcal{S}_m \colon \mathbb{R}^n \to \mathbb{R}^q$, $m \in \mathcal{M}$, of linear estimators of the q-dimensional target parameter $\boldsymbol{\phi}^* = W\boldsymbol{\theta}^* = W\mathcal{S}\boldsymbol{f}^* = \mathcal{K}\boldsymbol{f}^*$ for $\mathcal{K} = W\mathcal{S}$.
Suppose that $\{\widetilde{\boldsymbol{\phi}}_m, m \in \mathcal{M}\}$ is ordered by their complexity (variance):

$$\mathcal{K}_m \operatorname{Var}(\boldsymbol{\varepsilon}) \mathcal{K}_m^\top \leq \mathcal{K}_{m'} \operatorname{Var}(\boldsymbol{\varepsilon}) \mathcal{K}_{m'}^\top, \qquad m' > m.$$

One would like to pick up a smallest possible index $m \in \mathcal{M}$ which still provides a reasonable fit. The latter means that the bias component

$$\|b_m\|^2 = \|\boldsymbol{\phi}_m^* - \boldsymbol{\phi}^*\|^2 = \|(\mathcal{K}_m - \mathcal{K})\boldsymbol{f}^*\|^2$$

in the risk decomposition (3) is not significantly larger than the variance

$$\operatorname{tr}\{\operatorname{Var}(\widetilde{\boldsymbol{\phi}}_m)\} = \operatorname{tr}\{\mathcal{K}_m \operatorname{Var}(\boldsymbol{\varepsilon}) \mathcal{K}_m^\top\}.$$

If $m^\circ \in \mathcal{M}$ is such a "good" choice, then our ordering assumption yields that a further increase of the index m over m° only increases the complexity (variance) of the method without real gain in the quality of approximation.

This latter fact can be interpreted in term of pairwise comparison: whatever $m \in \mathcal{M}$ with $m > m^\circ$ we take, there is no significant bias reduction in using a larger model m instead of m°.

Leads to a multiple test procedure: for each pair $m > m^\circ$ from $\mathcal{M}$, we consider a hypothesis of no significant bias between the models m° and m, and let τ_{m,m° be the corresponding test.

The model m° is accepted if $\tau_{m,m^\circ} = 0$ for all $m > m^\circ$. Finally, the selected model is the "smallest accepted":

$$\widehat{m} \stackrel{\mathrm{def}}{=} \operatorname{argmin}\bigl\{m^\circ \in \mathcal{M} \colon \tau_{m,m^\circ} = 0,\, \forall m > m^\circ\bigr\}.$$

Usually the test τ_{m,m° can be written in the form

$$\tau_{m,m^\circ} = \mathbb{1}\bigl\{\mathbb{T}_{m,m^\circ} > \mathfrak{z}_{m,m^\circ}\bigr\}$$

for some *test statistics* $\mathbb{T}_{m,m^\circ}$ and for *critical values* $\mathfrak{z}_{m,m^\circ}$.

The information-based criteria like AIC or BIC use the likelihood ratio test statistics $\mathbb{T}_{m,m^\circ} = \sigma^{-2} \bigl\| \Psi^\top \bigl(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ} \bigr) \bigr\|^2$. A great advantage of such tests is that the test statistic $\mathbb{T}_{m,m^\circ}$ is pivotal (χ^2 with $m - m^\circ$ degrees of freedom) under the correct null hypothesis, this makes simple to compute the corresponding critical values.

Below we apply another choice based on the norm of differences $\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}$:

$$\mathbb{T}_{m,m^\circ} = \| \widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ} \| = \| \mathcal{K}_{m,m^\circ} \boldsymbol{Y} \|, \qquad \mathcal{K}_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_m - \mathcal{K}_{m^\circ} .$$

The main issue for such a method is a proper choice of the critical values $\mathfrak{z}_{m,m^\circ}$. One can say that the procedure is specified by a way of selecting these critical values.

Below we offer a novel way of doing this choice in a general situation by using a so called *propagation condition*: if a model m° is "good" it has to be accepted with a high probability.

This rule can be seen as analog of the family-wise level condition in a multiple test problem. Rejecting a "good" model is the family-wise error of first kind, and this error has to be controlled.

To specify precisely the meaning of a good model, consider for $m > m^\circ$ the decomposition

$$\mathbb{T}_{m,m^\circ} = \|\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}\| = \|\mathcal{K}_{m,m^\circ}\boldsymbol{Y}\| = \|\mathcal{K}_{m,m^\circ}(\boldsymbol{f}^* + \boldsymbol{\varepsilon})\| = \|b_{m,m^\circ} + \boldsymbol{\xi}_{m,m^\circ}\|,$$

where with $\mathcal{K}_{m,m^\circ} = \mathcal{K}_m - \mathcal{K}_{m^\circ}$

$$b_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_{m,m^\circ}\boldsymbol{f}^*, \qquad \boldsymbol{\xi}_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_{m,m^\circ}\boldsymbol{\varepsilon}.$$

It obviously holds $I\!\!E\boldsymbol{\xi}_{m,m^\circ} = 0$. Introduce $q \times q$-matrix $\mathbb{V}_{m,m^\circ}$ as the variance of $\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}$:

$$\mathbb{V}_{m,m^\circ} \stackrel{\text{def}}{=} \operatorname{Var}\bigl(\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}\bigr) = \operatorname{Var}\bigl(\mathcal{K}_{m,m^\circ}\boldsymbol{Y}\bigr) = \mathcal{K}_{m,m^\circ}\operatorname{Var}(\boldsymbol{\varepsilon})\,\mathcal{K}_{m,m^\circ}^\top.$$

Further,

$$I\!\!E\,\mathbb{T}_{m,m^\circ}^2 = \|b_{m,m^\circ}\|^2 + I\!\!E\|\boldsymbol{\xi}_{m,m^\circ}\|^2 = \|b_{m,m^\circ}\|^2 + \mathtt{p}_{m,m^\circ},$$

$$\mathtt{p}_{m,m^\circ} = \operatorname{tr}(\mathbb{V}_{m,m^\circ}) = I\!\!E\|\boldsymbol{\xi}_{m,m^\circ}\|^2.$$

The bias term $b_{m,m^\circ} \stackrel{\mathrm{def}}{=} \mathcal{K}_{m,m^\circ} \boldsymbol{f}^*$ is significant if its squared norm is competitive with the variance term $\mathtt{p}_{m,m^\circ} = \mathrm{tr}(\mathbb{V}_{m,m^\circ})$.

We say that m° is a "good" choice if there is no significant bias b_{m,m° for any $m > m^\circ$. This condition can be quantified as "bias-variance trade-off":

$$\|b_{m,m^\circ}\|^2 \leq \beta^2 \mathtt{p}_{m,m^\circ}, \qquad m > m^\circ \tag{4}$$

for a given parameter β.

Define the *oracle* m^* as the minimal m° with the property (4):

$$m^* \stackrel{\mathrm{def}}{=} \min\Big\{ m^\circ : \max_{m > m^\circ} \big\{ \|b_{m,m^\circ}\|^2 - \beta^2 \mathtt{p}_{m,m^\circ} \big\} \leq 0 \Big\}.$$

Let the noise distribution be known. A particular example is the case of Gaussian errors $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Then the distribution of the stochastic component $\boldsymbol{\xi}_{m,m^\circ}$ is known as well. Introduce for each pair $m > m^\circ$ from $\mathcal{M}$ a *tail function* $z_{m,m^\circ}(t)$ of the argument t such that

$$\mathbb{P}\Big(\|\boldsymbol{\xi}_{m,m^\circ}\| > z_{m,m^\circ}(t) \Big) = \mathrm{e}^{-t}. \tag{5}$$

Here we assume that the distribution of $\|\boldsymbol{\xi}_{m,m^\circ}\|$ is continuous and the value $z_{m,m^\circ}(t)$ is well defined. Otherwise one has to define $z_{m,m^\circ}(t)$ as a smallest value providing the prescribing error probability e^{-t}.

For checking the propagation condition, we need a uniform in $m > m^\circ$ version of the probability bound (5). Let

$$\mathcal{M}^+(m^\circ) \stackrel{\mathrm{def}}{=} \{m \in \mathcal{M} \colon m > m^\circ\}.$$

Given $\mathtt{x}$, by $q_{m^\circ} = q_{m^\circ}(\mathtt{x})$ denote the corresponding multiplicity correction:

$$\mathbb{P}\Big(\bigcup_{m \in \mathcal{M}^+(m^\circ)} \{\|\boldsymbol{\xi}_{m,m^\circ}\| \geq z_{m,m^\circ}(\mathtt{x} + q_{m^\circ})\} \Big) = \mathrm{e}^{-\mathtt{x}}.$$

A simple way of computing the multiplicity correction q_{m° is based on the Bonferroni bound: $q_{m^\circ} = \log(\#\mathcal{M}^+(m^\circ))$.

However, it is well known that the Bonferroni bound is very conservative and leads to very large correction q_{m°, especially if the random vectors $\boldsymbol{\xi}_{m,m^\circ}$ are strongly correlated.

As the joint distribution of the $\boldsymbol{\xi}_{m,m^\circ}$'s is precisely known, define the correction $q_{m^\circ} = q_{m^\circ}(\mathtt{x})$ just by condition

$$\mathbb{P}\Big(\bigcup_{m \in \mathcal{M}^+(m^\circ)} \big\{ \|\boldsymbol{\xi}_{m,m^\circ}\| \geq z_{m,m^\circ}(\mathtt{x} + q_{m^\circ}) \big\} \Big) = \mathrm{e}^{-\mathtt{x}}.$$

Finally we define the critical values $\mathfrak{z}_{m,m^\circ}$ by one more correction for the bias:

$$\mathfrak{z}_{m,m^\circ} \stackrel{\mathrm{def}}{=} z_{m,m^\circ}(\mathtt{x} + q_{m^\circ}) + \beta\sqrt{\mathtt{p}_{m,m^\circ}} \tag{6}$$

for $\mathtt{p}_{m,m^\circ} = \mathrm{tr}(\mathbb{V}_{m,m^\circ})$.

In practice $\mathtt{x} = 3$ and $\beta = 0$ provide a reasonable choice.

Define the selector $\widehat{m}$ by the "smallest accepted" (SmA) rule. Namely, with $\mathfrak{z}_{m,m^\circ}$ from (6), the acceptance rule reads as follows:

$$\bigl\{m^\circ \text{ is accepted}\bigr\} \Leftrightarrow \Bigl\{ \max_{m \in \mathcal{M}^+(m^\circ)} \bigl\{ \mathbb{T}_{m,m^\circ} - \mathfrak{z}_{m,m^\circ} \bigr\} \le 0 \Bigr\}.$$

The SmA rule is

$$\begin{aligned} \widehat{m} &\stackrel{\text{def}}{=} \text{"smallest accepted"} \\ &= \min\Bigl\{ m^\circ : \max_{m \in \mathcal{M}^+(m^\circ)} \bigl\{ \mathbb{T}_{m,m^\circ} - \mathfrak{z}_{m,m^\circ} \bigr\} \le 0 \Bigr\}. \end{aligned}$$

Our study mainly focuses on the behavior of the selector $\widehat{m}$. The performance of the resulting estimator $\widehat{\boldsymbol{\phi}} = \widetilde{\boldsymbol{\phi}}_{\widehat{m}}$ is a kind of corollary from statements about the selected model $\widehat{m}$. The ideal solution would be $\widehat{m} \equiv m^*$, then the adaptive estimator $\widehat{\boldsymbol{\phi}}$ coincides with the oracle estimate $\widetilde{\boldsymbol{\phi}}_{m^*}$.

The decomposition

$$\mathbb{T}_{m,m^\circ} = \|\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}\| = \|b_{m,m^\circ} + \boldsymbol{\xi}_{m,m^\circ}\| \le \|b_{m,m^\circ}\| + \|\boldsymbol{\xi}_{m,m^\circ}\|,$$

and the bounds

$$\mathbb{P}\Big(\bigcup_{m \in \mathcal{M}^+(m^\circ)} \big\{ \|\boldsymbol{\xi}_{m,m^\circ}\| \ge z_{m,m^\circ}(\mathtt{x} + q_{m^\circ}) \big\} \Big) = \mathrm{e}^{-\mathtt{x}},$$

$$\|b_{m,m^\circ}\|^2 \le \beta^2 \, \mathtt{p}_{m,m^\circ}, \qquad m > m^\circ$$

automatically ensures the desired *propagation property*.

Theorem. *Any good model* m° *will be accepted with probability at least* $1 - \mathrm{e}^{-\mathtt{x}}$ *:*

$$\mathbb{P}\big(m^* \text{ is rejected}\big) \le \mathrm{e}^{-\mathtt{x}}.$$

Corollary. By definition, the oracle m^* is also a “good” choice, thus accepted.

The oracle m^* is also a "good" choice, this yields

$$\mathbb{P}\big(m^* \text{ is rejected}\big) \le \mathrm{e}^{-\mathtt{x}}.$$

Therefore, the selector $\widehat{m}$ typically takes its value in $\mathcal{M}^-(m^*)$, where

$$\mathcal{M}^-(m^*) = \big\{m \in \mathcal{M}\colon m < m^*\big\}$$

is the set of all models in $\mathcal{M}$ smaller than m^*.

Zone of insensitivity: a subset $\mathcal{M}^\circ$ of $\mathcal{M}^-(m^*)$ of possible $\widehat{m}$-values.

The definition of m^* implies that there is a significant bias for each $m \in \mathcal{M}^-(m^*)$.

Intuition: Zone of insensitivity is composed of m-values for which the bias is significant but not very large.

Theorem. *For any subset* $\mathcal{M}^c \subseteq \mathcal{M}^-(m^*)$ *s.t.*

$$\|b_{m^*,m}\| > \mathfrak{z}_{m^*,m} + z_{m^*,m}(\mathtt{x}_\mathtt{s}), \qquad m \in \mathcal{M}^c,$$

for $\mathtt{x}_\mathtt{s} \stackrel{\mathrm{def}}{=} \mathtt{x} + \log(|\mathcal{M}^c|)$ *with* $|\mathcal{M}^c|$ *being the cardinality of* $\mathcal{M}^c$ *, it holds*

$$I\!P\big(\widehat{m} \in \mathcal{M}^c\big) \leq \mathrm{e}^{-\mathtt{x}}.$$

The SmA estimator $\widehat{\phi} = \widetilde{\phi}_{\widehat{m}}$ *satisfies the following bound:*

$$I\!P\Big(\big\|\widehat{\phi} - \widetilde{\phi}_{m^*}\big\| > \bar{\mathfrak{z}}_{m^*}\Big) \leq 2\mathrm{e}^{-\mathtt{x}},$$

where $\bar{\mathfrak{z}}_{m^*}$ *is defined with* $\mathcal{M}^\circ \stackrel{\mathrm{def}}{=} \mathcal{M}^-(m^*) \setminus \mathcal{M}^c$ *as*

$$\bar{\mathfrak{z}}_{m^*} \stackrel{\mathrm{def}}{=} \max_{m \in \mathcal{M}^\circ} \mathfrak{z}_{m^*,m}\,.$$

Outline

Consider the bootstrap estimates $\widetilde{\boldsymbol{\phi}}_m = W\widetilde{\boldsymbol{\theta}}^{\flat}_m$ in the form

$$\widetilde{\boldsymbol{\phi}}^{\flat}_m = W\bigl(\Psi_m\Psi_m^{\top}\bigr)^{-1}\Psi_m\mathcal{W}^{\flat}\boldsymbol{Y} = \mathcal{K}_m\mathcal{W}^{\flat}\boldsymbol{Y}.$$

Here $\mathcal{W}^{\flat}\boldsymbol{Y}$ means the vector with entries $w^{\flat}_i Y_i$ for $i \le n$, where $w^{\flat}_i$ are i.i.d. bootstrap weights with $I\!\!E^{\flat} w^{\flat}_i = \mathrm{Var}(w^{\flat}_i) = 1$. We are interested if the distribution of the differences

$$\widetilde{\boldsymbol{\phi}}^{\flat}_m - \widetilde{\boldsymbol{\phi}}^{\flat}_{m^\circ} = \mathcal{K}_{m,m^\circ}\mathcal{W}^{\flat}\boldsymbol{Y}, \quad , m > m^\circ$$

mimics their real world counterparts.

The identity $I\!\!E^{\flat}\mathcal{W}^{\flat} = I_n$ yields $I\!\!E^{\flat}\widetilde{\boldsymbol{\phi}}^{\flat}_{m,m^\circ} = \widetilde{\boldsymbol{\phi}}_{m,m^\circ}$, and the natural idea would be to use the difference

$$\widetilde{\boldsymbol{\phi}}^{\flat}_{m,m^\circ} - \widetilde{\boldsymbol{\phi}}_{m,m^\circ} = \mathcal{K}_{m,m^\circ}\bigl(\mathcal{W}^{\flat} - I_n\bigr)\boldsymbol{Y}$$

as a proxy for the stochastic component $\boldsymbol{\xi}_{m,m^\circ} = \mathcal{K}_{m,m^\circ}\boldsymbol{\varepsilon}$.

Unfortunately, this can only be justified if the bias component b_{m,m° of $\widetilde{\boldsymbol{\phi}}_{m,m^\circ}$ is small relative to its stochastic variance $\mathtt{p}_{m,m^\circ}$. But this is exactly what we would like to test!

To avoid this problem we apply a presmoothing which removes a pilot prediction of the regression function from the data. This presmoothing requires some minimal smoothness of the regression function, and this condition seems to be unavoidable if no information about noise is given: otherwise one cannot separate between signal and noise.

Suppose that a linear predictor $\widetilde{\boldsymbol{f}}_0 = \Pi \boldsymbol{Y}$ is given where Π is a sub-projector in the space $I\!R^n$. In most of cases one can take $\Pi = \Psi_{m^\dagger}^\top \left(\Psi_{m^\dagger} \Psi_{m^\dagger}^\top \right)^{-1} \Psi_{m^\dagger}$ where $m^\dagger$ is a large index, e.g. the largest index M in our collection.

Idea: computes the residuals $\breve{\boldsymbol{Y}} = \boldsymbol{Y} - \Pi \boldsymbol{Y}$ and uses them in place of the original data. This allows to remove the bias while keeping the noise variance only slightly changed.

For each bootstrap realization $\boldsymbol{w}^\flat = (w_i^\flat)$, we apply the procedure to the data vector $\mathcal{W}^\flat \breve{\boldsymbol{Y}}$ with entries $\breve{Y}_i w_i^\flat$ for $i \le n$. The bootstrap stochastic components $\boldsymbol{\xi}_{m,m^\circ}^\flat$ are defined as

$$\boldsymbol{\xi}_{m,m^\circ}^\flat \stackrel{\mathrm{def}}{=} \mathcal{K}_{m,m^\circ} \mathcal{E}^\flat \breve{\boldsymbol{Y}}, \qquad m > m^\circ,$$

where $\mathcal{E}^\flat = \mathcal{W}^\flat - I_n$ is the diagonal matrix of bootstrap errors $\varepsilon_i^\flat = w_i^\flat - 1 \sim \mathcal{N}(0,1)$.

The bootstrap quantiles $z^{\flat}_{m,m^\circ}(t)$ are given by the analog of (5):

$$I\!P^{\flat}\Big(\|\boldsymbol{\xi}^{\flat}_{m,m^\circ}\| > z^{\flat}_{m,m^\circ}(t)\Big) = \mathrm{e}^{-t}.$$

The multiplicity correction $q^{\flat}_{m^\circ} = q^{\flat}_{m^\circ}(\mathtt{x})$ is specified by the condition

$$I\!P^{\flat}\left(\bigcup_{m\in\mathcal{M}^{+}(m^\circ)}\big\{\|\boldsymbol{\xi}^{\flat}_{m,m^\circ}\| \geq z^{\flat}_{m,m^\circ}(\mathtt{x}+q^{\flat}_{m^\circ})\big\}\right) = \mathrm{e}^{-\mathtt{x}}.$$

Finally, the bootstrap critical values are fixed by the analog of (6):

$$\mathfrak{z}^{\flat}_{m,m^\circ} \stackrel{\mathrm{def}}{=} z^{\flat}_{m,m^\circ}(\mathtt{x}+q^{\flat}_{m^\circ}) + \beta\sqrt{\mathtt{p}^{\flat}_{m,m^\circ}}$$

for $\mathtt{p}^{\flat}_{m,m^\circ} = I\!E^{\flat}\|\boldsymbol{\xi}^{\flat}_{m,m^\circ}\|^2$. Remind that all these quantities are data-driven and depend upon the original data. Now we apply the SmA procedure with such defined critical values $\mathfrak{z}^{\flat}_{m,m^\circ}$.

- **Design Regularity** is measured by the value δ_Ψ

$$\delta_\Psi \stackrel{\text{def}}{=} \max_{i=1,\ldots,n} \|S^{-1/2}\Psi_i\|\sigma_i\,, \quad \text{where} \quad S \stackrel{\text{def}}{=} \sum_{i=1}^{n} \Psi_i\Psi_i^\top\sigma_i^2\,; \tag{7}$$

Obviously

$$\sum_{i=1}^{n} \|S^{-1/2}\Psi_i\|^2\sigma_i^2 \;=\; \operatorname{tr}\Bigl(\sum_{i=1}^{n} S^{-2}\Psi_i\Psi_i^\top\sigma_i^2\Bigr) = \operatorname{tr} I_p = p,$$

and therefore in typical situations the value δ_Ψ is of order $\sqrt{p/n}$.

- **Presmoothing bias** for a projector Π is described by the vector

$$\boldsymbol{B} \;=\; \Sigma^{-1/2}(\boldsymbol{f}^* - \Pi\boldsymbol{f}^*).$$

We will use the sup-norm $\|\boldsymbol{B}\|_\infty = \max_i |b_i|$ and the squared ℓ_2-norm $\|\boldsymbol{B}\|^2 = \sum_i b_i^2$ to measure the bias after presmoothing.

- **Stochastic noise after presmoothing** is described via the covariance matrix $\mathrm{Var}(\breve{\boldsymbol{\varepsilon}})$ of the smoothed noise $\breve{\boldsymbol{\varepsilon}} = \Sigma^{-1/2}(\boldsymbol{\varepsilon} - \Pi\boldsymbol{\varepsilon})$. Namely, this matrix is assumed to be sufficiently close to the unit matrix I_n, in particular, its diagonal elements should be close to one. This is measured by the operator norm of $\mathrm{Var}(\breve{\boldsymbol{\varepsilon}}) - I_n$ and by deviations of the individual variances $I\!\!E\breve{\varepsilon}_i^2$ from one:

$$\delta_1 \stackrel{\mathrm{def}}{=} \| \mathrm{Var}(\breve{\boldsymbol{\varepsilon}}) - I_n \|_{\mathrm{op}},$$

$$\delta_\varepsilon \stackrel{\mathrm{def}}{=} \max_i |I\!\!E\breve{\varepsilon}_i^2 - 1|.$$

In particular, in the case of homogeneous errors $\Sigma = \sigma^2 I_n$ and the smoothing operator Π as a p-dimensional projector, it holds

$$\begin{aligned} \mathrm{Var}(\breve{\boldsymbol{\varepsilon}}) &= (I_n - \Pi)^2 = I_n - \Pi \le I_n\,, \\ \delta_1 &= \| \mathrm{Var}(\breve{\boldsymbol{\varepsilon}}) - I_n \|_{\mathrm{op}} = \|\Pi\|_{\mathrm{op}} = 1, \\ \delta_\varepsilon &= \max_i |I\!\!E\breve{\varepsilon}_i^2 - 1| = \max_i |\Pi_{ii}|. \end{aligned}$$

- **Regularity of the smoothing operator** Π is required in Theorems 2, 3, and 4. This condition will be expressed via the norm of the rows $\Upsilon_i^\top$ of the matrix $\Upsilon \stackrel{\mathrm{def}}{=} \Sigma^{-1/2}\Pi\Sigma^{1/2}$ fulfill

$$\|\Upsilon_i^\top\| \leq \delta_\Psi, \qquad i = 1, \ldots, n. \tag{8}$$

This condition is in fact very close to the design regularity condition (7). To see this, consider the case of a homogeneous noise with $\Sigma = \sigma^2 I_n$ and $\Pi = \Psi^\top \left(\Psi\Psi^\top\right)^{-1}\Psi$. Then $\Upsilon = \Pi$ and (7) implies

$$\|\Upsilon_i^\top\| = \|\Psi^\top \left(\Psi\Psi^\top\right)^{-1}\Psi_i\| = \|\left(\Psi\Psi^\top\right)^{-1/2}\Psi_i\| \leq \delta_\Psi .$$

In general one can expect that (8) is fulfilled with some other constant which however, is of the same magnitude as δ_Ψ. For simplicity we use the same letter.

Let $\boldsymbol{Y} = \boldsymbol{f}^* + \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{f}^*, \Sigma)$ for $\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$.

$$\delta_\Psi \stackrel{\text{def}}{=} \max_{i=1,\ldots,n} \|S^{-1/2}\Psi_i\|\sigma_i\,, \quad \text{where} \quad S \stackrel{\text{def}}{=} \sum_{i=1}^n \Psi_i\Psi_i^\top\sigma_i^2\,;$$

$$\delta_\varepsilon = \max_i |I\!E\breve{\varepsilon}_i^2 - 1|, \qquad \boldsymbol{B} = \Sigma^{-1/2}(\boldsymbol{f}^* - \Pi\boldsymbol{f}^*).$$

Consider

$$\mathbb{Q} = \mathcal{L}\big(\boldsymbol{\xi}_{m,m^\circ}, m, m^\circ \in \mathcal{M}\big), \qquad \mathbb{Q}^\flat = \mathcal{L}^\flat\big(\boldsymbol{\xi}^\flat_{m,m^\circ}, m, m^\circ \in \mathcal{M}\big).$$

Theorem 2. *It holds on a random set* $\Omega_2(\mathtt{x})$ *with* $I\!P\big(\Omega_2(\mathtt{x})\big) \geq 1 - 3\mathrm{e}^{-\mathtt{x}}$ *:*

$$\|\mathbb{Q} - \mathbb{Q}^\flat\|_{\mathrm{TV}} \;\leq\; \frac{1}{2}\Delta_2(\mathtt{x}),$$

$$\Delta_2(\mathtt{x}) \stackrel{\text{def}}{=} 2\sqrt{\delta_\Psi^2\, p\, \mathtt{x}_n} + \sqrt{\delta_\varepsilon^2\, p} + \sqrt{\|\boldsymbol{B}\|_\infty^4\, p} + 4\,\delta_\Psi^2\,\|\boldsymbol{B}\|\,\big(1 + \sqrt{\mathtt{x}}\big).$$

where $\mathtt{x}_n = \mathtt{x} + \log(n)$ *.*

Theorem 3. (Bootstrap validity) *Assume the conditions of Theorem 2, and let the rows* $\Upsilon_i^\top$ *of the matrix* $\Upsilon \stackrel{\mathrm{def}}{=} \Sigma^{-1/2} \Pi \Sigma^{1/2}$ *fulfill* (8). *Then for each* $m^\circ \in \mathcal{M}$

$$\mathbb{P}\left(\max_{m > m^\circ} \left\{ \|\boldsymbol{\xi}_{m,m^\circ}\| - z^\flat_{m,m^\circ}(\mathtt{x} + q^\flat_{m^\circ}) \right\} \geq 0 \right) \leq 6\mathrm{e}^{-\mathtt{x}} + \sqrt{p}\, \Delta_0(\mathtt{x}),$$

where with $\mathtt{x}_n = \mathtt{x} + \log(n)$ *and* $\mathtt{x}_p = \mathtt{x} + \log(2p)$

$$\Delta_0(\mathtt{x}) \stackrel{\mathrm{def}}{=} \|\boldsymbol{B}\|_\infty^2 + \delta_\Psi^2 \|\boldsymbol{B}\| \sqrt{2\mathtt{x}} + 2\delta_\Psi \mathtt{x}_n + \delta_\Psi^2 \mathtt{x}_n + 2\delta_\Psi \sqrt{\mathtt{x}_p} + 2\delta_\Psi^2 \mathtt{x}_p.$$

The SmA procedure also involves the values $\mathtt{p}_{m,m^\circ}$ which are unknown and depend on the noise ε. The next result shows the bootstrap counterparts $\mathtt{p}^{\flat}_{m,m^\circ}$ can be well used in place of $\mathtt{p}_{m,m^\circ}$.

Theorem 4. *Assume the conditions of Theorem 2. Then it holds on a set* $\Omega_1(\mathtt{x})$ *with* $I\!P\bigl(\Omega_1(\mathtt{x})\bigr) \geq 1 - 3\mathrm{e}^{-\mathtt{x}}$ *for all pairs* $m < m^\circ \in \mathcal{M}$

$$\left| \frac{\mathtt{p}^{\flat}_{m,m^\circ}}{\mathtt{p}_{m,m^\circ}} - 1 \right| \leq \Delta_{\mathtt{p}} ,$$

$$\Delta_{\mathtt{p}} \stackrel{\mathrm{def}}{=} \|\boldsymbol{B}\|^2_{\infty} + 4\,\mathtt{x}_{\mathcal{M}}^{1/2}\,\delta_n^2\,\|\boldsymbol{B}\| + 4\mathtt{x}_{\mathcal{M}}^{1/2}\,\delta_n + 4\,\mathtt{x}_{\mathcal{M}}\,\delta_n^2 + \delta_{\varepsilon},$$

where $\mathtt{p}^{\flat}_{m,m^\circ} = I\!E^{\flat}\|\boldsymbol{\xi}^{\flat}_{m,m^\circ}\|^2$, $\mathtt{p}_{m,m^\circ} = I\!E\|\boldsymbol{\xi}_{m,m^\circ}\|^2$, *and* $\mathtt{x}_{\mathcal{M}} = \mathtt{x} + 2\log(|\mathcal{M}|)$.

The above results immediately imply all the oracle bounds for probabilistic loss with the obvious correction of the error terms.

Outline

Consider a regression problem for an unknown univariate function on $[0,1]$ with unknown inhomogeneous noise. The aim is to compare the bootstrap-calibrated procedure with the SmA procedure for the known noise and with the oracle estimator. We also check the sensitivity of the method to the choice of the presmoothing parameter $m^\dagger$.
We use a uniform design on $[0,1]$ and the Fourier basis $\{\psi_j(x)\}_{j=1}^{\infty}$ for approximation of the regression function f which is modelled in the form

$$f(x) = c_1\psi_1(x) + \ldots + c_p\psi_p(x),$$

where the $(c_j)_{1\le j\le p}$ are chosen randomly: with γ_j i.i.d. standard normal

$$c_j = \begin{cases} \gamma_j, & 1 \le j \le 10, \\ \gamma_j/(j-10)^2, & 11 \le j \le 200. \end{cases}$$

The noise intensity grows from low to high as x increases to one. We use $n_{\text{sim-bs}} = n_{\text{sim-theo}} = n_{\text{sim-calib}} = 1000$ samples for computing the bootstrap marginal quantiles and the theoretical quantiles and for checking the calibration condition. The maximal model dimension is $M = 34$ and we also choose $m^\dagger = M$. The calibration is run with $\mathtt{x} = 2$ and $\beta = 1$.

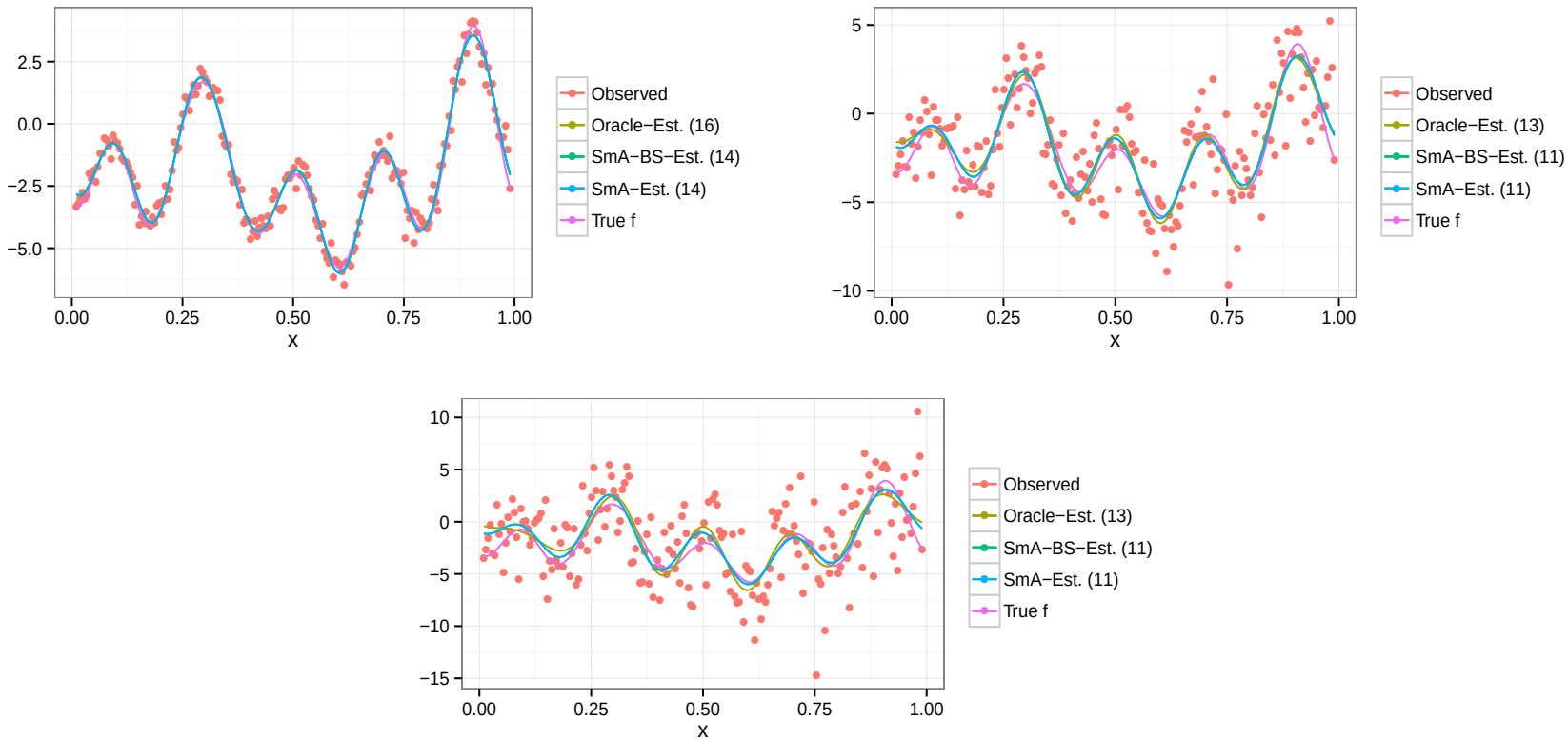

Figure : True functions and observed values plotted with oracle estimator, the known-variance SmA-Estimator (SmA-Est.) and the Bootstrap-SmA-Estimator (SmA-BS-Est.) for 3 different functions with different noise structure going from low noise to high noise. The numbers in parentheses indicate the chosen model dimension.

The oracles are respectively $m^* = 12$ for $n = 100, 200$ and $m^* = 10$ for $n = 50$.

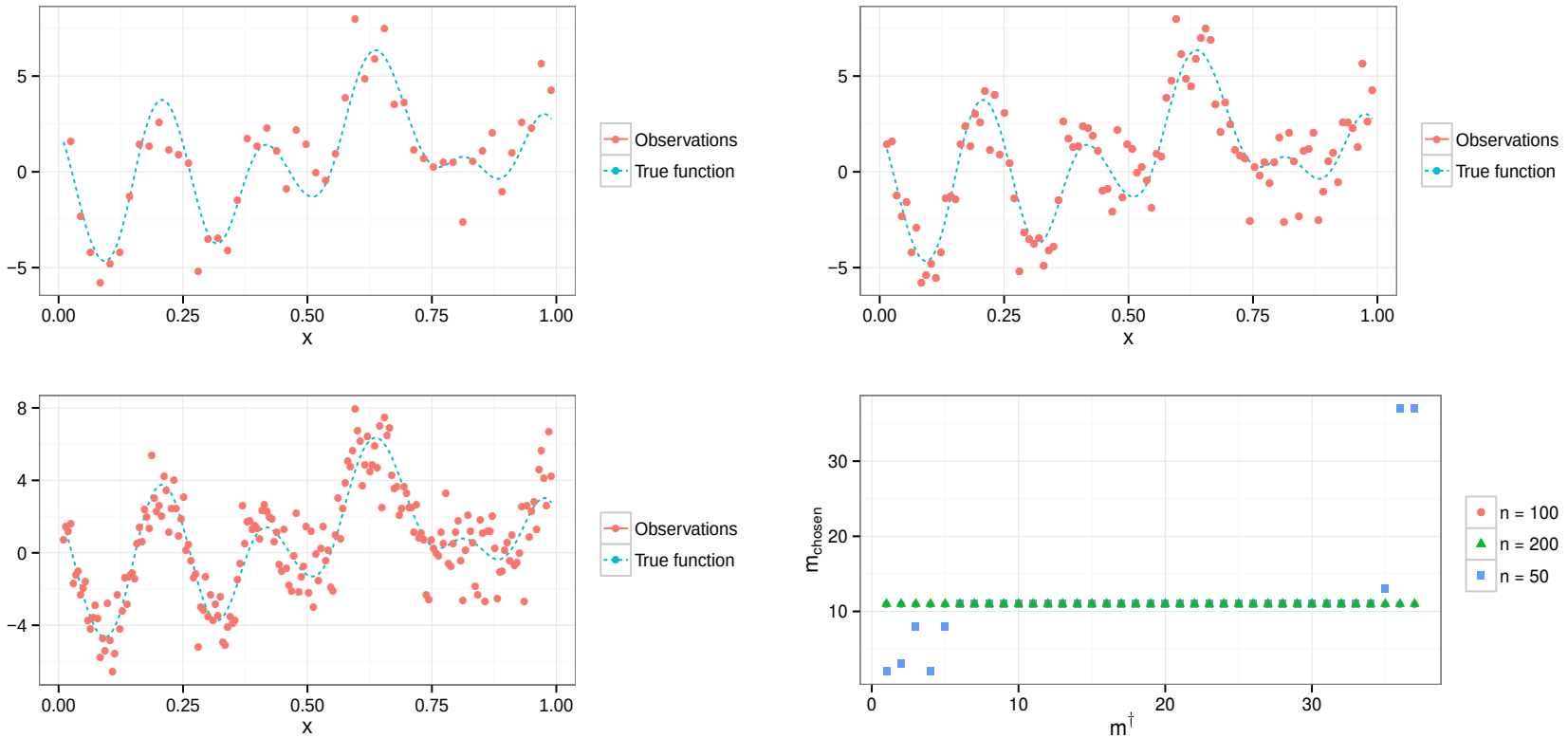

Figure : The first three plots show an exemplary function with $n = 50, 100, 200$ observations. The right plot shows the $\widehat{m}$ chosen by the Bootstrap-SmA-Method as a function of the calibration dimension $m^\dagger$ and the number of observations.

Figure 3 again demonstrates the dependence of the ratios on $m^\dagger$. It is remarkable that the ratio is varying very slowly above $m^* = 12$.

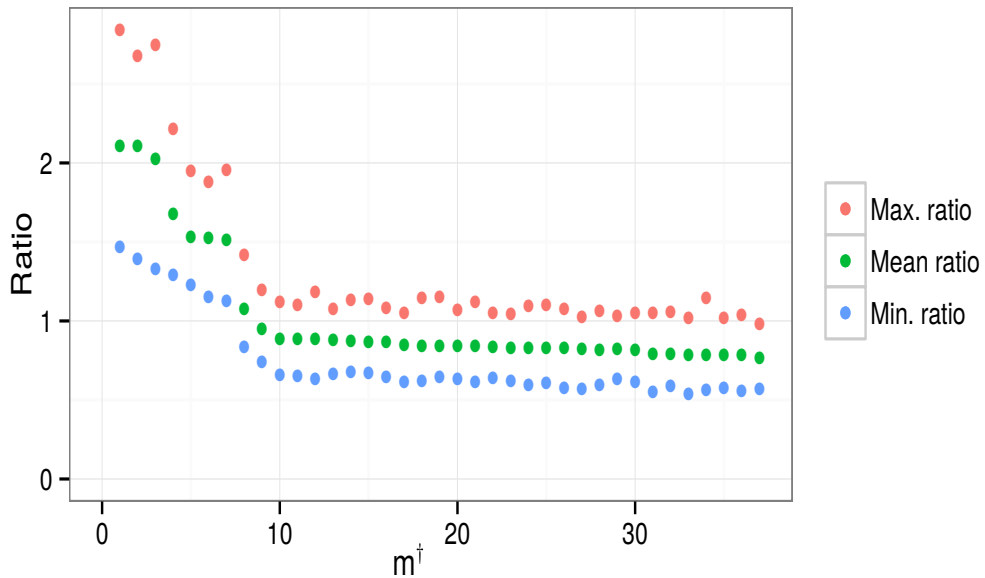

Figure : Maximal, minimal and mean ratio of the bootstrap and theoretical tail functions at $x = 2$, $|\mathfrak{z}^\flat_{m_1,m_2}/\mathfrak{z}_{m_1,m_2}|^2$ as a function of $m^\dagger$.

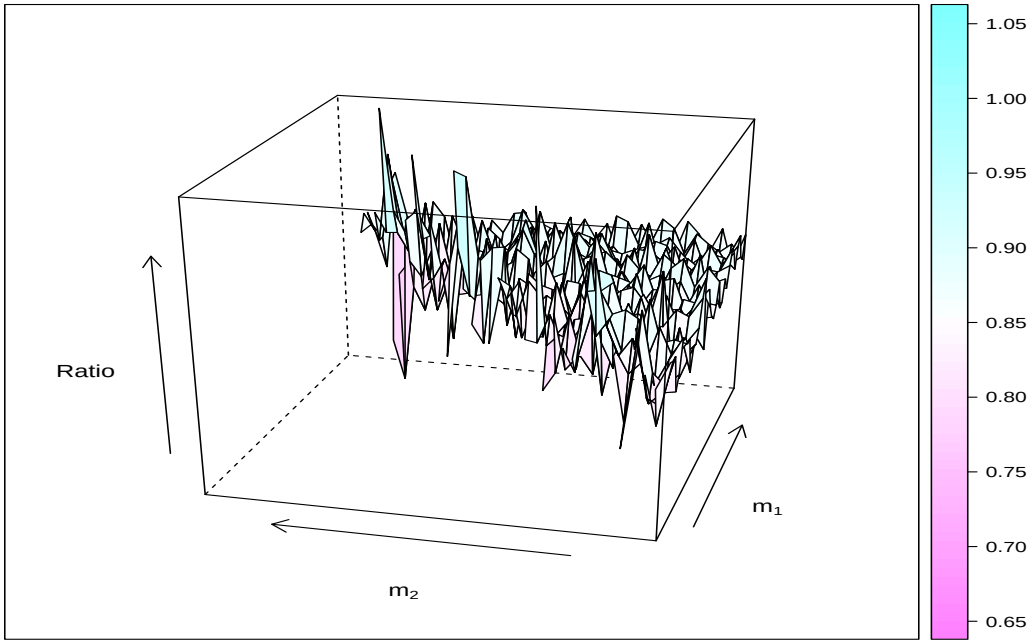

Figure : Ratio of quantiles $|\mathfrak{z}^{\flat}_{m_1,m_2}/\mathfrak{z}_{m_1,m_2}|^2$ for $m^\dagger = 20$ and $n = 200$ with the data and true function as in Fig. 2.

One realization

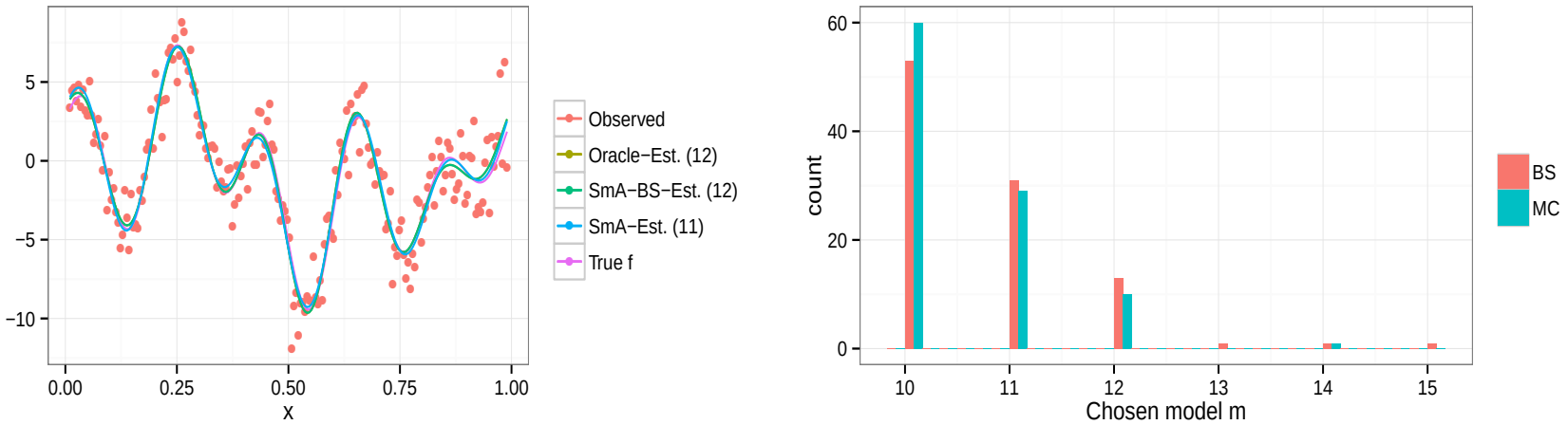

Figure : In the left plot, the true function and observed values are plotted for one realization together with the oracle estimator, the known-variance SmA-Estimator (SmA-Est.) and the Bootstrap-SmA-Estimator (SmA-BS-Est.). The numbers in parentheses indicate the chosen model dimension. In the right plot, histograms for the selected model are given for the bootstrap (BS) and the known-variance method (MC) for repeated observations of the same underlying function with a simulation size $n_{\text{hist}} = 100$.

Estimation of derivative

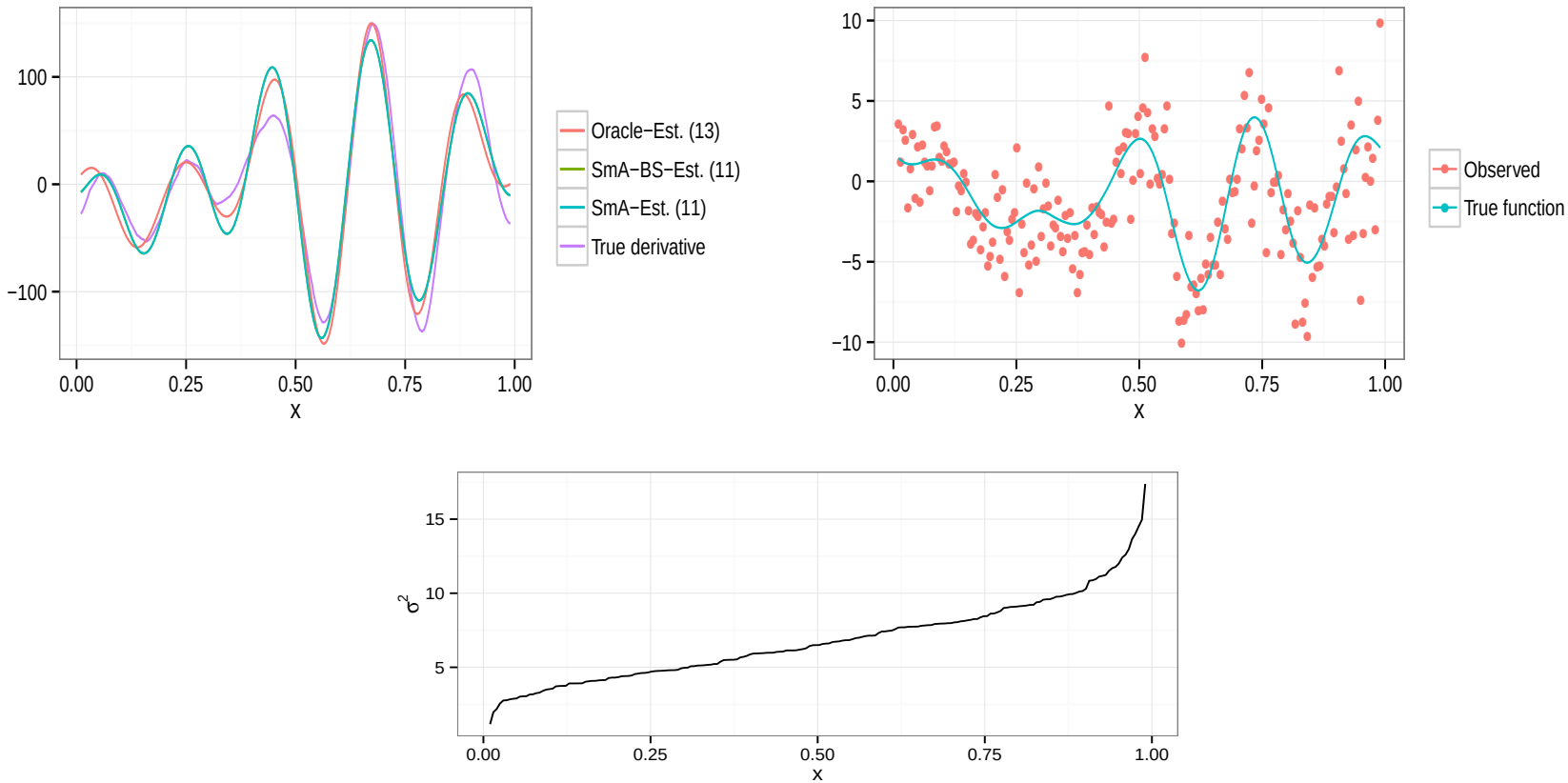

Figure : The upper left plot shows the true derivative, the oracle estimator, the known-variance SmA-Estimator (SmA-Est.) and the Bootstrap-SmA-Estimator (SmA-BS-Est.). The upper right plot shows the true function and the observations and in the lower plot one can find the standard deviation of the errors.

- A unified fully adaptive procedure for ordered model selection.
- Sharp oracle bounds
- Impact of the bias in the size of bootstrap confidence sets.

In progress:

- Model selection for unordered case like anisotropic classes. Theory applies but has to be extended;
- Active set selection. Theory applies but the problem of algorithmic efficient implementation.
- Large p s.t. p^2/n large; Use of sparse or complexity penalty.
- Extension to other problems like Hidden Markov Chain modeling;
- Multiscale change-point detection.

References

Arlot, S. (2009).

Model selection by resampling penalization.

Electron. J. Statist., 3:557–624.

Arlot, S. and Bach, F. R. (2009).

Data-driven calibration of linear estimators with minimal penalties.

In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 46–54. Curran Associates, Inc.

Barron, A., Birgé, L., and Massart, P. (1999).

Risk bounds for model selection via penalization.

Probab. Theory Related Fields, 113(3):301–413.

Birgé, L. (2001).

An alternative point of view on Lepski's method, volume Volume 36 of *Lecture Notes–Monograph Series*, pages 113–133.

Institute of Mathematical Statistics, Beachwood, OH.

Birgé, L. and Massart, P. (2007).

Minimal penalties for gaussian model selection.

Probability Theory and Related Fields, 138(1-2):33–73.

Cavalier, L., Golubev, G. K., Picard, D., and Tsybakov, A. B. (2002).

Oracle inequalities for inverse problems.

Ann. Statist., 30(3):843–874.

Cavalier, L. and Golubev, Y. (2006).

Risk hull method and regularization by projections of ill-posed inverse problems.

Ann. Statist., 34(4):1653–1677.

References

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014).

Anti-concentration and honest, adaptive confidence bands.

Ann. Statist., 42(5):1787–1818.

Dalalyan, A. S. and Salmon, J. (2012).

Sharp oracle inequalities for aggregation of affine estimators.

Ann. Statist., 40(4):2327–2355.

Gach, F., Nickl, R., and Spokoiny, V. (2013).

Spatially Adaptive Density Estimation by Localised Haar Projections.

Annales de l'Institut Henri Poincare - Probability and Statistics, 49(3):900–914.

DOI: 10.1214/12-AIHP485; arXiv:1111.2807.

Goldenshluger, A. (2009).

A universal procedure for aggregating estimators.

Ann. Statist., 37(1):542–568.

Kneip, A. (1994).

Ordered linear smoothers.

Ann. Statist., 22(2):835–866.

Lepski, O. V. (1990).

A problem of adaptive estimation in Gaussian white noise.

Teor. Veroyatnost. i Primenen., 35(3):459–470.

Lepski, O. V. (1991).

Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates.

Teor. Veroyatnost. i Primenen., 36(4):645–659.

References

Lepski, O. V. (1992).

Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates.

Teor. Veroyatnost. i Primenen., 37(3):468–481.

Lepski, O. V., Mammen, E., and Spokoiny, V. G. (1997).

Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors.

The Annals of Statistics, 25(3):929–947.

Lepski, O. V. and Spokoiny, V. G. (1997).

Optimal pointwise adaptive methods in nonparametric estimation.

Ann. Statist., 25(6):2512–2546.

Massart, P. (2007).

Concentration inequalities and model selection.

Number 1896 in Ecole d'Eté de Probabilités de Saint-Flour. Springer.

Spokoiny, V. and Vial, C. (2009).

Parameter tuning in pointwise adaptation using a propagation approach.

Ann. Statist., 37:2783–2807.

Spokoiny, V., Wang, W., and Härdle, W. (2013).

Local quantile regression (with rejoinder).

J. of Statistical Planing and Inference, 143(7):1109–1129.

ArXiv:1208.5384.

Tsybakov, A. (2000).

On the best rate of adaptive estimation in some inverse problems.

Comptes Rendus de l'Academie des Sciences - Series I - Mathematics, 330(9):835 – 840.