The Annals of Statistics
2019, Vol. 47, No. 3, 1351–1380
https://doi.org/10.1214/18-AOS1717

BOOTSTRAP TUNING IN GAUSSIAN ORDERED MODEL SELECTION[1]

BY VLADIMIR SPOKOINY[*,†,‡,§,¶] AND NIKLAS WILLRICH[*]

Weierstrass Institute Berlin[*], Humboldt University Berlin[†], IITP RAS[‡], HSE[§] and Skoltech Moscow[¶]

The paper focuses on the problem of model selection in linear Gaussian regression with unknown possibly inhomogeneous noise. For a given family of linear estimators $\{\widetilde{\boldsymbol{\theta}}_m, m \in \mathcal{M}\}$, ordered by their variance, we offer a new "smallest accepted" approach motivated by Lepski's device and the multiple testing idea. The procedure selects the smallest model which satisfies the acceptance rule based on comparison with all larger models. The method is completely data-driven and does not use any prior information about the variance structure of the noise: its parameters are adjusted to the underlying possibly heterogeneous noise by the so-called "propagation condition" using a wild bootstrap method. The validity of the bootstrap calibration is proved for finite samples with an explicit error bound. We provide a comprehensive theoretical study of the method, describe in details the set of possible values of the selected model $\widehat{m} \in \mathcal{M}$ and establish some oracle error bounds for the corresponding estimator $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$.

1. Introduction. Model selection is one of the key topics in mathematical statistics. A choice between models of differing complexity can often be viewed as a trade-off between overfitting the data by choosing a model which has too many degrees of freedom and smoothing out the underlying structure in the data by choosing a model which has too few degrees of freedom. This trade-off which shows up in most methods as the classical bias-variance trade-off is at the heart of every model selection method (as, e.g., in unbiased risk estimation, Kneip (1994) or in penalized model selection, Barron, Birgé and Massart (1999), Massart (2007)). This is also the case in Lepski's method, Lepskiĭ (1990, 1991, 1992), Lepski and Spokoiny (1997), Lepski, Mammen and Spokoiny (1997), Birgé (2001) and risk hull minimization, Cavalier and Golubev (2006). Many of these methods allow their strongest theoretical results only for highly idealized situations (e.g., sequence space models), are very specific to the type of problem under consideration (for instance, signal or functional estimation), require to know the noise behavior (like homogeneity) and the exact noise level. Moreover, they typically involve an unwieldy number of calibration constants whose choice is crucial to the

Received July 2015; revised April 2018.

[1]Supported by Russian Science Foundation Grant 14-50-00150. Financial support by the German Research Foundation (DFG) through the Research Training Group 1735 is gratefully acknowledged.

MSC2010 subject classifications. Primary 62G05; secondary 62G09, 62J15.

Key words and phrases. Smallest accepted, oracle, propagation condition.

applicability of the method and is not addressed by the theoretical considerations. For instance, any Lepski-type method requires to fix a numerical constant in the definition of the threshold, the theoretical results only apply if this constant is sufficiently large while the numerical results benefit from the choice of a rather small constant. Spokoiny and Vial (2009) offered a propagation approach to the calibration of Lepski's method in the case of the estimation of a one-dimensional quantity of interest. However, the proposal still requires the exact knowledge of the noise level and only applies to linear functional estimation. A similar approach has been applied to local constant density estimation with sup-norm risk in Gach, Nickl and Spokoiny (2013) and to local quantile estimation in Spokoiny, Wang and Härdle (2013).

In the case of unknown but homogeneous noise, generalized cross validation can be used instead of the unbiased risk estimation method. One can also apply one or another resampling method. Arlot (2009) suggested the use of resampling methods for the choice of an optimal penalization, following the framework of penalized model selection, Barron, Birgé and Massart (1999), Birgé and Massart (2007). The validity of a bootstrapping procedure for Lepski's method has also been studied in Chernozhukov, Chetverikov and Kato (2014) with new innovative technical tools with applications to honest adaptive confidence bands.

An alternative approach to adaptive estimation is based on aggregation of different estimates; see Goldenshluger (2009) and Dalalyan and Salmon (2012) for an overview of the existing results. However, the proposed aggregation procedures either require two independent copies of the data or involve a data splitting for estimating the noise variance. Each of these requirements is very restrictive for practical applications.

Another point to mention is that the majority of the obtained results on adaptive estimation focus on the quality of estimating the unknown response, that is, the loss is measured by the difference between the true response and its estimate. At the same time, inference questions like confidence estimation would require to know some additional information about the right model parameter. Only few results address the issue of estimating the oracle model. Moreover, there are some negative results showing that a construction of adaptive honest confidence sets is impossible without special conditions like self-similarity; see, for example, Giné and Nickl (2010).

This paper aims at developing a unified approach to the problem of ordered model selection with the focus on the quality of model selection rather than on accuracy of adaptive estimation. Our setup focuses on linear Gaussian regression and it equally applies to estimation of the whole parameter vectors, a subvector or linear mapping, as well as estimation of a linear functional. The proposed procedure and the theoretical study are also unified and do not distinguish between models and problems. The procedure does not use any prior information about the variance structure of the noise, the method automatically adjusts the parameters to the underlying possibly heterogeneous noise. The resampling technique allows

to achieve the same quality of estimation as if the noise structure were precisely known.

Consider a linear model $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ in $\mathbb{R}^n$ for an unknown parameter vector $\boldsymbol{\theta}^* \in \mathbb{R}^p$ and a given $p \times n$ design matrix Ψ. Suppose a family of linear smoothers $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$, $m \in \mathcal{M}$, to be fixed, where $\mathcal{M}$ is a set indexing the models considered and $\mathcal{S}_m$ 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. The proposed procedure called the "smallest accepted" (SmA) rule can be viewed as a calibrated Lepski-type method. The idea how the parameters of the method can be tuned, originates from Spokoiny and Vial (2009) and is related to a multiple testing problem. The whole procedure is based on a family of pairwise tests; each model is tested against all larger ones. Finally, the smallest accepted model is selected. The critical values for this multiple testing procedure are fixed using the so-called *propagation condition*. Unfortunately, the proposed approach requires the distribution of the errors $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \mathbb{E}\boldsymbol{Y}$ to be precisely known which is unrealistic in practical applications. Section 2.6 explains how the proposed procedure can be tuned in the case of Gaussian noise with unknown variance structure using a bootstrap method.

The paper presents a rigorous theoretical study of the proposed procedure for two cases. The first one corresponds to an idealistic situation that the noise distribution is precisely known; see Section 3.1. In particular, Theorem 3.1 presents finite sample results on the behavior of the proposed selector $\widehat{m}$ and the corresponding estimator $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$. It also describes a concentration set for the selected index $\widehat{m}$ and states a probabilistic oracle bound for the resulting estimator $\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{\widehat{m}}$. Usual rate results can be easily derived from these statements. Further results address the important quantity $\mathfrak{z}_{m^*}$ called "the payment for adaptation" which can be defined as the gap between oracle and adaptive bounds. Theorem 3.3 gives a general description of this quantity. Then we specify the results to important special cases like projection estimation and estimation of a linear functional. It appears, that in some cases the obtained results yield sharp asymptotic bounds. In some other cases, they lead to the usual log-price for data-driven model selection; Lepskiĭ (1992). An extension of the obtained probabilistic bounds to the case of a polynomial loss function is given in Section B of the Supplementary Material (Spokoiny and Willrich (2018)). The results are also specified to the particular problems of projection and linear functional estimation.

All the obtained results will be extended to regression models with unknown heterogeneous Gaussian noise (Section 3.6). Our main results about model selection in Gaussian regression with unknown heterogeneous noise are based on Theorem 3.6 which provides a kind of "bootstrap validity" statement: the bootstrap distribution mimics the unknown error variance with explicit error terms which can be controlled under usual regularity assumptions. This allows to extend the results obtained for the case of a known error distribution to the bootstrap calibrated procedure.

The paper is structured as follows. Section 2.1 explains our setup of ordered model selection, then Section 2.3 and Section 2.4 link the proposed approach to the multiple testing problem. The formal definition of the procedure is given in Section 2.5 for known noise and in Section 2.6 for the case of Gaussian errors with unknown variance. Section 3 states the main results, Section 4 illustrates the performance of the methods by numerical examples, while the proofs are gathered in the Appendix. The proofs of some technical results as well as some useful bounds for Gaussian quadratic forms and sums of random matrices are collected in the Supplementary Material (Spokoiny and Willrich (2018)).

2. SmA procedure. This section presents the proposed model selector. First, we specify our setup.

2.1. *Model and problem.* Consider the following linear regression model:

$$Y_i = \Psi_i^\top \boldsymbol{\theta}^* + \varepsilon_i, \qquad \mathbb{E}\varepsilon_i = 0, \qquad i = 1, \ldots, n,$$

with given design $\Psi_1, \ldots, \Psi_n$ in $\mathbb{R}^p$. Below we assume a deterministic design; otherwise, one can understand the results conditioned on the design. Further, $\boldsymbol{\theta}^*$ is an unknown vector in $\mathbb{R}^p$, and $\varepsilon_1, \ldots, \varepsilon_n$ are individual zero mean errors with finite variance. Our main results are stated under the assumption that individual errors $\varepsilon_i \stackrel{\text{def}}{=} Y_i - \mathbb{E}Y_i$ are independent normal and possibly heterogeneous, $\mathbb{E}\varepsilon_i^2 = \sigma_i^2$. However, Section 2.5 also discusses the case of an arbitrary but known error distribution.

The dimension p can be large, even $p = \infty$ can be incorporated. For notational simplicity, we proceed with p finite. The proposed approach can also be extended to the case when the linear parametric assumption $\mathbb{E}\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^*$ is not precisely fulfilled. Then, as usual, the target of estimation $\boldsymbol{\theta}^*$ can be defined as the vector of coefficients for the best approximation of the true response $\boldsymbol{f} \stackrel{\text{def}}{=} \mathbb{E}\boldsymbol{Y} = (f_1, \ldots, f_n)^\top$ by linear combinations of the feature vectors ψ_i which are the rows of the matrix Ψ. For the ease of notation, below we assume the linear parametric structure $f_i = \Psi_i^\top \boldsymbol{\theta}^*$. We write the underlying model in the vector form

$$\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\varepsilon} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}. \tag{2.1}$$

Let $\{\widetilde{\boldsymbol{\theta}}_m, m \in \mathcal{M}\}$ be a finite family of linear estimators $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$ of $\boldsymbol{\theta}^*$. Typical examples include projection estimation on a m-dimensional subspace or penalized estimators with a quadratic penalty function indexed by regularization parameter m, etc. To include specific problems like subvector/functional estimation, we also 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:

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

Alternatively, one can say that

$$\widetilde{\boldsymbol{\phi}}_m \stackrel{\mathrm{def}}{=} W\widetilde{\boldsymbol{\theta}}_m = W\widetilde{\boldsymbol{\theta}}_m = W\mathcal{S}_m\boldsymbol{Y} = \mathcal{K}_m\boldsymbol{Y}$$

with $\mathcal{K}_m = W\mathcal{S}_m$ is an estimator of the target $\boldsymbol{\phi}^* = W\boldsymbol{\theta}^*$ and $\varrho_m = \|\widetilde{\boldsymbol{\phi}}_m - \boldsymbol{\phi}^*\|^2$. Typical examples of W are as follows.

Estimation of the whole vector $\boldsymbol{\theta}^*$. Let W be the identity matrix $W = \boldsymbol{I}_p$ with $q = p$. This means that the *estimation loss* is measured by the usual squared Euclidean distance $\|\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*\|^2$.

Prediction. Let W be the square root of the $p \times p$ matrix $\mathbb{F} = \Psi\Psi^\top$, that is, $W^2 = \mathbb{F}$. The loss $\|W(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \|\Psi(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|$ is usually referred to as *prediction loss* because it measures the prediction ability of the true model by the model with the parameter $\boldsymbol{\theta}$.

Semiparametric estimation. Suppose that the target of estimation is not the whole vector $\boldsymbol{\theta}^*$ but some subvector $\boldsymbol{\theta}_0^*$ of dimension q. The matrix W can be defined as the projector Π_0 on the $\boldsymbol{\theta}_0^*$ subspace. The estimate $\Pi_0\widetilde{\boldsymbol{\theta}}_m$ is called the *profile maximum likelihood estimate*. The corresponding loss is equal to the squared Euclidean distance in this subspace:

$$\varrho_m = \|\Pi_0(\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\theta}^*)\|^2.$$

Alternatively, one can select W^2 as the efficient information matrix defined by relation $W^2 = (\Pi_0\mathbb{F}^{-1}\Pi_0^\top)^-$, where A^- means a pseudo-inverse of A.

Linear functional estimation. The choice of the weighting matrix W of rank one can be adjusted to address the problem of estimating some functionals of the whole parameter $\boldsymbol{\theta}^*$, for instance, the first coefficient θ_1^* or the sum of the θ_j^*'s.

In all cases, the most important feature of the estimators $\widetilde{\boldsymbol{\phi}}_m = \mathcal{K}_m\boldsymbol{Y}$ is *linearity*. It greatly simplifies the study of their properties including the prominent bias-variance decomposition of the risk of $\widetilde{\boldsymbol{\phi}}_m$. Namely, for the model (2.1) with $\mathbb{E}\boldsymbol{\varepsilon} = 0$ and $\boldsymbol{f} = \mathbb{E}\boldsymbol{Y}$, it holds

$$\begin{aligned} \mathbb{E}\widetilde{\boldsymbol{\phi}}_m &= \boldsymbol{\phi}_m^* \stackrel{\mathrm{def}}{=} \mathcal{K}_m\boldsymbol{f}, \\ \mathcal{R}_m &= \|\boldsymbol{\phi}_m^* - \boldsymbol{\phi}^*\|^2 + \operatorname{tr}\{\mathcal{K}_m \operatorname{Var}(\boldsymbol{\varepsilon})\mathcal{K}_m^\top\} = \|\boldsymbol{b}_m\|^2 + \mathfrak{p}_m, \end{aligned} \tag{2.2}$$

where $\|\boldsymbol{b}_m\|^2 \stackrel{\mathrm{def}}{=} \|\boldsymbol{\phi}_m^* - \boldsymbol{\phi}^*\|^2$ is the squared bias term and $\mathfrak{p}_m \stackrel{\mathrm{def}}{=} \operatorname{tr}\operatorname{Var}(\widetilde{\boldsymbol{\phi}}_m)$ is the variance term. This is the usual "bias-variance" decomposition of the squared risk $\mathcal{R}_m$. The optimal choice of m is often defined by risk minimization:

$$m_{\mathrm{opt}} \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_{m\in\mathcal{M}} \mathcal{R}_m = \operatorname*{argmin}_{m\in\mathcal{M}} (\|\boldsymbol{b}_m\|^2 + \mathfrak{p}_m). \tag{2.3}$$

Alternatively, one can define the best choice m^* via the "bias-variance trade-off;" see the definition below in (2.9). The *model selection* problem can be described as a data-based choice $\widehat{m}$ which leads to essentially the same quality of the adaptive estimator $\widetilde{\boldsymbol{\theta}}_{\widehat{m}}$ as for the optimal choice m^*.

2.2. *Ordered case.* Below we discuss the *ordered* case. For simplicity of presentation, we assume that $\mathcal{M}$ is a finite set of positive numbers, although the approach can be extended to situations with a countable and/or continuous and even unbounded set $\mathcal{M}$ using a discretization. Let $|\mathcal{M}|$ stand for the cardinality of $\mathcal{M}$. Typical examples of the parameter m are given by a chosen dimension (number of basis vectors) in projection estimation or by the bandwidth in kernel smoothing. In general, complexity can be naturally expressed via the variance of the stochastic term of the estimator $\widetilde{\boldsymbol{\phi}}_m$: the larger m, the larger is the variance term $\mathtt{p}_m = \operatorname{tr}\{\operatorname{Var}(\widetilde{\boldsymbol{\phi}}_m)\}$. In the case of projection estimation and a homogeneous noise $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$, this variance term is linear in m: $\mathtt{p}_m = \sigma^2 m$; see Section 3.3 for details. In general, dependence of the variance term on m is more complicated but monotonicity of $\mathtt{p}_m$ in m should be preserved. The related condition can be written as

$$\operatorname{tr}\bigl(\mathcal{K}_m \operatorname{Var}(\boldsymbol{\varepsilon})\mathcal{K}_m^\top\bigr) < \operatorname{tr}\bigl(\mathcal{K}_{m'} \operatorname{Var}(\boldsymbol{\varepsilon})\mathcal{K}_{m'}^\top\bigr), \qquad m' > m. \tag{2.4}$$

Monotonicity assumption yields, in particular, that the total number $|\mathcal{M}|$ of considered models $m \in \mathcal{M}$ is not large. Usually it is bounded by n^a for some $a < 1$, and the use of a geometric scaling allows to reduce $|\mathcal{M}|$ to $\mathtt{C} \log n$; see, for example, Lepskiĭ (1990), Lepski, Mammen and Spokoiny (1997). This is in striking difference with the case of unordered or partially ordered model selection problem when the number of models can be huge and the value $\log|\mathcal{M}|$ can be comparable with the sample size; see, for example, Birgé and Massart (2007). Further, it is implicitly assumed that the bias term $\|\boldsymbol{b}_m\|^2 = \|\boldsymbol{\phi}_m^* - \boldsymbol{\phi}^*\|^2$ becomes smaller as m increases. The smallest index $m_0 \in \mathcal{M}$ corresponds to the simplest (zero) model, usually with a large bias, while a large m ensures a good approximation quality $\boldsymbol{\phi}_m^* \approx \boldsymbol{\phi}^*$ and a small bias at cost of a big complexity measured by the variance term. In the case of projection estimation, the bias term in (2.2) describes the accuracy of approximating the response $\boldsymbol{f}$ by an m-dimensional linear subspace and this approximation improves as m grows. However, in general, in contrast to the case of projection estimation, one cannot require that the squared bias $\|\boldsymbol{b}_m\|^2$ monotonously decreases with m. An example is given below.

EXAMPLE 2.1. Suppose that a signal $\boldsymbol{\theta}^*$ is observed with noise: $Y_i = \theta_i^* + \varepsilon_i$. Consider the set of projection estimates $\widetilde{\boldsymbol{\theta}}_m$ on the first m coordinates and the target is $\phi^* \stackrel{\mathrm{def}}{=} W\boldsymbol{\theta}^* = \sum_j \theta_j^*$. If $\boldsymbol{\theta}^*$ is composed of alternating blocks of 1's and -1's with equal length, then the bias $|\phi^* - \phi_m^*|$ for $\phi_m^* = \sum_{j \le m} \theta_j^*$ is not monotonous in m.

2.3. *Smallest accepted (SmA) method in ordered model selection.* This section presents the basics of the SmA procedure, in particular, relations to the multiple testing problem.

Suppose we are given an ordered set of linear estimators $\widetilde{\boldsymbol{\phi}}_m$ of the q-dimensional target of estimation $\boldsymbol{\phi}^* = W\boldsymbol{\theta}^*$, that is, $\widetilde{\boldsymbol{\phi}}_m = W\mathcal{S}_m\boldsymbol{Y} = \mathcal{K}_m\boldsymbol{Y}$ with $q \times n$ matrices $\mathcal{K}_m = W\mathcal{S}_m$ for $m \in \mathcal{M}$, and (2.4) holds. Below we present a general approach to model selection problems based on multiple testing. In the problem of choosing m, we face a usual dilemma: an increase of complexity of the method m yields an increase of the variance term but probably improves the approximation quality measured by the bias term $\|\boldsymbol{b}_m\|^2$. Thus, we aim at picking up a possibly small index $m^\circ \in \mathcal{M}$ for which a further increase of the index m over m° only increases the complexity of the method without real gain in the quality of approximation. The 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°. Introduce for each pair $m > m^\circ$ from $\mathcal{M}$ a hypothesis H_{m,m° of "no significant difference between the models m° and m;" see the next section for a precise formulation. Let τ_{m,m° be the corresponding test. The model m° is *accepted* if $\tau_{m,m^\circ} = 0$ for all $m > m^\circ$. This can be viewed a multiple test of the set of hypotheses $\mathcal{H}_{m^\circ} = \{H_{m,m^\circ}, m > m^\circ\}$. 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

$$\tau_{m,m^\circ} = \mathbb{1}\{\mathbb{T}_{m,m^\circ} > \mathfrak{z}_{m,m^\circ}\} \tag{2.5}$$

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 $\mathbb{T}_{m,m^\circ} = \sigma^{-2}\|\Psi^\top(\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ})\|^2$. 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 null hypothesis $\mathbb{E}\widetilde{\boldsymbol{\theta}}_m = \mathbb{E}\widetilde{\boldsymbol{\theta}}_{m^\circ}$ and homogeneous Gaussian noise with known variance σ^2, this makes it simple to compute the corresponding critical values. However, under more general assumptions on the noise distribution, and it is more convenient to apply another choice corresponding to Lepski-type procedure and based on the norm of differences $\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}$:

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

where $\mathcal{K}_{m,m^\circ} \stackrel{\mathrm{def}}{=} \mathcal{K}_m - \mathcal{K}_{m^\circ}$. The main issue for such a method is a proper choice of the critical values $\mathfrak{z}_{m,m^\circ}$ in (2.5). One can say that the procedure is specified by a way of selecting these critical values. Below we fix these values by imposing a so-called *propagation property*: a "good" model m° for which all H_{m,m° with $m > m^\circ$ are true, has to be accepted with a high probability. This rule can be seen as an analogue of the family-wise error rate condition in a multiple testing problem.

2.4. *A "good" model.* This section aims at formalizing the above mentioned relations between model selection and multiple testing. We use below for each pair $m > m^\circ$ from $\mathcal{M}$ the decomposition of the test statistic $\mathbb{T}_{m,m^\circ}$:

$$\begin{aligned}\mathbb{T}_{m,m^\circ} &= \|\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}\| = \|\mathcal{K}_{m,m^\circ}\boldsymbol{Y}\| \\ &= \|\mathcal{K}_{m,m^\circ}(\boldsymbol{f} + \boldsymbol{\varepsilon})\| = \|\boldsymbol{b}_{m,m^\circ} + \boldsymbol{\xi}_{m,m^\circ}\|,\end{aligned} \tag{2.6}$$

with $\mathcal{K}_{m,m^\circ} = \mathcal{K}_m - \mathcal{K}_{m^\circ}$, where $\boldsymbol{b}_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_{m,m^\circ}\boldsymbol{f} \in \mathbb{R}^q$ is the deterministic *bias* vector, while $\boldsymbol{\xi}_{m,m^\circ} \stackrel{\text{def}}{=} \mathcal{K}_{m,m^\circ}\boldsymbol{\varepsilon} \in \mathbb{R}^q$ is the *stochastic* component. It obviously holds $\mathbb{E}\boldsymbol{\xi}_{m,m^\circ} = 0$. Introduce the $q \times q$-matrix $\mathbb{V}_{m,m^\circ}$ as the variance of $\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ} = \mathcal{K}_{m,m^\circ}\boldsymbol{Y}$:

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

If the noise $\boldsymbol{\varepsilon}$ is homogeneous with $\operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}_n$, it holds

$$\mathbb{V}_{m,m^\circ} = \sigma^2\mathcal{K}_{m,m^\circ}\mathcal{K}_{m,m^\circ}^\top.$$

Further,

$$\begin{aligned}\mathbb{E}\mathbb{T}_{m,m^\circ}^2 &= \|\boldsymbol{b}_{m,m^\circ}\|^2 + \mathbb{E}\|\boldsymbol{\xi}_{m,m^\circ}\|^2 = \|\boldsymbol{b}_{m,m^\circ}\|^2 + \mathfrak{p}_{m,m^\circ}, \\ \mathfrak{p}_{m,m^\circ} &\stackrel{\text{def}}{=} \mathbb{E}\|\boldsymbol{\xi}_{m,m^\circ}\|^2 = \operatorname{tr}(\mathbb{V}_{m,m^\circ}).\end{aligned} \tag{2.7}$$

For a fixed $m^\circ \in \mathcal{M}$, let

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

A "good" choice m° can be defined by the condition that, for each $m \in \mathcal{M}^+(m^\circ)$, the bias term $\|\boldsymbol{b}_{m,m^\circ}\|^2$ is not significantly larger than the variance term $\mathfrak{p}_{m,m^\circ}$. This condition can be quantified in the following "bias-variance trade-off" relation:

$$H_{m,m^\circ}\colon \|\boldsymbol{b}_{m,m^\circ}\|^2 \le \beta^2\mathfrak{p}_{m,m^\circ},$$

with a given parameter β. This can be viewed as a null hypothesis of "no significant difference" between models with parameters m° and m. For each candidate model m°, define a set of hypotheses

$$\mathcal{H}_{m^\circ} = \{H_{m,m^\circ}\colon \|\boldsymbol{b}_{m,m^\circ}\|^2 \le \beta^2\mathfrak{p}_{m,m^\circ}, m \in \mathcal{M}^+(m^\circ)\}. \tag{2.8}$$

A "good" model m° is one with all hypotheses in this set $\mathcal{H}_{m^\circ}$ fulfilled. Below this set of hypotheses will be considered for each m° separately. Now define the *oracle* m^* as the minimal m° under (2.8):

$$m^* \stackrel{\text{def}}{=} \min\Big\{m^\circ\colon \max_{m \in \mathcal{M}^+(m^\circ)}\{\|\boldsymbol{b}_{m,m^\circ}\|^2 - \beta^2\mathfrak{p}_{m,m^\circ}\} \le 0\Big\}. \tag{2.9}$$

Clearly, the notion of a "good" model depends on the value β, in particular, $m^* = m^*(\beta)$. Also m^* does not coincide with the risk minimizer m_{opt} from (2.3). However, both definitions exhibit bias-variance trade-off in (2.7).

2.5. *Calibration of the SmA procedure for the known noise distribution.* This section explains the choice of the critical values $\mathfrak{z}_{m,m^\circ}$ for the idealistic case when the noise distribution is precisely known. This greatly helps to explain the essence of the approach. Section 2.6 presents a data-driven procedure for the unknown noise variance using a resampling technique.

For a fixed m°, the related set of critical values $\mathfrak{z}_{m,m^\circ}$ should be fixed to ensure a prescribed family-wise error rate (FWER) $\mathrm{e}^{-\mathtt{x}}$ of the family of tests $\mathbb{1}(\mathbb{T}_{m,m^\circ} > \mathfrak{z}_{m,m^\circ})$ for $m \in \mathcal{M}$ with $m > m^\circ$. In the terminology of Romano and Wolf (2005), this is a *weak* FWER control.

Let us start with $\beta = 0$ corresponding to the set $\mathcal{H}_{m^\circ}$ of hypotheses H_{m,m° : $\boldsymbol{b}_{m,m^\circ} = 0$ for all $m > m^\circ$. In this situation, the test statistic $\mathbb{T}_{m,m^\circ}$ coincides under H_{m,m° with the norm of the stochastic term $\boldsymbol{\xi}_{m,m^\circ}$ whose distribution is precisely known under given noise. For instance, if errors $\boldsymbol{\varepsilon}$ are Gaussian, then the stochastic component $\boldsymbol{\xi}_{m,m^\circ}$ is a normal zero mean vector with the covariance matrix $\mathbb{V}_{m,m^\circ}$. Introduce for each pair $m > m^\circ$ from $\mathcal{M}$ a *tail function* $z_{m,m^\circ}(t)$ of the argument t such that

$$\mathbb{P}\bigl(\|\boldsymbol{\xi}_{m,m^\circ}\| > z_{m,m^\circ}(t)\bigr) = \mathrm{e}^{-t}. \tag{2.10}$$

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 the smallest value for which the deviation probability is smaller than e^{-t}. For multiple testing, we need a uniform in $m > m^\circ$ version of the probability bound (2.10). To guarantee the prescribed FWER for the set of hypotheses $\mathcal{H}_{m^\circ}$, introduce, given $\mathtt{x}$, the multiplicity correction $q_{m^\circ} = q_{m^\circ}(\mathtt{x})$:

$$\mathbb{P}\Biggl(\bigcup_{m \in \mathcal{M}^+(m^\circ)} \bigl\{\|\boldsymbol{\xi}_{m,m^\circ}\| \geq z_{m,m^\circ}(\mathtt{x} + q_{m^\circ})\bigr\}\Biggr) \leq \mathrm{e}^{-\mathtt{x}}. \tag{2.11}$$

A simple way of fixing the value q_{m° is based on the Bonferroni bound: $q_{m^\circ} = \log(|\mathcal{M}^+(m^\circ)|)$; cf. Spokoiny (1996) in context of adaptive testing. However, it is well known that the Bonferroni correction is very conservative and results in a large q_{m°; see, for example, Baraud, Huet and Laurent (2003). This is especially striking if the random vectors $\boldsymbol{\xi}_{m,m^\circ}$ are strongly correlated, which is exactly the case under consideration. As the joint distribution of the $\boldsymbol{\xi}_{m,m^\circ}$'s is precisely known, one can define the correction q_{m° just as the smallest value ensuring (2.11); cf. (5) in Baraud, Huet and Laurent (2003). This choice $z_{m,m^\circ}(\mathtt{x} + q_{m^\circ})$ of the critical values yields automatically the weak FWER bound for the set of hypotheses $\mathcal{H}_{m^\circ} = \{H_{m,m^\circ}, m > m^\circ\}$ with $\beta = 0$. Moreover, the FWER control would fail for any other uniformly smaller set of critical values.

In the case of β positive, we define the critical values $\mathfrak{z}_{m,m^\circ} = \mathfrak{z}_{m,m^\circ}(\mathtt{x})$ by one more correction for the bias term $\|\boldsymbol{b}_{m,m^\circ}\|$:

$$\mathfrak{z}_{m,m^\circ} \stackrel{\mathrm{def}}{=} z_{m,m^\circ}(\mathtt{x} + q_{m^\circ}) + \beta\sqrt{\mathfrak{p}_{m,m^\circ}} \tag{2.12}$$

for $\mathfrak{p}_{m,m^\circ} = \operatorname{tr}(\mathbb{V}_{m,m^\circ})$. The bound (2.11) automatically ensures the desired *propagation property*: any good model m° in the sense (2.8) will be *rejected* with probability at most $\mathrm{e}^{-\mathtt{x}}$ in the following sense:

$$\begin{aligned}\mathbb{P}\bigl(m^\circ \text{ is rejected}\bigr) &\stackrel{\text{def}}{=} \mathbb{P}\Bigl(\bigcup_{m\in\mathcal{M}^+(m^\circ)} \bigl\{\|\mathbb{T}_{m,m^\circ}\| \geq \mathfrak{z}_{m,m^\circ}(\mathtt{x})\bigr\}\Bigr) \\ &\leq \mathbb{P}\Bigl(\bigcup_{m\in\mathcal{M}^+(m^\circ)} \bigl\{\|\boldsymbol{\xi}_{m,m^\circ}\| \geq z_{m,m^\circ}(\mathtt{x}+q_{m^\circ})\bigr\}\Bigr) \leq \mathrm{e}^{-\mathtt{x}}.\end{aligned} \tag{2.13}$$

The last inequality follows from (2.6). One can say this is a built-in property of the procedure. By definition, the oracle m^* is also the smallest "good" choice, this yields due to (2.13)

$$\mathbb{P}\bigl(m^* \text{ is rejected}\bigr) \leq \mathrm{e}^{-\mathtt{x}}. \tag{2.14}$$

Definition (2.12) still involves two numerical constants $\mathtt{x}$ and β. It is quite common in the model selection literature to define the optimal choice of tuning parameters by minimization of the risk of the resulting procedure. Unfortunately, it does not apply in our setup which is based on multiple testing. Note however, that these values are not tuning parameters of the method, they rather serve to fix some expected features of the method. The value $\mathtt{x}$ defines the nominal FWER $\mathrm{e}^{-\mathtt{x}}$. Similar to the testing problem, there is no unique choice for $\mathtt{x}$, a usual choice of $\mathtt{x}$ in the range between 2 and 3 can be recommended. The value β controls the amount of admissible bias in the definition of a good model; cf. (2.8) and (2.9). The natural choice for β is $\beta = 1$ which balances the bias and variance terms in (2.8). Note, however, that the procedure and the theoretical results hold for any combination of these parameters. We only require that the value β is the same in the definition of a good model and in formula (2.12) for the critical values $\mathfrak{z}_{m,m^\circ}$. Our default choice is $\mathtt{x} = 2$, $\beta = 1$. An intensive numerical study indicates a very minor change in the estimation results for moderate deviations of these parameters around the mentioned default choice.

Define the selector $\widehat{m}$ by the "smallest accepted" (SmA) rule. Namely, with $\mathfrak{z}_{m,m^\circ}$ from (2.12), the acceptance rule reads as follows:

$$\bigl\{m^\circ \text{ is accepted}\bigr\} = \Bigl\{\max_{m\in\mathcal{M}^+(m^\circ)} \{\mathbb{T}_{m,m^\circ} - \mathfrak{z}_{m,m^\circ}\} \leq 0\Bigr\}. \tag{2.15}$$

The SmA choice is defined by the "smallest accepted" rule:

$$\widehat{m} \stackrel{\text{def}}{=} \min\Bigl\{m^\circ\colon \max_{m\in\mathcal{M}^+(m^\circ)} \{\mathbb{T}_{m,m^\circ} - \mathfrak{z}_{m,m^\circ}\} \leq 0\Bigr\}. \tag{2.16}$$

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 the statements about $\widehat{m}$. The desired solution would be $\widehat{m} \equiv m^*$, then the adaptive estimator $\widehat{\boldsymbol{\phi}}$ coincides with the oracle estimator $\widetilde{\boldsymbol{\phi}}_{m^*}$.

REMARK 2.1. The SmA procedure originates from Lepskiĭ (1990). However, Lepski's acceptance rule for a candidate m° is a bit stronger: it requires that each larger model $m > m^\circ$ is accepted as well, that is, all hypotheses $H_{m',m}$ for $m' > m \geq m^\circ$ are accepted. This allows to efficiently implement the procedure as a top-down algorithm: start from the largest model index m and check acceptance by the criterion (2.15) until rejection. Our acceptance rule is similar to Birgé (2001) and it can be implemented as a bottom-up algorithm: start from the smallest model and check each new candidate m° by rule (2.15) until the first acceptance. Note that the way of computing the critical values by multiplicity arguments can be used for the original Lepski's rule as well. It however requires an additional correction due to the more strict acceptance rule. More precisely, define for each t and each m° the correction $q_{m^\circ}(t)$ similar to (2.11):

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

Further, given x, define an additional correction $q^+ = q^+(\mathrm{x})$ by

$$\mathbb{P}\Big(\bigcup_{m^\circ \in \mathcal{M}} \bigcup_{m \in \mathcal{M}^+(m^\circ)} \{ \|\boldsymbol{\xi}_{m,m^\circ}\| \geq z_{m^\circ,m}\big(\mathrm{x} + q_{m^\circ}(\mathrm{x}) + q^+\big) \} \Big) \leq \mathrm{e}^{-\mathrm{x}}. \tag{2.17}$$

Finally define the critical values $\mathfrak{z}^L_{m,m^\circ} = \mathfrak{z}^L_{m,m^\circ}(\mathrm{x})$ in the form

$$\mathfrak{z}^L_{m,m^\circ} = z_{m,m^\circ}\big(\mathrm{x} + q_{m^\circ}(\mathrm{x}) + q^+\big) + \beta \sqrt{\mathfrak{p}_{m,m^\circ}}.$$

Again, this construction allows to build a set of critical values which guarantees the propagation property for Lepski's procedure. The correction (2.17) can be viewed as a special case of a classical proposal for simultaneous structured testing; see, for example, Marcus, Peritz and Gabriel (1976) or Romano and Wolf (2005) and of a sequential rejection principle from Goeman and Solari (2010).

2.6. *Bootstrap tuning*. This section explains how the proposed SmA procedure can be applied in the case of Gaussian *heterogeneous* noise with *unknown* covariance matrix $\Sigma = \mathrm{Var}(\boldsymbol{\varepsilon}) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. Let the observed data $\boldsymbol{Y}$ follow the model $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$. Assume to be given an ordered family of linear estimators $\widetilde{\boldsymbol{\phi}}_m = \mathcal{K}_m \boldsymbol{Y} = W \mathcal{S}_m \boldsymbol{Y}$ of the target $\boldsymbol{\phi}^* = W\boldsymbol{\theta}^*$, $m \in \mathcal{M}$. For each pair $m > m^\circ$ from $\mathcal{M}$, we consider the test statistic $\mathbb{T}_{m,m^\circ}$ and its decomposition from (2.6):

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

Calibration of the SmA model selection procedure requires to know the joint distribution of all corresponding stochastic terms $\|\boldsymbol{\xi}_{m,m^\circ}\|$ for $m > m^\circ$ which is uniquely determined by the noise covariance matrix Σ. In the case when this matrix is unknown, we are going to use a bootstrap procedure to approximate this distribution.

It allows to mimics the unknown heterogeneous noise, however, the bootstrap validity result requires that the parameter dimension of the largest considered model is not too big; see Section 3.7 for details. The proposed procedure relates to the concept of the *wild* bootstrap, Wu (1986), Beran (1986) or Härdle and Mammen (1993). In the framework of a regression problem, it suggests to model the unknown heteroscedastic noise using randomly weighted residuals from pilot estimation. We apply normal weights; for other weighting schemes see, for example, Mammen (1993).

Suppose we are given a pilot estimator (presmoothing) $\widetilde{\boldsymbol{f}}$ of the response vector $\boldsymbol{f} = \mathbb{E}\boldsymbol{Y} \in \mathbb{R}^n$. Define the residuals:

$$\check{\boldsymbol{Y}} \stackrel{\text{def}}{=} \boldsymbol{Y} - \widetilde{\boldsymbol{f}}.$$

About this pilot it is supposed that the related bias is negligible and the variance of $\check{\boldsymbol{Y}}$ is close to Σ. This presmoothing assumes some minimal regularity of the response $\boldsymbol{f}$ (usually expressed via minimal smoothness of the underlying regression function), and this condition seems to be unavoidable if no information about the noise is given: otherwise one cannot distinguish between signal and noise. Below we suppose that $\widetilde{\boldsymbol{f}}$ is a linear predictor, $\widetilde{\boldsymbol{f}} = \Pi\boldsymbol{Y}$, where Π is a sub-projector in the space $\mathbb{R}^n$. For example, one can take $\Pi = \Pi_{m^\dagger}$, where $\Pi_m = \Psi_m^\top(\Psi_m\Psi_m^\top)^- \Psi_m$, Ψ_m is the rank m feature matrix corresponding to the first m features, and $m^\dagger \in \mathcal{M}$ corresponds to a model with a possibly small bias, for example, the largest model M in our collection $\mathcal{M}$. The wild bootstrap proposes to resample from the heteroscedastic Gaussian noise with the covariance matrix

$$\check{\Sigma} = \operatorname{diag}(\check{\boldsymbol{Y}} \cdot \check{\boldsymbol{Y}}) = \operatorname{diag}(\check{Y}_1^2, \ldots, \check{Y}_n^2),$$

where $\check{\boldsymbol{Y}} \cdot \check{\boldsymbol{Y}}$ denotes the coordinate-wise product of the vector $\check{\boldsymbol{Y}}$ with itself and $\operatorname{diag}(\check{\boldsymbol{Y}} \cdot \check{\boldsymbol{Y}})$ denotes the diagonal matrix with entries $\check{Y}_i^2$. These entries depend on $\boldsymbol{Y}$ and thus are random. Therefore, the bootstrap distribution is a random measure on $\mathbb{R}^n$ and the aim of our study is to show that this random measure mimics well the underlying data distribution for typical realizations of $\boldsymbol{Y}$. Clearly, $\check{\Sigma} = \operatorname{diag}(\check{\boldsymbol{Y}} \cdot \check{\boldsymbol{Y}})$ is a very poor estimator of Σ. However, under realistic conditions on the pilot $\widetilde{\boldsymbol{f}}$ and on the model, it allows to obtain essentially the same results as in the case of known Σ.

Let $\boldsymbol{w}^\flat$ denote the n-vector of bootstrap standard Gaussian weights, $\boldsymbol{w}^\flat \sim \mathcal{N}(0, \boldsymbol{I}_n)$. Clearly, the product $\boldsymbol{\varepsilon}^\flat = \operatorname{diag}(\check{\boldsymbol{Y}})\boldsymbol{w}^\flat$ is conditionally on $\boldsymbol{Y}$ normal zero mean:

$$\boldsymbol{\varepsilon}^\flat = \operatorname{diag}(\check{\boldsymbol{Y}})\boldsymbol{w}^\flat \,\big|_{\boldsymbol{Y}} \sim \mathbb{P}^\flat \stackrel{\text{def}}{=} \mathcal{N}(0, \check{\Sigma}).$$

The bootstrap analogue of $\boldsymbol{\xi}_{m,m^\circ} = \mathcal{K}_{m,m^\circ}\boldsymbol{\varepsilon}$ reads

$$\boldsymbol{\xi}^\flat_{m,m^\circ} = \mathcal{K}_{m,m^\circ}\boldsymbol{\varepsilon}^\flat = \mathcal{K}_{m,m^\circ} \operatorname{diag}(\check{\boldsymbol{Y}})\boldsymbol{w}^\flat. \tag{2.18}$$

The idea is to calibrate the SmA procedure under the bootstrap measure $\mathbb{P}^{\flat}$ using $\|\boldsymbol{\xi}^{\flat}_{m,m^\circ}\|$ in place of $\|\boldsymbol{\xi}_{m,m^\circ}\|$. The bootstrap quantiles $z^{\flat}_{m,m^\circ}(t)$ are given by the analogue of (2.10):

$$\mathbb{P}^{\flat}\bigl(\|\boldsymbol{\xi}^{\flat}_{m,m^\circ}\| \geq z^{\flat}_{m,m^\circ}(t)\bigr) = \mathrm{e}^{-t}. \tag{2.19}$$

The use of a continuous distribution for the bootstrap weights $w_i^{\flat}$ allows us to uniquely define the values $z^{\flat}_{m,m^\circ}(t)$. If a discrete distribution of the weights is used, then, as usual, $z^{\flat}_{m,m^\circ}(t)$ is the minimal value for which the probability in the left hand-side of (2.19) does not exceed e^{-t}. The multiplicity correction $q^{\flat}_{m^\circ} = q^{\flat}_{m^\circ}(\mathrm{x})$ is specified by the condition

$$\mathbb{P}^{\flat}\Biggl(\bigcup_{m\in\mathcal{M}^+(m^\circ)} \{\|\boldsymbol{\xi}^{\flat}_{m,m^\circ}\| \geq z^{\flat}_{m,m^\circ}(\mathrm{x}+q^{\flat}_{m^\circ})\}\Biggr) = \mathrm{e}^{-\mathrm{x}}. \tag{2.20}$$

Finally, the bootstrap critical values are fixed by the analogue of (2.12):

$$\mathfrak{z}^{\flat}_{m,m^\circ} \stackrel{\mathrm{def}}{=} z^{\flat}_{m,m^\circ}(\mathrm{x}+q^{\flat}_{m^\circ}) + \beta^{\flat}\sqrt{\mathfrak{p}^{\flat}_{m,m^\circ}}, \tag{2.21}$$

where $\beta^{\flat}$ is a given positive constant and $\mathfrak{p}^{\flat}_{m,m^\circ} = \mathbb{E}^{\flat}\|\boldsymbol{\xi}^{\flat}_{m,m^\circ}\|^2$ is the conditional expectation of $\|\boldsymbol{\xi}^{\flat}_{m,m^\circ}\|^2$ w.r.t. the bootstrap measure:

$$\mathfrak{p}^{\flat}_{m,m^\circ} \stackrel{\mathrm{def}}{=} \mathrm{tr}\{\mathcal{K}^{\top}_{m,m^\circ}\,\mathrm{diag}(\check{\boldsymbol{Y}}\cdot\check{\boldsymbol{Y}})\mathcal{K}_{m,m^\circ}\}.$$

Now we apply the SmA procedure (2.16) with the data-driven critical values $\mathfrak{z}^{\flat}_{m,m^\circ}$ from (2.21).

3. Theoretical properties. This section contains the main theoretical properties of the proposed SmA procedure. We start again from the case of known noise. Then the results are extended to the bootstrap procedure.

3.1. *Known noise.* Let m^* be the oracle choice from (2.9), and let $\widehat{m}$ be the SmA selector from (2.16). Our study focuses on the properties of $\widehat{m}$. As a byproduct, we describe some oracle bounds on the loss of the corresponding adaptive procedure $\widehat{\boldsymbol{\phi}} = \widetilde{\boldsymbol{\phi}}_{\widehat{m}}$. The construction of $\widehat{m}$ ensures that the oracle m^* is accepted with high probability; see (2.14). Therefore, the selector $\widehat{m}$ with probability at least $1-\mathrm{e}^{-\mathrm{x}}$ takes its value in the set

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

of all models in $\mathcal{M}$ not greater than m^*. It remains to check the performance of the method in this region. The next step is to specify a subset of $\mathcal{M}^-$ which contains $\widehat{m}$-values with a high probability. By definition (2.9), m^* is the smallest index for which the bias terms $\|\boldsymbol{b}_{m,m^\circ}\|$ are uniformly bounded by $\beta\mathfrak{p}^{1/2}_{m,m^\circ}$, $m > m^\circ$. Therefore, for each $m^\circ < m^*$, there is at least one $m > m^\circ$ with $\|\boldsymbol{b}_{m,m^\circ}\| > \beta\mathfrak{p}^{1/2}_{m,m^\circ}$.

The next result shows that the test τ_{m,m° based on $\mathbb{T}_{m,m^\circ}$ rejects H_{m,m° with high probability if the condition $\|\boldsymbol{b}_{m,m^\circ}\| \le \beta \mathtt{p}_{m,m^\circ}^{1/2}$ is significantly violated (the "large bias" case). This observation allows us to describe the so-called *zone of insensitivity* $\mathcal{M}_{\mathrm{in}}$, where $\widehat{m}$ concentrates. The results in this subsection hold for a general noise distribution with zero mean and finite variance. In the subsequent subsections, we will then again assume Gaussianity of the errors.

THEOREM 3.1. *For the linear model $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with arbitrary but known distribution of $\boldsymbol{\varepsilon}$, suppose to be given a family of smoothers $\widetilde{\boldsymbol{\phi}}_m = \mathcal{K}_m \boldsymbol{Y}$, $m \in \mathcal{M}$, ordered by their variance due to (2.4). Let $z_{m,m^\circ}(\cdot)$ be the tail function from (2.10) for each pair $m > m^\circ \in \mathcal{M}$. Given $\mathtt{x}$ and β, let $\mathfrak{z}_{m,m^\circ}$ be due to (2.11) and (2.12), and let the oracle m^* be defined in (2.9). Then the property (2.14) is fulfilled for the SmA rule $\widehat{m}$. Let also $\mathcal{M}_{\boldsymbol{b}}^-$ be the subset of $\mathcal{M}^-$ defined by the "large bias" condition*:

$$\mathcal{M}_{\boldsymbol{b}}^- \stackrel{\mathrm{def}}{=} \bigl\{ m \in \mathcal{M}^- : \|\boldsymbol{b}_{m^*,m}\| > \mathfrak{z}_{m^*,m} + z_{m^*,m}(\overline{\mathtt{x}}) \bigr\},$$

where $\overline{\mathtt{x}} \stackrel{\mathrm{def}}{=} \mathtt{x} + \log|\mathcal{M}^-|$. Then it holds with $\mathcal{M}_{\mathrm{in}} \stackrel{\mathrm{def}}{=} \mathcal{M}^- \setminus \mathcal{M}_{\boldsymbol{b}}^-$

$$\mathbb{P}(\widehat{m} \in \mathcal{M}_{\mathrm{in}}) \ge 1 - 2\mathrm{e}^{-\mathtt{x}}.$$

REMARK 3.1. The *set of insensitivity* $\mathcal{M}_{\mathrm{in}} = \mathcal{M}^- \setminus \mathcal{M}_{\boldsymbol{b}}^-$ contains all indices $m^\circ < m^*$ for which the squared bias $\|\boldsymbol{b}_{m,m^\circ}\|^2$ exceeds at some point $m > m^\circ$ the value $\beta^2 \mathtt{p}_{m,m^\circ}$ but not essentially. Therefore, we cannot guarantee that the related test τ_{m,m° is powerful. The worst case setup corresponds to a flat bias profile with $\|\boldsymbol{b}_{m,m^\circ}\| \approx \beta \mathtt{p}_{m,m^\circ}^{1/2}$. Then the set of insensitivity $\mathcal{M}_{\mathrm{in}}$ can coincide with the whole range $\mathcal{M}^-$.

The next result describes the properties of the SmA estimator $\widehat{\boldsymbol{\phi}} = \widetilde{\boldsymbol{\phi}}_{\widehat{m}}$.

THEOREM 3.2. *Under conditions of Theorem 3.1, the SmA estimator $\widehat{\boldsymbol{\phi}} = \widetilde{\boldsymbol{\phi}}_{\widehat{m}}$ satisfies the following bound*:

$$\mathbb{P}\bigl(\|\widehat{\boldsymbol{\phi}} - \widetilde{\boldsymbol{\phi}}_{m^*}\| > \overline{\mathfrak{z}}_{m^*}\bigr) \le 2\mathrm{e}^{-\mathtt{x}}, \tag{3.1}$$

where $\overline{\mathfrak{z}}_{m^*}$ is defined as

$$\overline{\mathfrak{z}}_{m^*} \stackrel{\mathrm{def}}{=} \max_{m \in \mathcal{M}_{\mathrm{in}}} \mathfrak{z}_{m^*,m}. \tag{3.2}$$

This implies the probabilistic oracle bound: *with probability at least $1 - 2\mathrm{e}^{-\mathtt{x}}$*

$$\bigl\|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*\bigr\| \le \bigl\|\widetilde{\boldsymbol{\phi}}_{m^*} - \boldsymbol{\phi}^*\bigr\| + \overline{\mathfrak{z}}_{m^*}. \tag{3.3}$$

REMARK 3.2. The result (3.3) is called the *oracle bound* because it compares the loss of the data-driven selector $\widehat{m}$ and of the oracle choice m^*. The discrepancy $\overline{\mathfrak{z}}_{m^*}$ in (3.1) or (3.3) can be viewed as a price for a data-driven model choice or "payment for adaptation;" cf. Lepski, Mammen and Spokoiny (1997). An interesting feature of the presented result is that not only the oracle quality but also the payment for adaptation depend upon the unknown response $\boldsymbol{f}$ and the corresponding oracle choice m^*. The bound (3.3) is nearly sharp if the value $\overline{\mathfrak{z}}_{m^*}$ is smaller in order than $p_{m^*}^{1/2}$.

REMARK 3.3. The usual Lepski's risk upper bound is very similar to (3.3); cf. Lepski, Mammen and Spokoiny (1997). However, the related "payment for adaptation" $\mathfrak{z}$ is evaluated by rather crude Bonferroni-type arguments for the worst case, and it can be significantly larger than $\overline{\mathfrak{z}}_{m^*}$ from (3.2).

The procedure and the results can be extended to the case of polynomial loss; see Section B in the Supplementary Material (Spokoiny and Willrich (2018)).

3.2. *Analysis of the payment for adaptation* $\overline{\mathfrak{z}}_{m^*}$. We now return to the setting of Gaussian errors ε_i. The benefit of considering the Gaussian case is that each vector $\boldsymbol{\xi}_{m,m^\circ}$ is Gaussian as well, which simplifies the analysis of the tail function $z_{m,m^\circ}(\cdot)$. The bounds can be easily extended to sub-Gaussian errors.

With $\mathbb{V}_m \stackrel{\mathrm{def}}{=} \operatorname{Var}(\widetilde{\boldsymbol{\phi}}_m) = \mathcal{K}_m \operatorname{Var}(\boldsymbol{\varepsilon}) \mathcal{K}_m^\top$, denote for $m^\circ < m$,

$$\mathfrak{p}_m = \operatorname{tr}(\mathbb{V}_m), \qquad \lambda_m = \|\mathbb{V}_m\|_{\mathrm{op}},$$
$$\mathfrak{p}_{m,m^\circ} = \operatorname{tr}(\mathbb{V}_{m,m^\circ}), \qquad \lambda_{m,m^\circ} = \|\mathbb{V}_{m,m^\circ}\|_{\mathrm{op}}.$$

THEOREM 3.3. *Let the conditions of Theorem 3.1 be fulfilled, and let the errors ε_i be normal zero mean. Then the critical values $\mathfrak{z}_{m,m^\circ}$ given by (2.12) satisfy for all pairs $m > m^\circ$ in $\mathcal{M}$*

$$\mathfrak{z}_{m,m^\circ} \le (1+\beta)\sqrt{\mathfrak{p}_{m,m^\circ}} + \sqrt{2\lambda_{m,m^\circ}(\mathrm{x} + \log|\mathcal{M}|)}.$$

Suppose also that

$$\mathfrak{p}_{m^*,m} \le \mathfrak{p}_{m^*}, \qquad \lambda_{m^*,m} \le \lambda_{m^*} \qquad \forall m \in \mathcal{M}^-.$$

Then the value $\overline{\mathfrak{z}}_{m^*}$ follows the bound

$$\overline{\mathfrak{z}}_{m^*} \le (1+\beta)\sqrt{\mathfrak{p}_{m^*}} + \sqrt{2\lambda_{m^*}(\mathrm{x} + \log|\mathcal{M}|)}.$$

REMARK 3.4. The presented results help to understand the relation between the oracle risk $\mathcal{R}_{m^*}$ and the term $\overline{\mathfrak{z}}_{m^*}$. We know that $\mathcal{R}_{m^*} = \|\boldsymbol{b}_{m^*}\|^2 + \mathfrak{p}_{m^*} \ge \mathfrak{p}_{m^*}$. Consider separately two cases: $\mathfrak{p}_{m^*} \gg \lambda_{m^*}$ and $\mathfrak{p}_{m^*} \asymp \lambda_{m^*}$. In the first case which is the typical situation in model selection, it also holds $2\lambda_{m^*}(\mathrm{x} + \log|\mathcal{M}|) \ll \mathfrak{p}_{m^*}$

and the payment for adaptation is not essentially larger than the oracle risk. In fact, for the case with a narrow zone of insensitivity $\mathcal{M}_{\mathrm{in}} = \mathcal{M}^- \setminus \mathcal{M}^-_{\boldsymbol{b}}$, the value $\overline{\mathfrak{z}}_{m^*}$ is much smaller than $\mathtt{p}^{1/2}_{m^*}$; see Section 3.3 for details. The second case $\mathtt{p}_{m^*} \asymp \lambda_{m^*}$ is somewhat extreme and it corresponds to estimation of a linear functional or estimation for severely ill-posed problems; see Section 3.4 below. In this case, the squared payment for adaptation $\overline{\mathfrak{z}}^2_{m^*}$ can be larger than the oracle risk by a factor $\log|\mathcal{M}|$.

3.3. *Application to projection estimation.* This section discusses the case of projection estimation in the linear model $\boldsymbol{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ with homogeneous errors ε_i: $\operatorname{Var}(\varepsilon_i) = \sigma^2$. All the conclusions can be easily extended to heterogeneous errors whose variances are contained in some fixed interval. We also focus on probabilistic loss, the case of polynomial loss can be considered in the same way.

Let us assume that the features in Ψ are ordered and for each $m \in \mathbb{N}$, denote by Ψ_m the $p \times n$ matrix corresponding to the first m features and obtained from Ψ by letting to zero all the entries for the remaining features. The related estimator $\widetilde{\boldsymbol{\theta}}_m = \mathcal{S}_m \boldsymbol{Y}$ is the standard LSE with $\mathcal{S}_m = (\Psi_m \Psi_m^\top)^- \Psi_m$ and the prediction problem with $W = \Psi^\top$ yields $\mathcal{K}_m \boldsymbol{Y} = \Psi^\top \mathcal{S}_m \boldsymbol{Y} = \Pi_m \boldsymbol{Y}$, where $\Pi_m = \Psi_m^\top (\Psi_m \Psi_m^\top)^- \Psi_m$ is the projector in $\mathbb{R}^n$ onto the corresponding m -dimensional subspace. For homogeneous errors ε_i with $\operatorname{Var}(\varepsilon_i) = \sigma^2$, the variance $\mathbb{V}_m = \operatorname{Var}(\Pi_m \boldsymbol{Y})$ satisfies

$$\mathtt{p}_m = \operatorname{tr}\bigl\{\operatorname{Var}(\Pi_m \boldsymbol{Y})\bigr\} = \sigma^2 \operatorname{tr}(\Pi_m) = \sigma^2 m.$$

Moreover, for each pair $m > m^\circ$, it holds

$$\Psi^\top (\widetilde{\boldsymbol{\theta}}_m - \widetilde{\boldsymbol{\theta}}_{m^\circ}) = (\Pi_m - \Pi_{m^\circ}) \boldsymbol{Y}.$$

COROLLARY 3.4. *Consider the problem of projection estimation with homogeneous Gaussian errors ε_i and probabilistic loss. Then $\mathtt{p}_{m,m^\circ} = \sigma^2 (m - m^\circ)$, $\lambda_{m,m^\circ} = \sigma^2$, and*

$$\begin{aligned} \mathfrak{z}_{m,m^\circ} &\le \sigma(1+\beta)\sqrt{m - m^\circ} + \sigma\sqrt{2\mathtt{x} + 2\log|\mathcal{M}|}, \\ \overline{\mathfrak{z}}_{m^*} &\le \sigma(1+\beta)\sqrt{m^*} + \sigma\sqrt{2\mathtt{x} + 2\log|\mathcal{M}|}. \end{aligned} \tag{3.4}$$

REMARK 3.5. The first term in the expression for $\overline{\mathfrak{z}}_{m^*}$ is of order $\sqrt{m^*}$ and it is a leading one provided that the effective dimension m^* is essentially larger than $\log|\mathcal{M}|$. The ordering condition yields that the total number $|\mathcal{M}|$ of considered models is at most n. Moreover, a choice of the set $\mathcal{M}$ in a geometric scale yields that $|\mathcal{M}|$ is only logarithmic in the sample size n; cf. Lepskiĭ (1991), Lepski, Mammen and Spokoiny (1997). Then $\log|\mathcal{M}| \approx \log\log n$ and $\overline{\mathfrak{z}}_{m^*} \approx \sigma\sqrt{m^*}$ for $m^* \gg \log\log n$. For the oracle risk $\mathcal{R}_{m^*}$, it holds $\mathcal{R}_{m^*} = \mathtt{p}_{m^*} + \|\boldsymbol{b}_{m^*}\|^2 \ge \sigma^2 m^*$. Therefore, the payment for adaptation $\overline{\mathfrak{z}}_{m^*}$ is not larger in order than the square root of the oracle risk, and the result of Theorem 3.3 has a surprising corollary: if

the oracle dimension m^* is significantly larger than $\log\log n$, then the data-driven SmA estimator provides nearly the same accuracy as the oracle one.

REMARK 3.6. The payment for adaptation can be drastically reduced in the situations with a narrow zone of insensitivity. Suppose that the "large bias set" $\mathcal{M}_{\boldsymbol{b}}^-$ contains all indices $m \le m^\circ$ for a fixed $m^\circ < m^*$. For instance, this is the case when $\|\boldsymbol{b}_{m^*,m}\|^2 \ge \mathtt{C}\sigma^2(m^* - m + 2\mathtt{x} + 2\log|\mathcal{M}|)$ for some fixed sufficiently large constant $\mathtt{C}$ and all $m \le m^\circ$. Then by (3.4)

$$\bar{\mathfrak{z}}_{m^*} = \max_{m \in \mathcal{M}_{\mathrm{in}}} \mathfrak{z}_{m^*,m} \le \sigma(1+\beta)\sqrt{m^* - m^\circ} + \sigma\sqrt{2\mathtt{x} + 2\log|\mathcal{M}|}.$$

So, if $(m^* - m^\circ)/m^*$ is small, the payment for adaptation is smaller in order than the oracle risk, and the procedure is sharp adaptive. In particular, one can easily see that the self-similarity condition of Giné and Nickl (2010) ensures a rapid growth of the bias when the index m becomes smaller than m^*. This in turn yields a narrow zone of insensitivity, and hence, a sharp adaptive estimation.

REMARK 3.7. The popular Akaike criterion (AIC) defines $\widehat{m}$ as

$$\widehat{m} = \operatorname*{argmin}_m \{\|\boldsymbol{Y} - \Pi_m \boldsymbol{Y}\|^2 + 2\sigma^2 m\}.$$

One can easily see that this rule is equivalent to the SmA rule (2.16) with $\mathfrak{z}^2_{m,m^\circ} = 2\sigma^2(m - m^\circ)$. For this choice, one can prove a risk oracle bound under rather general conditions (see, e.g., Kneip (1994)); however, it does not deliver any information about the behavior of $\widehat{m}$, in particular, it does not guarantee the propagation property (2.14).

Oracle accuracy and asymptotic minimax risk. Here we briefly discuss the relation between the oracle bound (3.3) and minimax rates of estimation in regression with regular design and homogeneous noise. Suppose that the mean response vector $\boldsymbol{f}$ corresponds to the values of a smooth regression function $f(X_i)$ at some regular design points $X_1, \ldots, X_n \in [0,1]$. Let $\psi_1, \ldots, \psi_m, \ldots$, be a set of basis functions on $[0,1]$ like cosine, Demmler–Reinsch, or B-splines basis. We identify the function ψ_m with the vector $\boldsymbol{\psi}_m$ of its values at design points, $\boldsymbol{\psi}_m = (\psi_m(X_1, \ldots, \psi_m(X_n))^\top \in \mathbb{R}^n$. The operator Π_m projects onto the subspace spanned by the first m vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_m$. Then under the standard Sobolev smoothness condition on f, the bias $\boldsymbol{b}_m = \boldsymbol{f} - \Pi_m \boldsymbol{f}$ satisfies $n^{-1}\|\boldsymbol{b}_m\|^2 \le \mathtt{C} m^{-2s}$ and similarly $n^{-1}\|\boldsymbol{b}_{m',m}\|^2 \le \mathtt{C} m^{-2s}$ for $m' > m$. Together with $p_m = \sigma^2 m$, this yields that the conditions (2.9) are fulfilled with any fixed β and $m^* \approx \mathtt{C}(\beta) n^{1/(2s+1)}$ and $n^{-1}\mathcal{R}_{m^*} \le \mathtt{C}(\beta) n^{-2s/(2s+1)}$, where the constant $\mathtt{C}(\beta)$ only depensd on β. This is the optimal accuracy over the class of smooth function of the Sobolev degree s; see, for example, Ibragimov and Has'minskiĭ (1981) or Pinsker (1980). In view of Remark 3.5, the proposed selector ensures

the optimal estimation rate over a Sobolev smoothness class without knowing the parameters of the class up to the additive payment for adaptation $\overline{\mathfrak{z}}_{m^*}$. This easily implies the classical results on adaptive estimation: the SmA estimator is rate-adaptive over a wide range of smoothness classes such of degree s under the constraint $n^{1/(2s+1)} \geq \mathtt{C} \log\log n$.

3.4. *Linear functional estimation.* In this section, we discuss the problem of linear functional estimation. As previously, we assume a family of estimators $\widetilde{\phi}_m = \mathcal{K}_m \boldsymbol{Y}$, $m \in \mathcal{M}$, to be given, where the rank of each $\mathcal{K}_m$ is equal to 1. The ordering condition means that these estimators are ordered by their variance

$$\mathtt{p}_m = \operatorname{Var}(\mathcal{K}_m \boldsymbol{Y}) = \mathcal{K}_m \operatorname{Var}(\boldsymbol{\varepsilon}) \mathcal{K}_m^\top$$

which grows with m. Further, each stochastic component $\xi_{m,m^\circ} = \mathcal{K}_{m,m^\circ} \boldsymbol{\varepsilon}$ is one-dimensional, and it holds for $m > m^\circ$

$$\lambda_{m,m^\circ} = \mathtt{p}_{m,m^\circ} = \mathcal{K}_{m,m^\circ} \operatorname{Var}(\boldsymbol{\varepsilon}) \mathcal{K}_{m,m^\circ}^\top.$$

Note that in the case of Gaussian errors, ξ_{m,m° is also Gaussian: $\xi_{m,m^\circ} \sim \mathcal{N}(0, \mathtt{p}_{m,m^\circ})$. The tail function $z_{m,m^\circ}(\mathtt{x})$ of ξ_{m,m° can be upper-bounded by $\sqrt{2\mathtt{x}\mathtt{p}_{m,m^\circ}}$ yielding

$$\mathfrak{z}_{m,m^\circ} \leq \mathtt{p}_{m,m^\circ}^{1/2} \bigl(\beta + \sqrt{2\mathtt{x} + 2\log|\mathcal{M}|}\bigr), \tag{3.5}$$

$$\overline{\mathfrak{z}}_{m^*} \leq \mathtt{p}_{m^*}^{1/2} \bigl(\beta + \sqrt{2\mathtt{x} + 2\log|\mathcal{M}|}\bigr). \tag{3.6}$$

THEOREM 3.5. *Let the errors ε_i be Gaussian zero mean. Consider a problem of linear functional estimation of $\phi^* = W\boldsymbol{\theta}^*$ by a given family $\widetilde{\phi}_m = \mathcal{K}_m \boldsymbol{Y}$ with* $\operatorname{rank}(\mathcal{K}_m) = 1$, $m \in \mathcal{M}$. *Then the critical values $\mathfrak{z}_{m,m^\circ}$ from* (2.12) *fulfill* (3.5) *and the oracle inequality* (3.3) *holds with the payment for adaptation $\overline{\mathfrak{z}}_{m^*}$ obeying* (3.6).

REMARK 3.8. For the problem of linear functional estimation with probabilistic loss, the squared payment for adaptation $\overline{\mathfrak{z}}_{m^*}^2$ is by a factor $\log|\mathcal{M}|$ larger than the oracle variance $\mathtt{p}_{m^*}$. If $|\mathcal{M}|$ itself is logarithmic in the sample size n, we end up with the extra $\log\log n$—factor in the accuracy of adaptive estimation. This factor appears to be unavoidable; see, for example, Spokoiny and Vial (2009) in the context of estimating a linear functional.

3.5. *Validity of the bootstrap procedure. Conditions.* This and the next sections extend the results obtained for the case of known error distribution to the bootstrap procedure which does not use any information about the noise variance. The main result claims that the bootstrap choice still ensures the condition (2.11) and, therefore, all the obtained results including the oracle bounds, apply for this choice as well; see Theorem 3.7. Moreover, we evaluate the distance between the unknown underlying distribution $\mathbb{Q}$ of the set of random vectors $\boldsymbol{\xi}_{m,m^\circ}$ and their

bootstrap counterpart $\mathbb{Q}^\flat$. The latter is random, however, we show that with high probability, it is close to $\mathbb{Q}$. In what follows, we assume the model (2.1) with a heterogeneous Gaussian noise $\boldsymbol{\varepsilon}$. The results presented below rely on the following quantities.

Design regularity is measured by the value δ_Ψ

$$\delta_\Psi \stackrel{\text{def}}{=} \max_{i=1,\ldots,n} \|S^{-1/2}\Psi_i\|\sigma_i \qquad \text{where } S \stackrel{\text{def}}{=} \sum_{i=1}^n \Psi_i \Psi_i^\top \sigma_i^2; \tag{3.7}$$

Obviously,

$$\sum_{i=1}^n \|S^{-1/2}\Psi_i\|^2 \sigma_i^2 = \operatorname{tr}\biggl(\sum_{i=1}^n S^{-1}\Psi_i\Psi_i^\top\sigma_i^2\biggr) = \operatorname{tr} \boldsymbol{I}_p = p,$$

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

Presmoothing bias for $\boldsymbol{f} = \mathbb{E}\boldsymbol{Y}$ is described by the vector

$$\boldsymbol{B} = \Sigma^{-1/2}(\boldsymbol{f} - \Pi \boldsymbol{f}). \tag{3.8}$$

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

Regularity of the smoothing operator Π is required in Theorem 3.7. Namely, we assume that the rows $\Upsilon_i^\top$ of the matrix $\Upsilon \stackrel{\text{def}}{=} \Sigma^{-1/2}\Pi\Sigma^{1/2}$ fulfill

$$\|\Upsilon_i^\top\| \le \delta_\Pi, \qquad i = 1, \ldots, n. \tag{3.9}$$

This condition is in fact very close to the design regularity condition (3.7). To see this, consider the case of a homogeneous noise with $\Sigma = \sigma^2 \boldsymbol{I}_n$ and $\Pi = \Psi^\top(\Psi\Psi^\top)^{-1}\Psi$. Then $\Upsilon = \Pi$ and (3.7) implies

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

In general, one can expect that δ_Ψ and δ_Π are of the same order $\sqrt{p/n}$.

3.6. *Bootstrap validation.* This section states the main results justifying the proposed bootstrap procedure: the joint distribution $\mathbb{Q}^\flat$ of the bootstrap stochastic components $\boldsymbol{\xi}^\flat_{m,m^\circ}$ for $m > m^\circ$ from (2.18) nicely reproduces the underlying distribution $\mathbb{Q}$ of the $\boldsymbol{\xi}_{m,m^\circ}$'s, and hence, all the probabilistic results obtained in Section 3.1 for known noise continue to apply after bootstrap parameter tuning. The next result presents a bound on the total variation distance $\|\mathbb{Q} - \mathbb{Q}^\flat\|_{\mathrm{TV}}$ between $\mathbb{Q}$ and $\mathbb{Q}^\flat$. As $\mathbb{Q}^\flat$ is a random measure, the result only holds with high probability.

THEOREM 3.6. *Let* $\boldsymbol{Y} = \Psi^\top\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$ *be a Gaussian vector in* $\mathbb{R}^n$ *with independent components,* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$ *for* $\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, *and let also the*

feature matrix Ψ be such that the $p \times p$-matrix $S = \Psi \Sigma \Psi^\top$ is nondegenerated and (3.7) holds. For a given presmoothing operator $\Pi : \mathbb{R}^n \to R^n$, assume (3.9) to be fulfilled with $\Upsilon \stackrel{\text{def}}{=} \Sigma^{-1/2} \Pi \Sigma^{1/2}$. Let $\mathbb{Q} = \mathcal{L}(\boldsymbol{\xi}_{m,m^\circ}, m > m^\circ \in \mathcal{M})$ and let $\mathbb{Q}^\flat$ be the joint conditional distribution of the bootstrap stochastic terms $\boldsymbol{\xi}^\flat_{m,m^\circ}$ for $m > m^\circ \in \mathcal{M}$ given the data $\boldsymbol{Y}$. Then it holds on a random set Ω_n with $\mathbb{P}(\Omega_n) \geq 1 - 6/n$:

$$\|\mathbb{Q} - \mathbb{Q}^\flat\|_{\mathrm{TV}} \leq \frac{1}{2}\sqrt{p}\,\Delta_n, \tag{3.10}$$

$$\begin{aligned} \Delta_n \stackrel{\text{def}}{=} {} & \|\boldsymbol{B}\|^2_\infty + 4\bigl(\delta_\Pi \|\boldsymbol{B}\|_\infty + \delta_\Psi\bigr)\sqrt{\log n} \\ & + 4\bigl(\delta_\Pi + \delta_\Pi^2 + \delta_\Psi^2\bigr)\log n, \end{aligned} \tag{3.11}$$

where the bias $\boldsymbol{B}$ is given by (3.8).

The result (3.10) enables us to control the differences $\mathbb{Q}(A) - \mathbb{Q}^\flat(A)$ for fixed sets A. To justify the propagation property for the bootstrap-based set of critical values $z^\flat_{m,m^\circ}(\mathrm{x} + q^\flat_{m^\circ})$, given according to (2.18), (2.19), and (2.20) with $\check{\boldsymbol{Y}} = \boldsymbol{Y} - \Pi \boldsymbol{Y}$, we also need to take into account the $\boldsymbol{Y}$-dependence of $z^\flat_{m,m^\circ}(\mathrm{x} + q^\flat_{m^\circ})$. This is done by the following theorem.

THEOREM 3.7. *Assume the conditions of Theorem 3.6. Let Δ_n be from (3.11). Then for each $m^\circ \in \mathcal{M}$, it holds on a random set Ω_n with $\mathbb{P}(\Omega_n) \geq 1 - 6/n$:*

$$\begin{aligned} & \Bigl|\mathbb{P}\Bigl(\max_{m > m^\circ}\bigl\{\|\boldsymbol{\xi}_{m,m^\circ}\| - z^\flat_{m,m^\circ}\bigl(\mathrm{x} + q^\flat_{m^\circ}\bigr)\bigr\} \geq 0\Bigr) \\ & \qquad - \mathbb{P}^\flat\Bigl(\max_{m > m^\circ}\bigl\{\|\boldsymbol{\xi}^\flat_{m,m^\circ}\| - z^\flat_{m,m^\circ}\bigl(\mathrm{x} + q^\flat_{m^\circ}\bigr)\bigr\} \geq 0\Bigr)\Bigr| \leq \sqrt{p}\,\Delta_n. \end{aligned} \tag{3.12}$$

By construction, the values $z^\flat_{m,m^\circ}(\mathrm{x} + q^\flat_{m^\circ})$ are selected as minimal ones under the propagation constraint in the bootstrap world. The presented result shows that the use of these data-dependent critical values does not destroy the propagation condition in the real world.

Now we state a bootstrap version of Theorem 3.1. Note that the definition (2.9) of the oracle m^* involves a constant β, and exactly the same constant shows up in the definition (2.12) of the $\mathfrak{z}_{m,m^\circ}$'s for the case of known noise distribution. For the bootstrap procedure, the value $\beta^\flat$ in the definition (2.21) of the $\mathfrak{z}^\flat_{m,m^\circ}$'s has to be slightly larger than β from (2.9):

$$\beta^\flat \geq (1 - \Delta_n)^{-1/2}\beta.$$

If Δ_n is small, then one can fix $\beta^\flat \approx \beta$. Our default choice is again $\beta^\flat = 1$.

THEOREM 3.8. *Assume the conditions of Theorem 3.7. Given* $\mathtt{x}$ *and* $\beta^\flat$, *let the critical values* $\mathfrak{z}^\flat_{m,m^\circ}$ *be given by* (2.21). *If the value* β (*from the definition* (2.9) *of* m^*) *and* $\beta^\flat$ *satisfy* $\beta^\flat \geq (1-\Delta_n)^{-1/2}\beta$, *then*

$$\mathbb{P}\bigl(m^* \text{ is rejected}\bigr) \leq \mathrm{e}^{-\mathtt{x}} + \sqrt{p}\,\Delta_n,$$

and the bootstrap calibrated SmA estimator $\widehat{\boldsymbol{\phi}} = \widetilde{\boldsymbol{\phi}}_{\widehat{m}}$ *satisfies*

$$\mathbb{P}\bigl(\|\widehat{\boldsymbol{\phi}} - \widetilde{\boldsymbol{\phi}}_{m^*}\| > \overline{\mathfrak{z}}^\flat_{m^*}\bigr) \leq \mathrm{e}^{-\mathtt{x}} + 6n^{-1} + \sqrt{p}\,\Delta_n, \tag{3.13}$$

where $\overline{\mathfrak{z}}^\flat_{m^*}$ *satisfies on the set* Ω_n (*from Theorem 3.6*) *the bound*

$$\overline{\mathfrak{z}}^\flat_{m^*} \stackrel{\text{def}}{=} \max_{m \in \mathcal{M}^-} \mathfrak{z}^\flat_{m^*,m} \leq \sqrt{1+\Delta_n}\bigl\{(1+\beta)\sqrt{\mathfrak{p}_{m^*}} + \sqrt{2\lambda_{m^*}\bigl(\mathtt{x} + \log|\mathcal{M}|\bigr)}\bigr\}. \tag{3.14}$$

3.7. *Bootstrap validity and critical dimension.* Now we discuss the sense of the required conditions for bootstrap validity. The obtained results involve the error term $\sqrt{p}\,\Delta_n$ describing the accuracy of the bootstrap approximation. The Gaussian framework allows to reduce the proof of bootstrap validity to the comparison of two Gaussian measures and to get explicit error bounds. If the errors $\boldsymbol{\varepsilon}$ in the original model (2.1) are not Gaussian, the proof of bootstrap validity requires additional tools like high dimensional Gaussian approximation yielding much larger error bounds; cf. Spokoiny and Zhilova (2015).

Our results are only meaningful and the bootstrap approximation is accurate if the value $\sqrt{p}\,\Delta_n$ in (3.10) is small. One easily gets

$$\sqrt{p}\,\Delta_n \leq \mathtt{C}\, p^{1/2}\bigl\{\|\boldsymbol{B}\|^2_\infty + (\delta_\Psi + \delta_\Pi)\log n\bigr\},$$

where $\mathtt{C}$ is a generic notation for an absolute constant. So, the bootstrap approximation is valid if the values $p^{1/2}\delta_\Psi \log n$, $p^{1/2}\delta_\Pi \log n$, $\|\boldsymbol{B}\|^2_\infty p^{1/2}$ are sufficiently small. Now we spell this condition in the typical situation with $\delta_\Psi \asymp \sqrt{p/n}$ and $\delta_\Pi \asymp \sqrt{p/n}$. Then it suffices that the values $pn^{-1/2}\log(n)$ and $\|\boldsymbol{B}\|^2_\infty p^{1/2}$ are small. Suppose that

$$\|\boldsymbol{B}\|_\infty \leq \mathtt{C}\, p^{-s}. \tag{3.15}$$

Such bounds for $\boldsymbol{B} = \boldsymbol{f} - \Pi\boldsymbol{f}$ are often used in the approximation theory when the response vector $\boldsymbol{f}$ corresponds to a Hölder-smooth regression function with the smoothness parameter s observed with noise at design points. So, the bias component does not destroy the bootstrap validity result if p^{1-4s} is small. We summarize that the bootstrap procedure is justified for $s > 1/4$ if $p = p_n \to \infty$ but $p_n n^{-1/2}\log(n) \to 0$ as $n \to \infty$.

COROLLARY 3.9. *Assume the conditions of Theorem 3.7 and let* (3.15) *hold for* $s > 1/4$. *If* $p = p_n$ *fulfill* $p_n n^{-1/2}\log(n) \to 0$ *as* $n \to \infty$, *then the results of Theorem 3.6 and 3.7 apply with* $\Delta_n \to 0$ *as* $n \to \infty$.

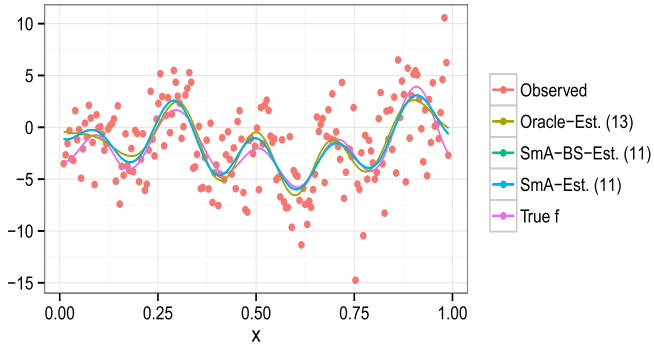

FIG. 1. *True function and observed data with oracle estimator, the known-variance SmA-Estimator (SmA-Est.) and the Bootstrap-SmA-Estimator (SmA-BS-Est.) for inhomogeneous noise. The numbers in parentheses indicate the chosen model dimension.*

4. Simulations. This section illustrates the performance of the proposed procedure by means of simulated examples. We consider a regression model $Y_i = f(X_i) + \varepsilon_i$ for an unknown univariate function on $[0, 1]$ with unknown inhomogeneous Gaussian noise $\boldsymbol{\varepsilon}$. 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 consider a sequence of equidistant design points $(x_i)_{1\le i\le n}$ on $[0, 1]$ and the Fourier basis $\{\psi_j(x)\}_{j=1}^\infty$ to define a sequence of projection estimators where m indicates the truncation dimension of the Fourier basis. The true function is generated by

$$f(x) = c_1\psi_1(x) + \cdots + c_n\psi_n(x),$$

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

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

The noise variances are obtained in the following way: one draws a vector from a normal distribution $\mathcal{N}(2, 0.4\boldsymbol{I}_n)$, takes the square of the coefficients of this vector, puts the coefficients in ascending order and then defines the resulting vector as σ^2. The covariance matrix will then be $\Sigma = \lambda_{\text{int}} \cdot \operatorname{diag}((\sigma_i^2)_{1\le i\le n})$, where λ_{int} governs the intensity of the noise and will be 0.2^2, 0.8^2 and 1.4^2, respectively, for low, medium and high noise level. To generate the noisy observations $n = 200$ will be used. When considering smaller sample sizes, we will take equidistant subsamples of the observations. As a default, the medium noise level will be used for simulations.

For each index m, we build the projection estimate using the first m basis functions. Solving the associated least squares problem gives $\boldsymbol{\theta}_m$ which we will assume to be in $\mathbb{R}^n$ by filling up the vectors of coefficients with zeros.

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 $M = 37$. Our default choice for calibration is $\varkappa = 2$, $\beta^\flat = 1$, and $m^\dagger = 20$.

We start by considering examples for $W = \Psi_n^\top$, that is, the estimation of the whole function vector with prediction loss. One can see in Figure 1 three examples with different intensity of the noise term comparing the Bootstrap-method to the oracle estimator and the known-variance SmA-Method.

Figure 2, top, illustrates the dependence of the SmA choice $\widehat{m}$ on the presmoothing dimension $m^\dagger$ and on the parameter $\beta^\flat$ for different values of the sample size n for one typical noise realization. We see that in the specific example we are considering, the impact of $m^\dagger$ decreases very fast with n. In particular, in the case $n = 200$, no variation in the choice of $\widehat{m}$ is observed in the whole range of $m^\dagger$. The oracles are respectively $m^* = 12$ for $n = 100, 200$ and $m^* = 10$ for $n = 50$. The impact of the parameter $\beta^\flat$ on the estimation results is illustrated on Figure 2, bottom, for $n = 200$. The method appears to be very stable with respect to the choice of $\beta^\flat$.

Figure 3 demonstrates the variability of the ratios $\mathfrak{z}^\flat_{m_1,m_2}/\mathfrak{z}_{m_1,m_2}$ w.r.t. $m^\dagger$. It is remarkable that it is very stable in the range $m^\dagger \geq 12$. Figure 4 shows the distribution of the selected index $\widehat{m}$ after $n_{\text{hist}} = 100$ simulations of the method with the same underlying function f observed with different realizations of the errors. Figure 5 shows the numerical results for the estimation of the first derivative $f'(x)$

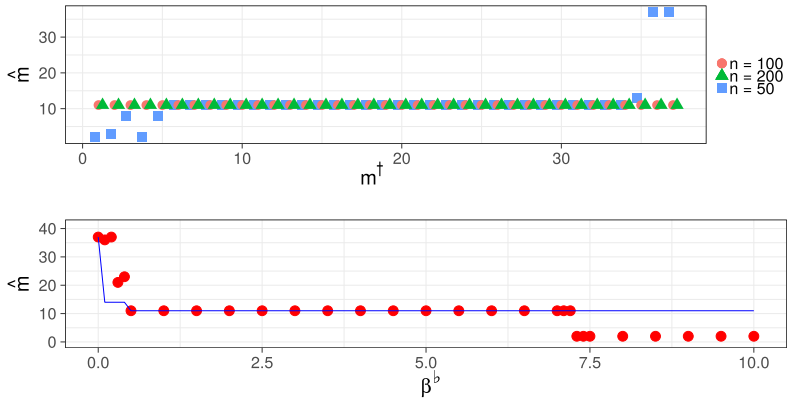

FIG. 2. (*Top*) *The value* $\widehat{m}$ *chosen by the Bootstrap-SmA-Method as a function of the presmoothing dimension* $m^\dagger$ *for* $n = 50, 100, 200$. (*Bottom*) $\widehat{m}$ *as a function of* β *for* $n = 200$. *The blue line indicates the oracle* m^* *according to the definition* (2.9).

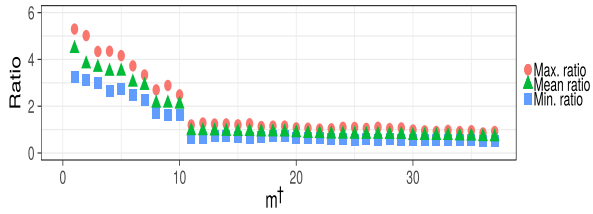

FIG. 3. *Maximal, minimal and mean ratio of the bootstrap and theoretical critical values* $|\mathfrak{z}^{\flat}_{m,m^\circ}/\mathfrak{z}_{m,m^\circ}|^2$ *as a function of* $m^\dagger$.

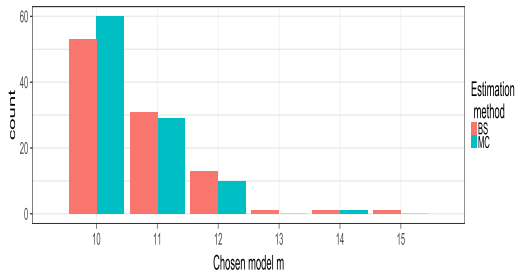

FIG. 4. *Histograms for the selected model by the bootstrap* (*BS*) *and the known-variance method* (*MC*), *simulation size* $n_{hist} = 100$.

in the same model as above. This means taking $W = (\psi_i'(x_j))_{1\le i,j\le n}$. The bootstrap SmA-procedure is well competitive with the procedure based on a known noise structure and the method does a good job of mimicking the oracle in various settings.

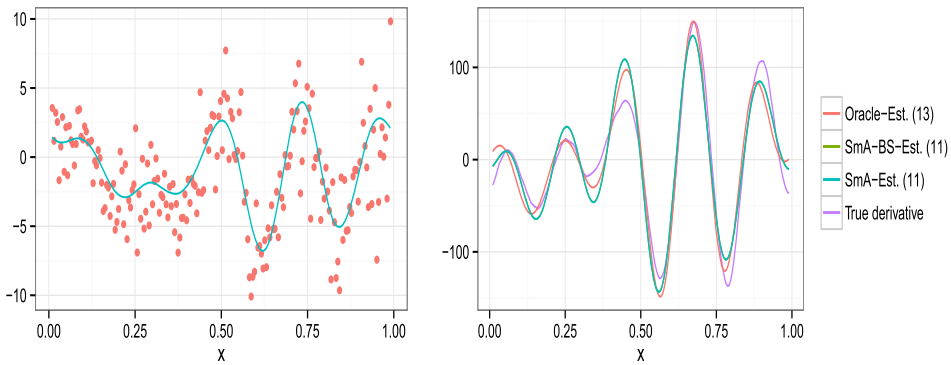

FIG. 5. *Left*: *the true function and the observations*. *Right*: *the true derivative*, *the oracle estimator*, *the known-variance SmA-Estimator* (*SmA-Est.*) *and the Bootstrap-SmA-Estimator* (*SmA-BS-Est.*).

APPENDIX: PROOFS

The Appendix collects the proofs of announced results.

A.1. Proof of Theorems 3.1 and 3.2. The propagation property (2.14) claims that the oracle model m^* will be accepted with high probability. This yields that the selected model is not larger than m^*, that is, $\widehat{m} \le m^*$ with a probability at least $1 - \mathrm{e}^{-\mathtt{x}}$. Below we consider only this event. Let $m \in \mathcal{M}^-$. Acceptance of m requires in particular that $\mathbb{T}_{m^*,m} \le \mathfrak{z}_{m^*,m}$. The representation $\mathbb{T}_{m^*,m} = \|\boldsymbol{b}_{m^*,m} + \boldsymbol{\xi}_{m^*,m}\|$ implies

$$\mathbb{P}(\mathbb{T}_{m^*,m} \le \mathfrak{z}_{m^*,m}) \le \mathbb{P}\bigl(\|\boldsymbol{\xi}_{m^*,m}\| \ge \|\boldsymbol{b}_{m^*,m}\| - \mathfrak{z}_{m^*,m}\bigr).$$

If $m \in \mathcal{M}_{\boldsymbol{b}}^-$, this yields with $\overline{\mathtt{x}} = \mathtt{x} + \log|\mathcal{M}^-|$

$$\begin{aligned}\mathbb{P}(m \text{ is accepted}) &\le \mathbb{P}\bigl(\|\boldsymbol{b}_{m^*,m} + \boldsymbol{\xi}_{m^*,m}\| \le \mathfrak{z}_{m^*,m}\bigr)\\ &\le \mathbb{P}\bigl(\|\boldsymbol{\xi}_{m^*,m}\| \ge z_{m^*,m}(\overline{\mathtt{x}})\bigr) \le \mathrm{e}^{-\overline{\mathtt{x}}}.\end{aligned}$$

This helps to bound the probability of the event $\{\widehat{m} \in \mathcal{M}_{\boldsymbol{b}}^-\}$:

$$\mathbb{P}\bigl(\widehat{m} \in \mathcal{M}_{\boldsymbol{b}}^-\bigr) \le \sum_{m \in \mathcal{M}_{\boldsymbol{b}}^-} \mathbb{P}\bigl(\|\boldsymbol{b}_{m^*,m} + \boldsymbol{\xi}_{m^*,m}\| \le \mathfrak{z}_{m^*,m}\bigr) \le \sum_{m \in \mathcal{M}_{\boldsymbol{b}}^-} \mathrm{e}^{-\overline{\mathtt{x}}} \le \mathrm{e}^{-\mathtt{x}}.$$

Therefore, the probability that the SmA-selector picks up a value $m > m^*$ or $m \in \mathcal{M}_{\boldsymbol{b}}^-$ is very small:

$$\mathbb{P}\bigl(\widehat{m} \in \mathcal{M}^+(m^*) \cup \mathcal{M}_{\boldsymbol{b}}^-\bigr) \le 2\mathrm{e}^{-\mathtt{x}}.$$

It remains to study the case when $\widehat{m} = m$ for some $m \in \mathcal{M}_{\mathrm{in}}$. We can use that this m is accepted, which implies by definition

$$\mathbb{T}_{m^*,m} = \|\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^*}\| \le \mathfrak{z}_{m^*,m}.$$

This yields (3.1). The bound (3.3) now follows by the triangle inequality.

A.2. Proof of Theorem 3.3. Below we use the deviation bound (C.2) for a Gaussian quadratic form from Theorem C.1 in the Supplementary Material (Spokoiny and Willrich (2018)). Note that similar results are available for non-Gaussian quadratic forms under exponential moment conditions; see, for example, Spokoiny (2012). The result (C.2) combined with the Bonferroni correction $q_{m^\circ} = \log|\mathcal{M}^+(m^\circ)| \le \log|\mathcal{M}|$ yields the following upper bound for the critical values $\mathfrak{z}_{m,m^\circ}$:

$$\begin{aligned}\mathfrak{z}_{m,m^\circ} &\le z_{m,m^\circ}(\mathtt{x} + q_{m^\circ}) + \beta \mathfrak{p}_{m,m^\circ}^{1/2}\\ &\le (1+\beta)\sqrt{\mathfrak{p}_{m,m^\circ}} + \sqrt{2\lambda_{m,m^\circ}\bigl(\mathtt{x} + \log|\mathcal{M}^+(m^\circ)|\bigr)}\\ &\le (1+\beta)\sqrt{\mathfrak{p}_{m,m^\circ}} + \sqrt{2\lambda_{m,m^\circ}\bigl(\mathtt{x} + \log|\mathcal{M}|\bigr)}.\end{aligned} \tag{A.1}$$

For the payment for adaptation $\overline{\mathfrak{z}}_{m^*}$, the result (A.1) and the conditions $\mathfrak{p}_{m^*,m} \le \mathfrak{p}_{m^*}$ and $\lambda_{m^*,m} \le \lambda_{m^*}$ imply the required upper bound:

$$\overline{\mathfrak{z}}_{m^*} \le (1+\beta)\sqrt{\mathfrak{p}_{m^*}} + \sqrt{2\lambda_{m^*}(\mathtt{x} + \log|\mathcal{M}|)}.$$

A.3. Proof of Theorem 3.6. Any statement on the use of bootstrap-tuned parameters faces the same fundamental problem: the bootstrap distribution is random and depends on the underlying sample. When we use such bootstrap-based values for the original procedure, we have to account for this dependence. The statement of Theorem 3.6 is even more involved due to the presmoothing step and multiplicity correction (2.20). The proof will be split into a couple of steps. First, we evaluate the effect of the presmoothing bias and variance and reduce the study to an artificial situation where one uses the errors ε_i for resampling in place of the residuals $\check{Y}_i$. Then we compare $\mathbb{Q}$ and $\mathbb{Q}^\flat$ using the Pinsker inequality.

Below we write Ψ in place of Ψ_M, where M is the largest model in the collection. This does not conflict with our general setup, it is implicitly assumed that the largest model coincides with the original one. By p, we denote the corresponding parameter dimension, that is, Ψ is a $p \times n$ matrix. Further, the feature matrix Ψ_m can be written as the product $\Psi_m = \Gamma_m \Psi$, where Γ_m is the projector on the subspace of the feature space $\mathbb{R}^p$ spanned by the features from the model m: $\Gamma_m = \Psi_m \Psi_m^\top (\Psi_m \Psi_m^\top)^-$. This allows to represent each estimator $\widetilde{\boldsymbol{\phi}}_m$ in the form

$$\widetilde{\boldsymbol{\phi}}_m = W\widetilde{\boldsymbol{\theta}}_m = W\mathcal{S}_m \boldsymbol{Y} = W(\Psi_m\Psi_m^\top)^- \Psi_m \boldsymbol{Y} = \mathcal{T}_m \Psi \boldsymbol{Y},$$

$$\mathcal{T}_m \stackrel{\mathrm{def}}{=} W(\Psi_m\Psi_m^\top)^- \Gamma_m = W(\Psi_m\Psi_m^\top)^-.$$

This implies the following representation of the stochastic components $\boldsymbol{\xi}_{m,m^\circ}$ of the difference $\widetilde{\boldsymbol{\phi}}_m - \widetilde{\boldsymbol{\phi}}_{m^\circ}$: with $\nabla = \Psi\boldsymbol{\varepsilon}$, it holds

$$\boldsymbol{\xi}_{m,m^\circ} = \mathcal{T}_{m,m^\circ}\Psi\boldsymbol{\varepsilon} = \mathcal{T}_{m,m^\circ}\nabla, \qquad \mathcal{T}_{m,m^\circ} \stackrel{\mathrm{def}}{=} \mathcal{T}_m - \mathcal{T}_{m^\circ}. \tag{A.2}$$

Thus, each stochastic vector $\boldsymbol{\xi}_{m,m^\circ}$ is a linear function of the vector ∇. A similar representation holds true in the bootstrap world:

$$\boldsymbol{\xi}^\flat_{m,m^\circ} = \mathcal{T}_{m,m^\circ}\Psi\,\mathrm{diag}(\check{\boldsymbol{Y}})\boldsymbol{w}^\flat = \mathcal{T}_{m,m^\circ}\nabla^\flat, \qquad \nabla^\flat \stackrel{\mathrm{def}}{=} \Psi\,\mathrm{diag}(\check{\boldsymbol{Y}})\boldsymbol{w}^\flat. \tag{A.3}$$

Here, the original errors $\boldsymbol{\varepsilon}$ are replaced by their bootstrap surrogates $\boldsymbol{\varepsilon}^\flat = \mathrm{diag}(\check{\boldsymbol{Y}})\boldsymbol{w}^\flat$. Therefore, it suffices to compare the distribution of $\nabla = \Psi\boldsymbol{\varepsilon}$ with the conditional distribution of $\nabla^\flat = \Psi\,\mathrm{diag}(\check{\boldsymbol{Y}})\boldsymbol{w}^\flat$ given $\boldsymbol{Y}$. Then the results will be automatically extended to any deterministic mapping of these two vectors. Normality of the errors $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ implies that $\nabla = \Psi\boldsymbol{\varepsilon}$ is also normal zero mean:

$$\nabla \sim \mathcal{N}(0, S), \qquad S \stackrel{\mathrm{def}}{=} \Psi\Sigma\Psi^\top, \qquad \Sigma = \mathrm{Var}(\boldsymbol{\varepsilon}) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2).$$

Similarly, given the data $\boldsymbol{Y}$, the vector $\nabla^\flat$ is conditionally normal zero mean with the conditional variance

$$S^\flat \stackrel{\mathrm{def}}{=} \mathrm{Var}^\flat(\nabla^\flat) = \Psi\,\mathrm{diag}(\check{Y}_1^2, \ldots, \check{Y}_n^2)\Psi^\top = \Psi\,\mathrm{diag}(\check{\boldsymbol{Y}} \cdot \check{\boldsymbol{Y}})\Psi^\top.$$

Therefore, it remains to compare two p-dimensional Gaussian distributions with different covariance matrices. We apply Pinsker's inequality (see Lemma E.1 in the Supplementary Material, Spokoiny and Willrich (2018)) which only relies on the values $\|\mathcal{B}\|_{\mathrm{op}}$ and $\|\mathcal{B}\|_{\mathrm{Fr}} = \sqrt{\mathrm{tr}(\mathcal{B}^2)}$ for a random $p \times p$ matrix $\mathcal{B}$ given by

$$\mathcal{B} \stackrel{\mathrm{def}}{=} S^{-1/2}(S^\flat - S)S^{-1/2}. \tag{A.4}$$

PROPOSITION A.1. *There is a random set Ω_n with $\mathbb{P}(\Omega_n) \geq 1 - 6/n$ such that it holds on Ω_n with Δ_n given in* (3.11):

$$\|\mathcal{B}\|_{\mathrm{op}} \leq \Delta_n, \qquad \|\mathcal{B}\|_{\mathrm{Fr}} \leq \sqrt{p}\Delta_n, \tag{A.5}$$

PROOF. Define a $p \times n$ matrix $\mathcal{U} = S^{-1/2}\Psi\Sigma^{1/2}$ so that $\mathcal{U}\mathcal{U}^\top = \boldsymbol{I}_p$. We will use the decomposition

$$\Sigma^{-1/2}\check{\boldsymbol{Y}} = \Sigma^{-1/2}(\boldsymbol{Y} - \Pi\boldsymbol{Y}) = \Sigma^{-1/2}(\boldsymbol{\varepsilon} - \Pi\boldsymbol{\varepsilon}) + \Sigma^{-1/2}(\boldsymbol{f} - \Pi\boldsymbol{f}) = \boldsymbol{\eta} + \boldsymbol{B}$$

with $\boldsymbol{\eta} \stackrel{\mathrm{def}}{=} \Sigma^{-1/2}(\boldsymbol{\varepsilon} - \Pi\boldsymbol{\varepsilon})$ and the result follows from Proposition D5 in the Supplementary Material (Spokoiny and Willrich (2018)) with $\delta = \delta_\Psi$ and $\delta_n = \delta_\Pi$ and $y = \log n$ yielding $y + \log n = 2\log n$ and $y + \log p \leq 2\log n$, and thus $\Delta(y) \leq \Delta_n$. □

On the set Ω_n, the claim (3.10) follows from $\|\mathcal{B}\|_{\mathrm{Fr}} \leq \sqrt{p}\Delta_n$ by Lemma E.1 in the Supplementary Material (Spokoiny and Willrich (2018)) with $\boldsymbol{b} = \boldsymbol{b}^\flat = 0$.

A.4. Proof of Theorem 3.7. The result of Theorem 3.6 explains why the known bootstrap distribution can be used as a proxy for the unknown error distribution. However, it cannot be applied directly to (3.12) because the quantities $z^\flat_{m,m^\circ}(\mathtt{x})$ and $q^\flat_{m^\circ}$ are random and depend on the original data. This especially concerns the multiplicity correction $q^\flat_{m^\circ}$ which is based on the joint distribution of the vectors $\boldsymbol{\xi}^\flat_{m,m^\circ}$ from (2.18) and is defined in (2.20). The latter distribution is a Gaussian random measure in the bootstrap world. To cope with the problem of this cross-dependence, we apply geometric arguments to sandwich (with high probability) the random measure from (2.20) in two deterministic measures; see section. The statement of Theorem 3.7 can be derived from Theorem A.1 of the Supplementary Material (Spokoiny and Willrich (2018)) using the bound $\|\mathcal{B}\|_{\mathrm{Fr}} \leq \sqrt{p}\Delta_n$ and $\|\mathcal{B}\|_{\mathrm{op}} \leq \Delta_n$ of Proposition A.1 and conditioning on the set Ω_n.

A.5. Proof of Theorem 3.8. Due to Theorem 3.7, the bootstrap stochastic terms $\boldsymbol{\xi}^\flat_{m,m^\circ}$ nicely mimic (in distribution) their real world counterparts $\boldsymbol{\xi}_{m,m^\circ}$. The SmA procedure also involves the values $\mathfrak{p}_{m,m^\circ} = \mathbb{E}\|\boldsymbol{\xi}_{m,m^\circ}\|^2$, which are unknown and depend on the noise $\boldsymbol{\varepsilon}$. The bootstrap procedure utilizes their versions $\mathfrak{p}^\flat_{m,m^\circ} = \mathbb{E}^\flat\|\boldsymbol{\xi}^\flat_{m,m^\circ}\|^2$. This is justified by the next lemma.

LEMMA A.2. *On the set Ω_n shown in Theorem 3.6, the values for* $\mathfrak{p}^{\flat}_{m,m^\circ} \stackrel{\mathrm{def}}{=} \mathrm{tr}\{\mathrm{Var}^{\flat}(\boldsymbol{\xi}^{\flat}_{m,m^\circ})\}$ *and* $\lambda^{\flat}_{m,m^\circ} \stackrel{\mathrm{def}}{=} \lambda_{\max}\{\mathrm{Var}^{\flat}(\boldsymbol{\xi}^{\flat}_{m,m^\circ})\}$ *for all pairs $m < m^\circ \in \mathcal{M}$ fulfill*

$$\mathfrak{p}_{m,m^\circ}(1-\Delta_n) \le \mathfrak{p}^{\flat}_{m,m^\circ} \le \mathfrak{p}_{m,m^\circ}(1+\Delta_n),$$

$$\lambda_{m,m^\circ}(1-\Delta_n) \le \lambda^{\flat}_{m,m^\circ} \le \lambda_{m,m^\circ}(1+\Delta_n).$$

PROOF. Similar to the proof of Theorem 3.6, we use that $\boldsymbol{\xi}_{m,m^\circ} = \mathcal{T}_{m,m^\circ}\nabla$ and $\boldsymbol{\xi}^{\flat}_{m,m^\circ} = \mathcal{T}_{m,m^\circ}\nabla^{\flat}$ for the same deterministic linear mapping $\mathcal{T}_{m,m^\circ}$; see (A.2) and (A.3). On the set Ω_n the variances $S = \mathrm{Var}(\nabla)$ and $S^{\flat} = \mathrm{Var}(\nabla^{\flat})$ are related by (A.5) for $\mathcal{B}$ from (A.4). This easily implies

$$(1-\Delta_n)S \le S^{\flat} \le (1+\Delta_n)S$$

and thus, the desired bounds follow. □

The relation $\beta^{\flat} \ge (1-\Delta_n)^{-1/2}\beta$ helps to bound on the set Ω_n

$$\|\boldsymbol{b}_{m,m^\circ}\| \le \beta\sqrt{\mathfrak{p}_{m,m^\circ}} \le \beta^{\flat}\sqrt{\mathfrak{p}^{\flat}_{m,m^\circ}}$$

and one can upper bound the probability of the event $\{m^*$ is rejected$\}$ similar to the case of known noise distribution. The oracle inequality (3.13) follows from the acceptance rule under the conditions that $\widehat{m} \le m^*$ and $\widehat{m}$ is accepted; cf. the proof of Theorem 3.2. The bound (3.14) follows from Lemma A.2 and arguments of Theorem 3.3 applied to the bootstrap quantities $\mathfrak{z}^{\flat}_{m,m^\circ}$.

Acknowledgments. The authors are extremely grateful to two anonymous referees whose suggestions led to a significant improvement of the papers.

SUPPLEMENTARY MATERIAL

Some auxiliary results (DOI: 10.1214/18-AOS1717SUPP; .pdf). The supplement collects some useful technical facts and extensions.

REFERENCES

ARLOT, S. (2009). Model selection by resampling penalization. *Electron. J. Stat.* **3** 557–624. MR2519533

BARAUD, Y., HUET, S. and LAURENT, B. (2003). Adaptive tests of linear hypotheses by model selection. *Ann. Statist.* **31** 225–251. MR1962505

BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028

BERAN, R. (1986). Discussion: Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14** 1295–1298.

BIRGÉ, L. (2001). An alternative point of view on Lepski's method. In *State of the Art in Probability and Statistics* (*Leiden*, 1999). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **36** 113–133. IMS, Beachwood, OH. MR1836557

BIRGÉ, L. and MASSART, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138** 33–73. MR2288064

CAVALIER, L. and GOLUBEV, Y. (2006). Risk hull method and regularization by projections of ill-posed inverse problems. *Ann. Statist.* **34** 1653–1677. MR2283712

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.* **42** 1787–1818. MR3262468

DALALYAN, A. S. and SALMON, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.* **40** 2327–2355. MR3059085

GACH, F., NICKL, R. and SPOKOINY, V. (2013). Spatially adaptive density estimation by localised Haar projections. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 900–914. MR3112439

GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170. MR2604707

GOEMAN, J. J. and SOLARI, A. (2010). The sequential rejection principle of familywise error control. *Ann. Statist.* **38** 3782–3810. MR2766868

GOLDENSHLUGER, A. (2009). A universal procedure for aggregating estimators. *Ann. Statist.* **37** 542–568. MR2488362

HÄRDLE, W. and MAMMEN, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21** 1926–1947. MR1245774

IBRAGIMOV, I. A. and HAS'MINSKIĬ, R. Z. (1981). *Statistical Estimation*: *Asymptotic Theory*. *Applications of Mathematics* **16**. Springer, New York. MR0620321

KNEIP, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22** 835–866. MR1292543

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. *Ann. Statist.* **25** 929–947. MR1447734

LEPSKI, O. V. and SPOKOINY, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* **25** 2512–2546. MR1604408

LEPSKIĬ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatn. Primen.* **35** 459–470. MR1091202

LEPSKIĬ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatn. Primen.* **36** 645–659. MR1147167

LEPSKIĬ, O. V. (1992). Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates. *Teor. Veroyatn. Primen.* **37** 468–481. MR1214353

MAMMEN, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.* **21** 255–285. MR1212176

MARCUS, R., PERITZ, E. and GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660. MR0468056

MASSART, P. (2007). *Concentration Inequalities and Model Selection*. *Lecture Notes in Math.* **1896**. Springer, Berlin. MR2319879

PINSKER, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Inf. Transm.* **16** 52–68. MR0624591

ROMANO, J. P. and WOLF, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica* **73** 1237–1282. MR2149247

SPOKOINY, V. G. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24** 2477–2498. MR1425962

SPOKOINY, V. (2012). Parametric estimation. Finite sample theory. *Ann. Statist.* **40** 2877–2909. MR3097963

SPOKOINY, V. and VIAL, C. (2009). Parameter tuning in pointwise adaptation using a propagation approach. *Ann. Statist.* **37** 2783–2807. MR2541447

SPOKOINY, V., WANG, W. and HÄRDLE, W. K. (2013). Local quantile regression. *J. Statist. Plann. Inference* **143** 1109–1129. MR3049611

SPOKOINY, V. and WILLRICH, N. (2019). Supplement to "Bootstrap tuning in Gaussian ordered model selection." DOI:10.1214/18-AOS1717SUPP.

SPOKOINY, V. and ZHILOVA, M. (2015). Bootstrap confidence sets under model misspecification. *Ann. Statist.* **43** 2653–2675. MR3405607

WU, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14** 1261–1350. MR0868303

WEIERSTRASS-INSTITUTE BERLIN
MOHRENSTR. 39
10117 BERLIN
GERMANY
E-MAIL: spokoiny@wias-berlin.de
willrich@wias-berlin.de