Submitted to the Annals of Statistics

ADAPTIVE ESTIMATION IN A LINEAR INVERSE PROBLEM.

By Vladimir Spokoiny and Céline Vial

Weierstrass-Institute and Humboldt University Berlin,
Modal'X and Université Européenne de Bretagne

Abstract

This paper discusses the problem of estimating a linear functional in a linear inverse problem. We consider an adaptive procedure originated from Lepski (1990) which selects in a data-driven way one estimate out of a class of the given estimates ordered by their variability. The main problem with using this and similar procedures is that the question of selecting the involved tuning parameters was not carefully addressed. At the same time, the numerical results indicate that a careful choice of the parameters of the procedure is extremely important for getting the reasonable quality of estimation. The main contribution of this paper is the new approach for choosing the parameters of the procedure by providing the prescribed behavior of the resulting estimate in the simple parametric situation. We establish a non-asymptotical "oracle" bound which shows that the estimation risk is, up to a logarithmic multiplier, equal to the risk of the "oracle" estimate which is optimally selected for the given family. A numerical study demonstrates the nice performance of the resulting procedure in a number of simulated examples.

1. Introduction. This paper discusses the problem of statistical estimation in a linear inverse problem. Such problems are usually considered as more complex than the usual nonparametric regression estimation due to the poor rate of estimation. Moreover, the difficulty which is usually associated with the attained estimation accuracy increases with the degree of illposedness. The origin for the poor rate in inverse problem estimation is mostly due to a bad quality of inversion of the underlying compact operator.

Similarly to the standard regression set-up, the overall error of estimation in an inverse problem is obtained as a sum of the error of approximation and the stochastic error. The approximation error is typically controlled by

AMS 2000 subject classifications: Primary 62G05, 62G05; secondary 62G10, 62G10
Keywords and phrases: linear functional, inverse problem, propagation, oracle

the so called "source condition" and this error decreases with the size (dimension) of the approximating model. On the contrary, the stochastic error rapidly increases with the size of the approximating model and the optimal accuracy is obtained by selecting the smoothness/regularization parameter to balance these two errors. In this paper we focus on the statistical analysis. This means that the method of approximation or regularization is selected in advance and one only has to adjust the degree of regularization from the data. The efficiency of any adaptive (data driven) method can be measured by the ratio of the risk of the proposed method to the "oracle" risk which corresponds to the optimal choice of the regularization parameter for the model at hand. One message of this note is that this statistical part of the linear inverse problem is actually not harder than in the classical nonparametric inference. Moreover, in the inverse problem set-up it is typically easier to do a statistical adaption because the likelihood profile is not so flat as in the classical nonparametric regression.

Below we consider one method of adaptive nonparametric estimation in the linear inverse problem which originates from Lepski (1990). Similar estimation procedures in context of linear inverse problem can be found in Goldenshluger (1999), Goldenshluger and Pereversev (2000), Cavalier, Golubev, Picard and Tsybakov (2002), Tsybakov (2000), Cavalier and Tsybakov (2001,2002), among other. A common drawback of all these proposals is that they involve some other parameter(s) like a threshold whose choice in practical applications is not really addressed. In some sense, the original problem of selecting one (regularization) parameter is simply transferred into another problem of selecting the threshold. The theoretical results claiming the optimal minimax rate of estimation, however, have been established under the condition that the threshold is sufficiently large and they tell nothing if this condition is not fulfilled. At the same time, the experience in practical implementations of the adaptive methods strongly support taking rather small thresholds. Applying a large threshold typically leads to a conservative procedure and oversmoothing effects. In this sense, one can say that there is some critical gap between the theory and practical applications.

Our paper aims at developing a "constructive" theory for this problem which closes the gap between theory and practical implementation and explains in details how the procedure can be implemented in practical situations to deliver reasonable results without any further adjustment. We offer a new approach for selecting the tuning parameters of the method based on the so called "propagation" condition which postulates the desirable performance of the method in the simple parametric situation, see Section 2.1 for a detailed explanation. The idea is similar to the problem of hypothesis

testing for which the critical value of a test is selected by bounding the first kind error probability under the null hypothesis. Note that this approach can be directly applied to many other procedures including local model selection, stagewise aggregation, local change point analysis which are studied in details in Spokoiny (2008) in a much more general set-up.

Golubev (2004) proposed another "risk envelope" approach to select the threshold for a special sequence space model and a particular linear functional. We consider this example in Section 1.3. The common point between Golubev (2004) and our proposal is the selection of the parameters of the method by a Monte Carlo simulation from the model with zero response. However, the motivation and theoretical analysis for our study is quite different from the one in Golubev (2004). Recently Cavalier and Golubev (2006) extended developed a similar "risk hull" approach for the choice of the penalization for estimation of the whole parameter vector.

Theoretical properties of the proposed method are presented in Section 3. The main result states the "oracle" property of the proposed estimate: the risk of the adaptive estimate is within a log-multiple as small as the risk of the "oracle" estimate for the given model. The results are established in the precise nonasymptotic way under mild regularity conditions. Our simulation study in Section 4 confirms a nice finite sample performance of the procedure for a rather big class of different models and problems.

1.1. *Model and problem.* Consider a general set-up of a linear inverse problem when the observed data Y from a Hilbert space $\mathcal{H}_Y$ are modelled by a linear operator equation

$$Y = AX + \varepsilon \tag{1.1}$$

where X is the unknown parameter vector from some Hilbert space $\mathcal{H}_X$, $A : \mathcal{H}_X \to \mathcal{H}_Y$ is a linear operator, and ε is a random Gaussian noise in $\mathcal{H}_Y$ with the known correlation structure given by the covariance operator Σ. The goal is to estimate a linear functional $\theta = \theta(X)$ which can be represented in the form $\langle \vartheta, X \rangle$ for some known element $\vartheta \in \mathcal{H}_X$. Some examples of this scheme are given in the next sections. A naive estimation approach is based on the explicit least square solution of the problem (1.1):

$$\widetilde{\theta} = \langle \vartheta, (A^*A)^- A^* Y \rangle = \langle A(A^*A)^- \vartheta, Y \rangle = \langle \phi, Y \rangle$$

where A^* is the conjugate operator to A, C^- means a pseudo-inverse of C and $\phi = A(A^*A)^- \vartheta$. However, this approach cannot be efficiently applied if A is a compact operator because the inverse of A^*A does not exists or is an unbounded operator. One can regularize the problem if some

additional information about smoothness of the element X is available. This allows to replace $(A^*A)^-$ by its regularization $g_\alpha(A^*A)$ where g_α means some regularized inversion and α is the corresponding parameters. See, e.g., Goldenshluger and Pereversev (1999) for typical examples. The quality of estimation heavily depends on the choice of the regularization parameter α and its choice is a challenging problem. Usually one fixes a finite ordered set of values $\alpha_1 < \alpha_2 < \ldots < \alpha_K$ and considers the corresponding estimates

$$\widetilde{\theta}_k = \langle \phi_k, Y \rangle, \qquad \phi_k = A g_{\alpha_k}(A^*A)\vartheta.$$

Now the original problem can be reformulated as follows: given a set of estimates $\widetilde{\theta}_k$ for known vectors ϕ_k, build an estimate $\widehat{\theta}$ of the functional θ which performs nearly as good as the best in this family. We present one practical example for the considered set-up and one special problem considered in Golubev (2004). More examples include positron emission tomography problem, Cavalier (2001), functional data analysis, Cai and Hall (2006), among many others.

1.2. *Example: Option pricing.* Let S_t be a stock process for which the market of the corresponding vanilla (call) type options is liquid (heavily traded). The problem of pricing an option with another pay-off function $c(y,T)$ can be naturally treated as a linear inverse problem. Indeed, if $p_T(\cdot)$ means the state price density as a function of the strike price K and the time to maturity T, then the (discounted) option price can be computed as the linear functional $\int c(y,T)p(y,T)dy$, see e.g. Ait-Sahalia and Lo (1998) and references therein. The same is applied to the vanilla option where the pay-off function is $c(K,T) = (S_T - K)_+$. Therefore, the observed vanilla option prices $C_1, \ldots, C_n$ for the strikes $K_1, \ldots, K_n$ and the fixed time to maturity T can be described by the model

$$C_i = \int (y - K_i)_+ \, p(y,T)dy + \varepsilon_i$$

with the individual error ε_i.

Here X means the state price density $p(\cdot,T)$ while A maps it into the finite dimensional space $\mathbb{R}^n$: AX is the vector of integrals $\int_y (y - K_i)_+ \, p(y,T)dy$. The objective functional is $\int c(y,T)\, p(y,T)dy$. This problem becomes ill-posed if the pay-off $c(y,T)$ is positive for $y < 0$ or for y which belongs to the out-of-the-money regions where no or only very few options are traded.

1.3. *Example: "sequence space" model.* We consider the statistical problem when the observations $y_1, \ldots, y_M$ follow the "sequence space" equation

$$y_i = \mu_i + \sigma_i \varepsilon_i \,, \qquad i = 1, \ldots, M, \tag{1.2}$$

where ε_i are independent standard normal and the standard deviations σ_i are known while the mean values μ_i are unknown. This "sequence space" model is prototypical for many realistic models like regression or statistical inverse problems, see e.g. Cavalier and Tsybakov (2001,2002) for examples and details. The variances σ_i^2 are usually constant for the regression set-up or grow with i for ill-posed inverse problems.

One particular problem in this set-up can be to estimate the sum

$$\theta = \mu_1 + \ldots + \mu_M \,,$$

where M can be equal to infinity. If μ_i are the Fourier coefficients of some function $f(\cdot)$ and $M = \infty$, then θ means the value $f(0)$.

The "naive" estimate $\widetilde{\theta} = \sum_{i=1}^{M} y_i$, even for a finite M, has a very large variance $\sum_{i=1}^{M} \sigma_i^2$ and hence, can be highly inefficient. The "smoothing" idea leads to the set of the "spectral cut-off" estimates

$$\widetilde{\theta}_k = \langle \phi_k, X \rangle = \sum_{i=1}^{m_k} y_i \,,$$

where $\phi_k = (1, \ldots, 1, 0, \ldots, 0)$ is the vector with the first m_k entries equal to one and the others equal to zero, while m_k is a fixed decreasing sequence of indices $M \ge m_1 > m_2 > \ldots > m_K \ge 1$.

One can easily compute for $k = 1, \ldots, K$

$$\theta_k \stackrel{\mathrm{def}}{=} \boldsymbol{E}\, \widetilde{\theta}_k = \sum_{i=1}^{m_k} \mu_i \,, \qquad v_k \stackrel{\mathrm{def}}{=} \operatorname{Var} \widetilde{\theta}_k = \sum_{i=1}^{m_k} \sigma_i^2 \,,$$

The major difficulty in applying the "smoothing" approach is the proper choice of the parameter k. Small values of k lead to a huge variance v_k of the estimate $\widetilde{\theta}_k$ while large k-values can result in a big bias $b_k = \theta - \theta_k = \sum_{i=m_k+1}^{M} \mu_i$. The "oracle" choice balances the approximation and stochastic errors. However, this "ideal" choice assumes that the bias (the approximation error) is known. The problem we consider in this paper is to develop an adaptive (data-driven) choice which mimics the "oracle" and achieves the best possible performance among the set of estimates $\widetilde{\theta}_k$.

1.4. *Some properties of the estimates* $\widetilde{\theta}_k$. The definition of the estimate $\widetilde{\theta}_k = \langle \phi_k, Y \rangle$ and the model equation (1.1) yield the decomposition

$$\widetilde{\theta}_k = \langle \phi_k, AX \rangle + \langle \phi_k, \varepsilon \rangle = \theta_k + \xi_k \, .$$

The next properties of $\widetilde{\theta}_k$ are direct corollaries of this decomposition.

THEOREM 1.1. *It holds for any* $k \le K$

$$\begin{aligned} \boldsymbol{E}\widetilde{\theta}_k &= \theta_k \, , \\ \operatorname{Var} \widetilde{\theta}_k &= \phi_k^\top \Sigma \phi_k \, . \end{aligned}$$

Moreover, $\widetilde{\theta}_k - \theta_k = \langle \phi_k, \varepsilon \rangle = \xi_k$ *is a Gaussian zero mean r.v. with the variance* $v_k = \operatorname{Var} \widetilde{\theta}_k = \phi_k^\top \Sigma \phi_k$ *satisfying for any* $r > 0$ *and any* $\lambda < 1/2$

$$\begin{aligned} \boldsymbol{E} \bigl| v_k^{-1} \bigl(\widetilde{\theta}_k - \theta_k \bigr)^2 \bigr|^r &= \mathfrak{c}_r \, , \\ \boldsymbol{E} \exp\bigl\{ \lambda v_k^{-1} \bigl(\widetilde{\theta}_k - \theta_k \bigr)^2 \bigr\} &= (1 - 2\lambda)^{-1/2} \end{aligned}$$

where $\mathfrak{c}_r = \boldsymbol{E}|\xi|^{2r}$ *and* ξ *is standard normal.*

Due to this result, $\widetilde{\theta}_k$ is a good estimate of θ if the "bias" $\bigl|\theta_k - \theta\bigr|$ is sufficiently small. In particular, in the "no bias" situation $\theta_k = \theta$ the estimate $\widetilde{\theta}_k$ leads to the accuracy of order $v_k^{1/2}$ and one can build confidence intervals for the parameter θ_k in the form

$$\mathcal{E}_k(\mathfrak{z}) = \bigl\{ u : v_k^{-1} \bigl(\widetilde{\theta}_k - u \bigr)^2 \le \mathfrak{z} \bigr\}. \tag{1.3}$$

If $\mathfrak{z}$ is sufficiently large then the result of Theorem 1.1 ensures that $\mathcal{E}_k(\mathfrak{z})$ contains θ_k with a high probability.

2. Description of the method. This section presents the considered adaptive estimation procedure. Our starting point is the given family of estimates $\widetilde{\theta}_k$ for $k = 1, \ldots, K$ ordered by their variability so that the variance v_k of $\widetilde{\theta}_k$ decreases with k. We aim to select a data-driven index $\widehat{k}$ or equivalently the estimate $\widehat{\theta} = \widetilde{\theta}_{\widehat{k}}$ which minimizes the corresponding estimation risk. The method we apply originates from Lepski (1990). We however, consider a slightly different interpretation of the procedure which is based on the multiple testing idea. Similar and more general ideas lead to a general local model selection procedure which applies for a broad class of nonparametric models and studied in details in Spokoiny (2008).

For a given sequence of estimates $\widetilde{\theta}_k = \langle \phi_k, X \rangle$ consider the sequence of nested hypothesis $H_k : \theta_1 = \ldots = \theta_k = \theta$. The procedure is sequential: we

start with $k = 2$ and at every step k the hypothesis H_k is tested against $H_1, \ldots, H_{k-1}$. If H_k is not rejected then we continue with the next larger k. The final estimate corresponds to the latest accepted hypothesis. For testing H_k against H_l with $l < k$, we check that the new estimate $\widetilde{\theta}_k$ belongs to the confidence intervals built on the base of $\widetilde{\theta}_l$. More precisely, we apply the test statistics:

$$T_{lk} = (\widetilde{\theta}_l - \widetilde{\theta}_k)^2 / v_l\,, \qquad l < k,$$

where v_l is the variance of $\widetilde{\theta}_l$. Big values of T_{lk} indicate a significant difference between the estimates $\widetilde{\theta}_l$ and $\widetilde{\theta}_k$. Due to the definition (1.3), the event $A_{lk} = \{T_{lk} \le \mathfrak{z}_l\}$ means that $\widetilde{\theta}_k$ belongs to the confidence set $\mathcal{E}_l(\mathfrak{z}_l)$ based on $\widetilde{\theta}_l$. The estimate $\widetilde{\theta}_k$ (or the hypothesis H_k) is accepted if H_{k-1} was accepted and $T_{lk} \le \mathfrak{z}_l$ for all $l < k$, that is, the new estimate $\widetilde{\theta}_k$ belongs to the intersection of all the confidence intervals $\mathcal{E}_l(\mathfrak{z}_l)$ built on the previous steps of the procedure. The formal definition is given by

$$\widehat{k} = \max\{k \le K : \quad T^*_{lk} \le \mathfrak{z}_l \quad l = 1, \ldots, k-1\}, \qquad T^*_{lk} = \max_{l<j\le k} T_{lj}\,.$$

Here the "critical values" $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ are the parameters of the procedure. Their choice is discussed in Section 2.1.

The random index $\widehat{k}$ means the largest accepted k. The adaptive estimate $\widehat{\theta}$ is $\widetilde{\theta}_{\widehat{k}}$:

$$\widehat{\theta} = \widetilde{\theta}_{\widehat{k}}\,.$$

We also define the adaptive estimate $\widehat{\theta}_k$ as the latest accepted after the first k steps:

$$\widehat{\theta}_k = \widetilde{\theta}_{\min\{\widehat{k},k\}}\,.$$

The described procedure involves $K - 1$ parameters and their automatic choice is ultimately required for practical applications of the method. Our next step is the procedure for an automatic selection of the critical values $\mathfrak{z}_k$.

2.1. *Choice of the critical values $\mathfrak{z}_k$ using a "propagation condition".* The critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ are selected by the reasoning similar to the standard approach of hypothesis testing theory: to provide the prescribed performance of the procedure under the simplest (null) hypothesis. In the considered set-up, the null means $X \equiv 0$. In this case it is natural to expect that the estimate $\widehat{\theta}_k$ coming out of the first steps of the procedure until the index k is close to the nonadaptive counterpart $\widetilde{\theta}_k$. This particularly means

that the probability of rejecting one of the estimates $\widetilde{\theta}_2, \ldots, \widetilde{\theta}_k$ under the null hypothesis should be very small.

To give a precise definition we need to specify a loss function. Suppose that the risk of estimation for an estimate $\widehat{\theta}$ of θ is measured by $\boldsymbol{E}|\widehat{\theta}-\theta|^{2r}$ for some $r > 0$. Under the null hypothesis $X \equiv 0$, every estimate $\widetilde{\theta}_k$ fulfills $\widetilde{\theta}_k = \langle \phi_k, \varepsilon \rangle$ and hence, it is a zero mean normal variable with the variance v_k. Therefore,

$$\boldsymbol{E}_0 |v_k^{-1} (\widetilde{\theta}_k - \theta)^2|^r = \mathfrak{c}_r$$

where $\mathfrak{c}_r = E|\xi|^{2r}$ and ξ is standard normal. We require that the parameters $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ of the procedure are selected in such a way that

$$\boldsymbol{E}_0 |v_k^{-1} (\widehat{\theta}_k - \widetilde{\theta}_k)^2|^r \le \alpha \mathfrak{c}_r, \qquad k = 2, \ldots, K. \tag{2.1}$$

Here α is the preselected constant which is similar to the confidence level of a testing procedure. This gives us $K-1$ conditions to fix $K-1$ parameters. As in the testing problem, we are interested to select the critical values as small as possible under the constraint (2.1). Note that the choice of the indicator loss function $\mathbf{1}(\widehat{\theta}_k \neq \widetilde{\theta}_k)$ would lead to the usual error of the first kind for the multiple testing procedure. We select another stronger loss function because it is better assigned for the purpose of statistical estimation problem.

Our definition still involves two parameters α and r. It is important to mention that their choice is subjective and there is no way for an automatic selection. A proper choice of the power r for the loss function as well as the "confidence level" α depends on the particular application and on the additional subjective requirements to the procedure. Taking a large r and small α would result in an increase of the critical values and therefore, improves the performance of the method in the parametric situation at cost of some loss of sensitivity to deviations from the parametric situation. This behaviour is analogous to the hypothesis testing problem where a small α reduces the first kind error at costs of the test power. Theorem 3.1 presents some upper bounds for the critical values $\mathfrak{z}_k$ as functions of α and r in the form $a_0 + a_1 \log \alpha^{-1} + a_2 r(K-k)$ with some coefficients a_0, a_1 and a_2. We see that these bounds linearly depend on r and on $\log \alpha^{-1}$. For our examples, we apply a relatively small value $r = 1/2$. We also apply $\alpha = 1$ although the other values in the range $[0.5, 1]$ lead to very similar results. It is worth mentioning that both the procedure and the theoretical study apply and lead to reasonable results whatever r and α are. This makes a striking difference with many other proposals, see the references in the

introduction, for selecting the tuning parameter(s). Typically one requires that the critical value (threshold) $\mathfrak{z}$ is sufficiently large and the theory is only valid under this constraint.

The set of conditions (2.1) do not directly define the critical values $\mathfrak{z}_k$. We present below one sequential method for fixing $\mathfrak{z}_k$ one after another starting from $\mathfrak{z}_1$. The idea is to provide that the relative impact of each $\mathfrak{z}_k$ in the total risk in (2.1) is the same for every $k \le K-1$. We start with $\mathfrak{z}_1$ and set $\mathfrak{z}_2 = \ldots = \mathfrak{z}_{K-1} = \infty$. This effectively means that every new estimate $\widetilde{\theta}_k$ is only compared with $\widetilde{\theta}_1$. We run the procedure with such critical values. The resulting adaptive estimate after step k is denoted by $\widehat{\theta}_k(\mathfrak{z}_1)$. We select $\mathfrak{z}_1$ as the minimal value providing that

$$\boldsymbol{E}_0 |v_k^{-1}\{\widehat{\theta}_k(\mathfrak{z}_1) - \widetilde{\theta}_k\}^2|^r \le \alpha \mathfrak{c}_r/(K-1), \qquad k = 2, \ldots, K. \tag{2.2}$$

Such a value exists because the choice $\mathfrak{z}_1 = \infty$ leads to $\widehat{\theta}_k = \widetilde{\theta}_k$ for all k.

Similarly, we specify $\mathfrak{z}_2$ by considering the situation with the previously fixed $\mathfrak{z}_1$, some finite $\mathfrak{z}_2$ and all the remaining critical values equal to infinity, and so on. For the formal definition, suppose that $\mathfrak{z}_1, \ldots, \mathfrak{z}_{j-1}$ have been already fixed for some $j > 1$ and define for any $\mathfrak{z}_j$ the adaptive estimates $\widehat{\theta}_k(\mathfrak{z}_1, \ldots, \mathfrak{z}_j)$ for $k > j$ which come out of the procedure with the critical value $\mathfrak{z}_1, \ldots, \mathfrak{z}_j, \infty, \ldots, \infty$. We select $\mathfrak{z}_j$ as the minimal value providing that

$$\boldsymbol{E}_0 |v_k^{-1}\{\widehat{\theta}_k(\mathfrak{z}_1, \ldots, \mathfrak{z}_j) - \widetilde{\theta}_k\}^2|^r \le j\alpha \mathfrak{c}_r/(K-1), \; k = j+1, \ldots, K. \tag{2.3}$$

Such a value exists because the choice $\mathfrak{z}_j = \infty$ leads to $\widehat{\theta}_k(\mathfrak{z}_1, \ldots, \mathfrak{z}_j) = \widehat{\theta}_k(\mathfrak{z}_1, \ldots, \mathfrak{z}_{j-1})$ and even a stronger condition has been already checked at the previous step.

The condition (2.2) describes the impact of the first critical value in the risk (2.1) while (2.3) describes the accumulated impact of the first j critical values. The factor $j/(K-1)$ in the right hand side of (2.3) is chosen to ensure that every $\mathfrak{z}_k$ has the same impact.

3. Theoretical study. This section presents some properties of the adaptive estimate $\widehat{\theta}$. We suppose that the parameters $\mathfrak{z}_k$ of the procedure are selected in such a way that the condition (2.1) is fulfilled. The main result is the "oracle" property of the adaptive estimate $\widehat{\theta}$ which claims that the risk of adaptive estimation is up to a logarithmic multiplier as good as the risk of the ideal ("oracle") estimate. In the proof we distinguish between 3 cases: parametric, local parametric and nonparametric. The parametric case means that $\theta_k \stackrel{\text{def}}{=} \boldsymbol{E}\widetilde{\theta}_k \equiv \theta$ for all $k \le K$. This case easily reduces to the null hypothesis $\theta_1 = \ldots = \theta_K = 0$ and the "oracle" property of

the adaptive estimate $\widehat{\theta}$ is granted by the construction, more precisely, by the "propagation condition" (2.1). The local parametric case means that for some $k < K$ holds $\theta_1 = \ldots = \theta_k = \theta$. In this case, the construction ensures the "oracle" property for the adaptive estimate $\widehat{\theta}_k$ obtained after the first k steps of the procedure. Then we show that a similar "oracle" property of the estimate $\widehat{\theta}$ can be obtained in the nonparametric situation under the so called "small modeling bias" condition. This condition is used to give a formal definition of the "oracle" choice. The final "oracle" result for the adaptive estimate $\widehat{\theta}$ is obtained by combining the previously established "propagation" result under the small modeling bias condition with the "stability" property which is granted by the adaptive procedure itself.

First we present some bounds on $\mathfrak{z}_k$ that ensure (2.1). Next we study the properties of $\widehat{\theta}$ in the parametric and local parametric situation. Then we extend this result to the nonparametric situation using the so called "small modeling bias" condition. Finally we present the "stability" property and state the "oracle" result.

3.1. *Bounds for the critical values.* This section presents some upper and lower bounds for the critical values $\mathfrak{z}_k$. The results are established under the following condition on the variances v_k.

(*MD*) for some constants $\mathfrak{u}_0, \mathfrak{u}$ with $1 < \mathfrak{u}_0 \le \mathfrak{u}$, the variances v_k satisfy

$$v_{k-1} \le \mathfrak{u} v_k, \qquad \mathfrak{u}_0 v_k \le v_{k-1}, \qquad 2 \le k \le K.$$

Our first result presents some upper bound for the parameters $\mathfrak{z}_k$ under condition (MD). The proof is given in the Appendix.

THEOREM 3.1. *Assume (MD). Let $\theta_k = \theta$ for all $k \ge 1$. Then there are three constants a_0, a_1 and a_2 depending on r and $\mathfrak{u}_0$, $\mathfrak{u}$ only such that the choice*

$$\mathfrak{z}_k = a_0 + a_1 \log \alpha^{-1} + a_2 r \log(v_k / v_K)$$

ensures (2.1) for all $k \le K$. Particularly, $\boldsymbol{E}_0 |v_K^{-1} (\widetilde{\theta}_K - \widehat{\theta})^2|^r \le \alpha \mathfrak{c}_r$.

REMARK 3.2. The result of Theorem 3.1 presents some upper bounds for the critical values. These upper bounds will be used for our theoretical study, however, they do not appear in the proposed adaptive procedure. An interesting observation is that these upper bounds linearly decrease with k. The reason for a decrease is relatively clear. Under the null hypothesis the procedure should not terminate at intermediate steps and the "oracle"

estimate is $\widetilde{\theta}_K$. An early stop ("false alarm") $\widehat{k} = k$ for $k < K$ results in selecting the estimate $\widetilde{\theta}_k$ which has much larger variability than $\widetilde{\theta}_K$. The smaller k is, the larger is the associated loss in the estimation quality. Therefore, the test at the early stage of the procedure should be rather conservative while a "false alarm" at the final steps of the procedure is not so critical and we are more interested to improve sensitivity by applying non-conservative critical values.

Our next result shows that the linear growth of the critical values $\mathfrak{z}_k$ with $K-k$ is not only sufficient but also necessary for providing (2.1). To highlight the contribution of every particular value $\mathfrak{z}_k$, we consider the situation when all the previous parameters are equal to infinity: $\mathfrak{z}_1 = \ldots = \mathfrak{z}_{k-1}$. This effectively means that the procedure cannot terminate at the first $k-1$ steps due to a possibly wrong choice of the corresponding critical values.

THEOREM 3.3. *Assume* (MD). *Let, for a fixed* k, *it holds* $\mathfrak{z}_1 = \ldots = \mathfrak{z}_{k-1} = \infty$, *and*

$$\boldsymbol{E}_0 \widetilde{\theta}_k \widetilde{\theta}_{k+1} \le \rho \sqrt{v_k v_{k+1}} \tag{3.1}$$

for some $\rho < 1$. *Then the condition (2.1)implies that*

$$\mathfrak{z}_k \ge c_* r \log(v_k / v_K)$$

for some positive constant c_* *depending on* ρ *and the constants* $\mathfrak{u}$ *and* $\mathfrak{u}_0$ *from condition* (MD) *only.*

The proof is again moved to the Appendix.

REMARK 3.4. The condition (3.1) means that the correlation between estimates $\widetilde{\theta}_k$ and $\widetilde{\theta}_{k+1}$ is bounded away from 1. Note that for the example of "sequence space model" this condition is fulfilled with $\rho = \mathfrak{u}$.

3.2. *Behavior in the local parametric situation.* The parametric situation can be understood as the case when $\theta_1 = \theta_2 = \ldots = \theta_K$. In this case the estimate $\widetilde{\theta}_K$ is unbiased and has the smallest variance and hence, the smallest risk described by the formula $\boldsymbol{E}|v_K^{-1}(\widetilde{\theta}_K - \theta)^2|^r = \mathfrak{c}_r$. A natural requirement to any adaptive procedure is to provide a similar accuracy of the adaptive estimate under the parametric hypothesis. Similarly, the local parametric situation corresponds to the case when $\theta_1 = \ldots = \theta_k = \theta$ for some $k \le K$. In this case it is natural to require that the adaptive estimate $\widehat{\theta}_k$ after k steps is close to its non-adaptive counterpart $\widetilde{\theta}_k$. This property is actually provided by the construction of the critical values.

THEOREM 3.5. *Let* $\theta_1 = \theta_2 = \ldots = \theta_K = \theta$. *Then it holds*

$$\boldsymbol{E}|v_K^{-1}(\widehat{\theta} - \widetilde{\theta}_K)^2|^r \le \alpha \mathfrak{c}_r .$$

Moreover, if $\theta_1 = \theta_2 = \ldots = \theta_k = \theta$ *for some* $k \le K$, *then*

$$\boldsymbol{E}|v_k^{-1}(\widehat{\theta}_k - \widetilde{\theta}_k)^2|^r \le \alpha \mathfrak{c}_r .$$

PROOF. Only the differences $\widetilde{\theta}_l - \widetilde{\theta}_k$ appear in the definition of the test statistics T_{lk}. In view of the decomposition $\widetilde{\theta}_k = \theta + \xi_k$, see Theorem 1.1, the value θ cancels there. Similarly, the adaptive estimate $\widehat{\theta}_k$ coincides with one of $\widetilde{\theta}_1, \ldots, \widetilde{\theta}_k$ and the value θ cancels in the difference $\widehat{\theta}_k - \widetilde{\theta}_k$ as well. Hence, we can assume $\theta = 0$ and $\widetilde{\theta}_k = \xi_k$. Then the results follow from the constraints (2.1) on the critical values $\mathfrak{z}_k$. □

3.3. *"Small modeling bias" condition and "propagation" property.* Theorem 3.5 describes the performance of the estimate $\widetilde{\theta}_k$ under the parametric or local parametric assumption. Now we aim to extend this result to the general nonparametric situation when the identities $\theta_1 = \theta_2 = \ldots = \theta_k = \theta$ are only approximately fulfilled and the deviation from the null hypothesis H_k is not significant.

As mentioned in Section 2.1, the choice of critical values $\mathfrak{z}_k$ is determined by the joint distribution of the test statistics $T_{lk} = v_l^{-1}(\widetilde{\theta}_l - \widetilde{\theta}_k)^2$ under the parametric hypothesis $X \equiv 0$. An extension of this result to the nonparametric situation leads to considering the similar distribution in the general case. Let $\boldsymbol{P}_k$ mean the joint distribution of $\widetilde{\theta}(k) = (\widetilde{\theta}_1, \ldots, \widetilde{\theta}_k)^\top$ for $k \ge 1$. By Theorem 1.1 this is a Gaussian vector. Hence, its distribution is described by the mean and the variance. By Theorem 1.1 $\boldsymbol{E}\,\widetilde{\theta}(k) = \theta(k) = (\theta_1, \ldots, \theta_k)^\top$. Let also B_k means the covariance matrix of the vector $\widetilde{\theta}(k)$. Then $\boldsymbol{P}_k$ is the normal distribution with the mean $\theta(k)$ and the covariance matrix B_k, $\boldsymbol{P}_k = \mathcal{N}(\theta(k), B_k)$. Similarly, if $\boldsymbol{P}_{\theta,k}$ is the distribution of $\widetilde{\theta}(k)$ under the local parametric situation $\theta_1 = \ldots = \theta_k = \theta$, then $P_{\theta,k} = \mathcal{N}(\theta_0(k), B_k)$, where $\theta_0(k) = (\theta, \ldots, \theta)^\top$.

LEMMA 3.6. *For* $k \ge 1$, *define* $b(k) = (b_1, \ldots, b_k)^\top$ *with* $b_k = \theta_k - \theta$ *and*

$$\Delta_k \stackrel{\mathrm{def}}{=} b^\top(k) B_k^{-1} b(k).$$

Then the Kullback-Leibler divergence $\mathcal{K}(\boldsymbol{P}_k, \boldsymbol{P}_{\theta,k})$ *between* $\boldsymbol{P}_k$ *and* $\boldsymbol{P}_{\theta,k}$ *fulfills*

$$\mathcal{K}(\boldsymbol{P}_k, \boldsymbol{P}_{\theta,k}) \stackrel{\mathrm{def}}{=} \boldsymbol{E}_k \log\Big(\frac{d\boldsymbol{P}_k}{d\boldsymbol{P}_{\theta,k}}\Big) = \Delta_k/2$$

and the values Δ_k *grow with* k. *It also holds for any* $s > 1$

$$\frac{1}{s} \log \boldsymbol{E}_{\theta,k} \Big(\frac{d\boldsymbol{P}_k}{d\boldsymbol{P}_{\theta,k}} \Big)^s = \frac{\Delta_k (s-1)}{2} \, .$$

Moreover, if ζ *is measurable function of* $\widetilde{\theta}_1, \ldots, \widetilde{\theta}_k$, *then it holds with* $s' = s/(s-1)$

$$\boldsymbol{E}\zeta \le \big(\boldsymbol{E}_{\theta,k} \zeta^{s'} \big)^{1/s'} \exp\{ \Delta_k (s-1)/2 \}.$$

In particular, for $s = 2$ *it holds* $\boldsymbol{E}\zeta \le \big(\mathrm{e}^{\Delta_k} \boldsymbol{E}_{\theta,k} \zeta^2 \big)^{1/2}$.

Proof. Define $Z_k = d\boldsymbol{P}_k / d\boldsymbol{P}_{\theta,k}$. Then $\log Z_k = b^\top(k) B_k^{-1/2} \xi_k + b^\top(k) B_k^{-1} b(k)/2$ with $\xi_k \sim \mathcal{N}(0,1)$ and $\boldsymbol{E}_k \log(Z_k) = \Delta_k/2$. This immediately implies that Δ_k monotonously increase with k, that is, $\Delta_k \le \Delta_{k'}$ for $k < k'$. Similarly,

$$\begin{aligned} \boldsymbol{E}_{\theta,k} Z_k^s &= \boldsymbol{E}_{\theta,k} \exp\{ s b^\top(k) B_k^{-1/2} \xi_k - b^\top(k) B_k^{-1} b(k) s/2 \} \\ &= \exp\{ b^\top(k) B_k^{-1} b(k) (s^2 - s)/2 \}. \end{aligned}$$

Next, let ζ be a measurable function of the vector $\widetilde{\theta}(k)$. It holds $\boldsymbol{E}\zeta = \boldsymbol{E}_{\theta,k} \zeta Z_k$. By the Hölder inequality

$$\boldsymbol{E}_{\theta,k} \zeta Z_k \le \big(\boldsymbol{E}_{\theta,k} \zeta^{s'} \big)^{1/s'} \big(\boldsymbol{E}_{\theta,k} Z_k^s \big)^{1/s}$$

and the assertion follows. □

Due to Lemma 3.6, the value Δ_k can be used to measure the distance between the two models: one corresponds to the local parametric situation with $\theta_1 = \theta_2 = \ldots = \theta_k = \theta$ and the other one describes the distribution of the same vector $\widetilde{\theta}(k)$ in the general nonparametric situation. We call this value Δ_k the "modeling bias" because it describes how much we have to pay in the risk for using the "wrong" parametric model in place of the underlying nonparametric one. The "small modeling bias" (SMB) condition simply means that the value Δ_k is sufficiently small.

The result of Lemma 3.6 implies that the bound for the risk of estimation $\boldsymbol{E}_0 \{ v_k^{-1} (\widetilde{\theta}_k - \theta)^2 \}^r$ under the parametric hypothesis translates under the SMB condition into the bound for the risk $\boldsymbol{E} \{ v_k^{-1} (\widetilde{\theta}_k - \theta)^2 \}^{r/s'}$. Similarly one can bound $\boldsymbol{E} \{ v_k^{-1} (\widehat{\theta}_k - \widetilde{\theta}_k)^2 \}^{r/s'}$.

In what follows we apply the result of Lemma 3.6 with $s = s' = 2$ which nicely simplifies the notation. Note, however, that any $s > 1$ can be used. For instance, taking a large s leads to the value of s' close to one.

THEOREM 3.7. *For any* $r > 0$, *it holds for every* $k \le K$

$$\begin{aligned} \boldsymbol{E}\{v_k^{-1}(\widetilde{\theta}_k - \theta)^2\}^{r/2} &\le \sqrt{e^{\Delta_k}\mathfrak{c}_r}. \\ \boldsymbol{E}\{v_k^{-1}(\widetilde{\theta}_k - \widehat{\theta}_k)^2\}^{r/2} &\le \sqrt{e^{\Delta_k}\alpha\mathfrak{c}_r}. \end{aligned}$$

PROOF. The bound follows directly from Lemma 3.6 and Theorem 3.5. □

We call this result the "propagation" property because it ensures that with a high probability the procedure does not terminate as long as the "small modeling bias" condition is fulfilled. Note that a similar property has been proved for the original procedure in Lepski (1990), see also Lepski (1991, 1992) however, under the additional condition that the critical values $\mathfrak{z}_k$ are sufficiently large. We instead use the "propagation condition" (2.1).

3.4. *"Stability after propagation" and "oracle" results.* Due to the "propagation" result of Theorem 3.7, the procedure performs well as long as the SMB condition is fulfilled which means that the value Δ_k remains bounded by some (small) constant. We formalize this condition in the form $\Delta_k \le \Delta$. Here Δ is an arbitrary number which will determine the "oracle" choice. We will show in Section 3.5 that in typical situations this value Δ is similar to the ratio of the squared bias to the variance variance of $\widetilde{\theta}$. Note however, that the value Δ only appears in our theoretical study, it does not affect the procedure and we do not need to fix this value. The result apply whatever $\Delta > 0$.

To establish the accuracy result for the final estimate $\widehat{\theta}$, we have to check that the adaptive estimate $\widehat{\theta}_k$ does not vary much at the steps "after propagation" when the "modeling bias" Δ_k becomes large.

THEOREM 3.8. *It holds for every* $k < K$

$$v_k^{-1}(\widetilde{\theta}_k - \widehat{\theta})^2 \mathbf{1}(\widehat{k} > k) \le \mathfrak{z}_k. \tag{3.2}$$

PROOF. The result follows by the definition of $\widehat{\theta} = \widetilde{\theta}_{\widehat{k}}$ and $\widehat{\theta}_k = \widetilde{\theta}_{\min\{\widehat{k},k\}}$ because $\widehat{k}$ is accepted and $\min\{\widehat{k}, k\} \le \widehat{k}$. □

REMARK 3.9. An interesting feature of this "stability" result is that it is fulfilled not only with a high probability, it always applies. This property follows directly from the construction of the procedure.

Combination of the "propagation" and "stability" statements implies the main result concerning the properties of the adaptive estimate $\widehat{\theta}$. In the formulation of this and the further results we assume some constant $\Delta > 0$ to be fixed.

THEOREM 3.10. *Let k^* be the maximal value k such that $\Delta_k \le \Delta$. Then*

$$\boldsymbol{E}\bigl|v_{k^*}^{-1}\bigl(\widetilde{\theta}_{k^*} - \widehat{\theta}\bigr)^2\bigr|^{r/2} \le \sqrt{\alpha \mathfrak{c}_r e^{\Delta}} + \mathfrak{z}_{k^*}^{r/2}.$$

PROOF. The events $\mathbf{1}\bigl(\widehat{k} > k^*\bigr)$ and $\mathbf{1}\bigl(\widehat{k} \le k^*\bigr)$ do not overlap and $\widehat{\theta} = \widehat{\theta}_{k^*}$ for $\widehat{k} \le k^*$. This yields the representation

$$\boldsymbol{E}\bigl|v_{k^*}^{-1}\bigl(\widetilde{\theta}_{k^*} - \widehat{\theta}\bigr)^2\bigr|^{r/2} = \boldsymbol{E}\bigl|v_{k^*}^{-1}\bigl(\widetilde{\theta}_{k^*} - \widehat{\theta}\bigr)^2\bigr|^{r/2}\mathbf{1}\bigl(\widehat{k} > k^*\bigr) + \boldsymbol{E}\bigl|v_{k^*}^{-1}\bigl(\widetilde{\theta}_{k^*} - \widehat{\theta}_{k^*}\bigr)^2\bigr|^{r/2}.$$

Now the result follows from Theorems 3.7 and 3.8. □

We briefly comment on the meaning of the "oracle" result. Theorem 3.7 ensures that the estimation loss $v_k^{-1}\bigl(\widetilde{\theta}_k - \theta\bigr)^2$ is bounded with a high probability provided that the "modeling bias" Δ_k is not too big. The "oracle" choice k^* is the largest one for which the "small modeling bias" condition $\Delta_k \le \Delta$ holds leading to the accuracy $|\widetilde{\theta}_{k^*} - \theta|$ of order $v_{k^*}^{1/2}$. We aim to build an adaptive estimate which delivers the same quality as the "oracle" one. Theorem 3.10 claims that the difference $\widehat{\theta} - \widetilde{\theta}_{k^*}$ between the adaptive estimate $\widehat{\theta}$ and "oracle" is indeed of order $v_{k^*}^{1/2}$ up to the factor $\sqrt{\mathfrak{z}_{k^*}}$ which can be viewed as the "price" for adaptation. Another "price" we pay is that the "oracle" result is stated for the polynomial loss of power r while the "oracle" is trained under the parametric model for the loss of power $2r$.

We also present a corollary of the "oracle" result concerning the risk of the adaptive estimate $\widehat{\theta}$ for the special case with $r = 1$. The other values of r can be considered as well, one only has to update the constants depending on r. We also assume that $\alpha \le 1$.

COROLLARY 3.11. *Let k^* be the the largest k with $\Delta_k \le \Delta$. Then*

$$v_{k^*}^{-1/2}\boldsymbol{E}\bigl|\widehat{\theta} - \theta\bigr| \le 2\sqrt{e^{\Delta}} + \sqrt{\mathfrak{z}_{k^*}}.$$

PROOF. Just observe that

$$\bigl|\widehat{\theta} - \theta\bigr| \le \bigl|\widetilde{\theta}_{k^*} - \theta\bigr| + \bigl|\widetilde{\theta}_{k^*} - \widehat{\theta}\bigr|$$

and the result follows from Theorem 3.10 in view of $\mathfrak{c}_1 = 1$. □

REMARK 3.12. Recall that in the parametric situation, the risk $\boldsymbol{E}\bigl|v_{k^*}^{-1}\bigl(\widetilde{\theta}_{k^*}-\theta\bigr)^2\bigr|$ of $\widetilde{\theta}_{k^*}$ is bounded by $\mathfrak{c}_1 = 1$, cf. Theorem 1.1. In the nonparametric situation, the result is only slightly worse. It bounds the absolute loss $\bigl|\widehat{\theta}-\theta\bigr|$ instead of squared loss and there is an additional factor $\sqrt{e^{\Delta}}$ which takes into account the modeling bias. There is also an additional term proportional to $\sqrt{\mathfrak{z}_{k^*}}$ which can be considered as the payment for adaptation. Due to Theorem 3.1, $\mathfrak{z}_{k^*}$ is bounded from above by $a_0 + a_1 \log \alpha^{-1} + a_2 \log(v_{k^*}/v_K)$.

REMARK 3.13. The "oracle" results are especially popular in (global) model selection, see e.g. Birgé, L.; Massart, P. (1993, 1998), Birgé, L. (2006), Juditsky, Rigollet, Tsybakov (2006) and references therein. The corresponding result compare the risk of adaptive estimate with the risk of the "oracle" which is the risk minimizer over the considered family of estimates. Our results are stated for the local model selection in the sense that we estimate a linear functional rather than the whole model. This makes a significant difference between two problems. In the local model selection the estimates have to be ordered by their variability, the definition of the "oracle" relies to this ordering, and the results are stated about the difference between the adaptive and "oracle" estimates. Consider, for instance, an example for a sequence space model in which that the sequence $\theta_k = \sum_{i=1}^{m_k} Y_i$ is very irregular but by chance $\theta_K = \theta$. Then the best linear estimate is $\widetilde{\theta}_K$ but it is not the "oracle" one because the modeling bias Δ_K is big. More concise relations between these two problems have to be elaborated elsewhere.

3.5. *"Small modeling bias" condition versus "bias-variance trade-off"*. The standard approach for selecting the optimal index k is based on balancing an upper bound $\bar{b}_k$ for the bias $b_k = |\theta_k - \theta|$ and the standard deviation $v_k^{1/2}$ of the estimate $\widetilde{\theta}_k$, see e.g. Goldenshluger (1998) and Goldenshluger and Pereversev (1999). This section shows that under some additional technical assumptions this approach is nearly equivalent to the "small modeling bias" condition advocated in this paper.

In addition to (MD) we suppose the following properties of the covariance matrices B_k. Let $B_{k,\mathrm{diag}}$ be the diagonal matrix with the same diagonal entries as for B_k. Define also $D_k = B_k^{1/2}$ and $D_{k,\mathrm{diag}} = B_{k,\mathrm{diag}}^{1/2}$. The required conditions reads as follows:

$(\boldsymbol{Dk})$ It holds for some constant $\mathfrak{s}$ and all $k \le K$

$$D_k^{-1} \preceq \mathfrak{s} D_{k,\mathrm{diag}}^{-1}.$$

Here the notation $A \preceq B$ for two matrices A, B means that $|Av| \le |Bv|$ for any vector v. If B is symmetric and invertible, this is equivalent to

the condition that the maximal eigenvalue of the matrix $B^{-1}A^\top AB^{-1}$ is bounded by $\mathfrak{s}^2$.

Condition (Dk) allows to rewrite the "small modeling bias" condition $\bigl|D_k^{-1}b(k)\bigr|^2 \le \Delta$ in the following form:

$$\bigl|D_{k,\mathrm{diag}}^{-1}b(k)\bigr|^2 \le \Delta/\mathfrak{s}^2$$

or, equivalently,

$$\sum_{l=1}^{k} b_l^2/v_l \le \Delta/\mathfrak{s}^2. \tag{3.3}$$

Let now $\bar{b}_k$ be a monotonously increasing upper bound for b_k, that is, $\bar{b}_l = \max_{k\le l} b_k$. The "balance" relation (bias-variance trade-off) is usually written in the form

$$\bar{b}_{k^*} \le C_b v_{k^*}^{1/2} \tag{3.4}$$

for some fixed constant C_b. The next result shows that this relation implies the "small modeling bias" condition (3.3).

THEOREM 3.14. *Suppose (MD) and (Dk) . Then for the index k^* defined by the balance relation (3.4), the "small modeling bias" condition is also fulfilled with $\Delta = \mathfrak{s}^2 C_{\mathfrak{u}_0} C_b^2$.*

PROOF. It suffices to note that under the conditions of the theorem,

$$\sum_{l=1}^{k} b_l^2/v_l \le \bar{b}_k^2 \sum_{l=1}^{k} v_l^{-1} \le C_{\mathfrak{u}_0}\bar{b}_k^2 v_k^{-1}$$

$C_{\mathfrak{u}_0} = (1-\mathfrak{u}_0^{-1})^{-1}$. Now condition (Dk) provides

$$\bigl|D_k^{-1}b(k)\bigr|^2 \le \mathfrak{s}^2\bigl|D_{k,\mathrm{diag}}^{-1}b(k)\bigr|^2 \le \mathfrak{s}^2 C_{\mathfrak{u}_0}\bar{b}_k^2 v_k^{-1} \le \mathfrak{s}^2 C_{\mathfrak{u}_0} C_b^2$$

thus yielding (3.3). □

Combination of the results of Theorem 3.14 and Corollary 3.11 yields the following

COROLLARY 3.15. *Suppose (MD) and (Dk) and let the index k^* be defined by the balance relation (3.4). Then for $\Delta = \mathfrak{s}^2 C_{\mathfrak{u}_0} C_b^2$ and any $r > 0$*

$$\begin{aligned} \boldsymbol{E}\bigl|v_{k^*}^{-1}\bigl(\widehat{\theta} - \widetilde{\theta}_{k^*}\bigr)^2\bigr|^{r/2} &\le \sqrt{e^{\Delta}\alpha\mathfrak{c}_r} + \mathfrak{z}_{k^*}^{r/2}, \\ v_{k^*}^{-1/2}\boldsymbol{E}\bigl|\widehat{\theta} - \theta\bigr| &\le 2\sqrt{e^{\Delta}} + \sqrt{\mathfrak{z}_{k^*}}. \end{aligned}$$

We conclude this section by a small discussion about between of the "oracle" result and minimax rate of convergence. Most of theoretical results in the modern statistical literature are stated about the asymptotic minimax rate of estimation on the functional classes, see e.g. Goldenshluger (1999), Goldenshluger and Pereversev (2000), Cavalier, Golubev, Picard and Tsybakov (2002), Cavalier and Tsybakov (2001, 2002). The rate optimal procedures can be constructed as linear spectral cut-off estimates satisfying the "bias-variance" relation which balances the variance of the stochastic component of the estimate with the upper bound for the squared bias on the considered functional classes. An immediate corollary of Theorem 3.14 is that the proposed adaptive estimate which selects one out of the family of the spectral cut-off estimates $\widetilde{\theta}_k$ is rate optimal (up to a logarithmic multiplier) for all such set-ups, because it also provides the accuracy corresponding to the "balance" relation. A precise formulation of this result lies beyond the focus of this paper.

3.6. *Application to the "sequence space" model.* This section specifies the general results to the "sequence space" example considered in Section 1.3. In this case,

$$\widetilde{\theta}_k = y_1 + \ldots + y_{m_k}, \qquad v_k = \sigma_1^2 + \ldots + \sigma_{m_k}^2 \tag{3.5}$$

with $m_1 > m_2 > \ldots > m_K \geq 1$. We additionally assume that σ_i^2 are monotonous in i. The condition (MD) means in this situation that the indices m_k properly decrease to provide an exponential decrease of the sums v_k in k. The next result shows that this condition ensures (Dk).

LEMMA 3.16. *For the model (3.5), the condition* (MD) *implies* (Dk) *with the constant* $\mathfrak{s} = (1 - 1/\mathfrak{u}_0)^{-3/2}$.

PROOF. It suffices to show that the minimal eigenvalue of the matrix $M_k = D_{k,\mathrm{diag}}^{-1} B_k D_{k,\mathrm{diag}}^{-1}$ is bounded away from zero, or, equivalently, the largest eigenvalue of M_k^{-1} is bounded from above: $\|M_k^{-1}\|_\infty \leq (1-1/\mathfrak{u}_0)^{-3}$. Clearly $\boldsymbol{E}_0 \widetilde{\theta}_j \widetilde{\theta}_l = \boldsymbol{E}_0 \widetilde{\theta}_l^2 = v_l$ for $j \leq l$, and M_k is the symmetric matrix composed by the elements of the form $\rho_{jl} = v_j^{-1/2} v_l^{-1/2} \boldsymbol{E}_0 \widetilde{\theta}_j \widetilde{\theta}_l = (v_j/v_l)^{1/2}$ for $j \leq l$. In other words, M_k is the covariance matrix for the set of random variables $\eta_l = \widetilde{\theta}_l / v_l^{1/2}$ for $l = 1, \ldots, k$.

Define $\xi_l = v_l^{-1/2}(\widetilde{\theta}_l - \widetilde{\theta}_{l+1})$ for $l < k$ and $\xi_k = v_k^{-1/2} \eta_k$. The random variables ξ_l are independent zero mean normal with the variance $s_l \stackrel{\mathrm{def}}{=} \boldsymbol{E}\xi_l^2 = v_l^{-1}(v_l - v_{l+1})$ for $l < k$ and $s_k = 1$. The condition (MD) implies

for all $l \le k$ that $(1 - 1/\mathfrak{u}_0) \le s_l \le (1 - 1/\mathfrak{u})$. Define $\xi^{(k)} = (\xi_1, \ldots, \xi_k)^\top$ and $\eta^{(k)} = (\eta_1, \ldots, \eta_k)^\top$. The identities $\xi_l = \eta_l - \eta_{l+1}(v_{l+1}/v_l)^{1/2}$ for $l < k$ can be written as $\xi^{(k)} = A_k \eta^{(k)}$ where where line l of the matrix A_k only contains only two nonzero entries: $a_{l,l} = 1$ and $a_{l,l+1} = -v_{l+1}^{1/2}/v_l^{1/2}$ for $l = 1, \ldots, k-1$. Again, the condition (MD) implies that $\|I - A_k\|_\infty \le 1/\mathfrak{u}_0$ and $\|A_k^{-1}\|_\infty = \|\{I - (I - A_k)\}^{-1}\| \le (1 - 1/\mathfrak{u}_0)^{-1}$. Similar bound holds for $A_k^\top$. Obviously $\boldsymbol{E}_0 \xi^{(k)} (\xi^{(k)})^\top = \Sigma_k \stackrel{\mathrm{def}}{=} \operatorname{diag}(s_1, \ldots, s_k)$. This yields

$$\Sigma_k = \boldsymbol{E} A_k \eta^{(k)} (\eta^{(k)})^\top A_k^\top = A_k M_k A_k^\top .$$

This easily yields

$$\|M_k^{-1}\|_\infty \le \|A_k^{-1}\|_\infty^2 \cdot \|\Sigma_k^{-1}\|_\infty \le (1 - 1/\mathfrak{u}_0)^{-3}$$

and the result follows. □

The estimate $\widetilde{\theta}_k$ has the bias $b_k = \theta_k - \theta = -\sum_{i=m_k+1}^{M} \mu_i$. The bias-variance relation (3.4) balances the non-decreasing envelope $\bar{b}_k = \max_{l \le k} |b_l|$ with the variance v_k^2 leading to the oracle choice k^*. Corollary 3.15 ensures for the adaptive estimate $\widehat{\theta}$ the accuracy of order $v_{k^*}^{-1/2}$ up to the multiplicative factor $\sqrt{\mathfrak{z}_{k^*}}$.

4. Simulation. This section illustrates the performance of the proposed procedure by means of two simulated examples. The first correspond to a severely ill-posed inverse problem and the second to an ill-posed problem. We focus on two important features of our procedure: "propagation property" and "adaptivity". The "propagation" property means that the selected models only in very few cases is larger than the "oracle" one, that means, the "false alarm" situation when the procedure stops but the modeling bias is still small is very rare. The "adaptivity" means that the ratio of the risk of the adaptive estimate to the risk of "oracle" is bounded by some fixed constant.

For simplicity we consider "sequence space" models, i.e. the data Y_i are generated by the following model : $Y_i = \mu_i + \sigma_i \delta \epsilon_i$, for $i = 1, \ldots, n$ for $n = 50$ and we assume that ϵ_i are i.i.d. standard normal. In each example the values $(\mu_i)_{i=1,\ldots,n}$ are generated randomly from a centered gaussian with a decreasing variance i^{-3} and we consider ten different models of this type. The error level δ is equal to $10^{-4}, 10^{-5}$ or 10^{-6}. In every example, the target is the sum of the parameters μ_i, that is, $\theta = \sum_{i=1}^{n} \mu_i$. This set-up is friendly advised by F. Bauer, see e.g. Bauer (2007).

We apply the proposed procedure to the family of "weak" estimates $\widetilde{\theta}_k = \sum_{i=1}^{m_k} Y_i$. The "metaparameters" are set as $\alpha = 1$ and $r = 1/2$. Our numerical result (not reported here) confirm that the critical values slightly increase with r and decrease with α, however, the final results are very insensitive to the choice of these metaparameters.

In the first example we choose $\sigma_i = a^i$ for $i = 1, \ldots, n$, where $a = n^{2/n}$. We consider the estimates $\widetilde{\theta}_k = \sum_{i=1}^{m_k} Y_i$ with $m_k = [n - 2 * (k-1)]$, for $k = 1, \ldots, K$ and $K = 20$, then $m_K = 12$.

The critical values $\mathfrak{z}_k$ are computed from 50000 Monte Carlo replications from the null hypothesis (pure noise model) using the sequential procedure from Section 2.1, see Table 1.

TABLE 1
Critical values computed under the null hypothesis from 50000 replications, when $K = 20$ *and* $(\sigma_i = ((n^{2/n})^i)_{i=1,\ldots,n}$ *using the sequential procedure.*

r	α	$\mathfrak{z}_1$	$\mathfrak{z}_2$	$\mathfrak{z}_3$	$\mathfrak{z}_4$	$\mathfrak{z}_5$	$\mathfrak{z}_6$	$\mathfrak{z}_7$	$\mathfrak{z}_8$	$\mathfrak{z}_9$	$\mathfrak{z}_{10}$
0.50	1.0	15.5	13.0	12.8	12.2	11.5	11.3	10.9	9.8	9.2	8.6
		$\mathfrak{z}_{11}$	$\mathfrak{z}_{12}$	$\mathfrak{z}_{13}$	$\mathfrak{z}_{14}$	$\mathfrak{z}_{15}$	$\mathfrak{z}_{16}$	$\mathfrak{z}_{17}$	$\mathfrak{z}_{18}$	$\mathfrak{z}_{19}$	
		8.3	7.6	7.0	6.6	5.9	5.2	4.5	3.6	2.5	

Figure 1 compares the results for our adaptive estimate with the "oracle" one. The "oracle" value k^* is defined as $\max\{k : \Delta_k < 1\}$. The results for other values of Δ, e.g. $\Delta = 0.5$ or $\Delta = 2$ are very similar and we do not report them here. Each row corresponds to a different level of the noise δ. The panel (a) draws the ratio of the adaptive risk $\boldsymbol{E}|\widehat{\theta} - \theta|$ obtained from 500 realizations to the corresponding "oracle" risk $\boldsymbol{E}|\widetilde{\theta}_{k^*} - \theta|$ for the ten different models. In the panel (b) we show the box-plot of $\widehat{k}$ from 500 replications and the "oracles" values k^* (triangles) for the ten different models describes above. One can see that the adaptive risk is in the most of cases not more than twice larger than the the oracle risk. The oracle choice k^* is usually smaller than the adaptively selected $\widehat{k}$ which illustrates "propagation" property: procedure does not stop until k^*. It is also worth noticing that both the "oracle" choice k^* and the adaptive values $\widehat{k}$ decrease with the noise, i.e. the smaller is the noise the more coefficients y_i are taken for estimating the sum $\theta = \sum_i \mu_i$.

In the second example we consider a model with $(\sigma_i = i^2)_{i=1,\ldots,n}$ and apply the estimates $\widetilde{\theta}_k = \sum_{i=1}^{m_k} Y_i$ with $m_k = [n/(2^{1/5})^{k-1}]$, for $k = 1, \ldots, K$ and $K = 15$, leading to $m_K = 7$. The critical values $\mathfrak{z}_k$ are computed from 50000 Monte Carlo replications under the null hypothesis, see Table 2.

Figure 2 presents the results comparing the performance of the adaptive

TABLE 2
Critical values computed under the null hypothesis from 50000 *replications, when* $K = 15$ *and* $(\sigma_i = i^2)_{i=1,\dots,n}$ *using the sequential procedure.*

r	α	$\mathfrak{z}_1$	$\mathfrak{z}_2$	$\mathfrak{z}_3$	$\mathfrak{z}_4$	$\mathfrak{z}_5$	$\mathfrak{z}_6$	$\mathfrak{z}_7$
0.5	1.0	11.4	10.7	9.9	9.2	8.7	8.3	7.3
		$\mathfrak{z}_8$	$\mathfrak{z}_9$	$\mathfrak{z}_{10}$	$\mathfrak{z}_{11}$	$\mathfrak{z}_{12}$	$\mathfrak{z}_{13}$	$\mathfrak{z}_{14}$
		6.6	6.1	5.7	4.9	3.7	2.9	2.0

and "oracle" estimates in the second example. The set-up is the same as in the first example and the results are very similar.

Then we would like to see the behavior of our procedure when $r = 1$, that is, when the risk is of the form $\boldsymbol{E}|\widehat{\theta} - \theta|^2$. We study the same two examples and Tables 3 and 4 give the critical values $\mathfrak{z}_k$. The adaptive index $\widehat{k}$ similarly to the earlier results, is a little bit larger than the oracle value k^*.

TABLE 3
Critical values computed under the null hypothesis from 50000 *replications, when* $K = 20$ *and* $(\sigma_i = (n^{2/n})^i)_{i=1,\dots,n}$ *using the sequential procedure.*

r	α	$\mathfrak{z}_1$	$\mathfrak{z}_2$	$\mathfrak{z}_3$	$\mathfrak{z}_4$	$\mathfrak{z}_5$	$\mathfrak{z}_6$	$\mathfrak{z}_7$	$\mathfrak{z}_8$	$\mathfrak{z}_9$	$\mathfrak{z}_{10}$
1	1.0	22.5	19.0	16.4	17.2	16.2	15.6	16.8	14.4	13.4	13.2
		$\mathfrak{z}_{11}$	$\mathfrak{z}_{12}$	$\mathfrak{z}_{13}$	$\mathfrak{z}_{14}$	$\mathfrak{z}_{15}$	$\mathfrak{z}_{16}$	$\mathfrak{z}_{17}$	$\mathfrak{z}_{18}$	$\mathfrak{z}_{19}$	
1	1.0	12.9	11.9	10.2	9.3	8.3	7.3	5.8	4.7	3.4	

TABLE 4
Critical values computed under the null hypothesis from 50000 *replications, when* $K = 15$ *and* $(\sigma_i = i^2)_{i=1,\dots,n}$ *using the sequential procedure.*

r	α	$\mathfrak{z}_1$	$\mathfrak{z}_2$	$\mathfrak{z}_3$	$\mathfrak{z}_4$	$\mathfrak{z}_5$	$\mathfrak{z}_6$	$\mathfrak{z}_7$
1	1.0	16.5	15.7	14.0	14.0	19.4	13.3	10.9
		$\mathfrak{z}_8$	$\mathfrak{z}_9$	$\mathfrak{z}_{10}$	$\mathfrak{z}_{11}$	$\mathfrak{z}_{12}$	$\mathfrak{z}_{13}$	$\mathfrak{z}_{14}$
1	1.0	9.4	8.9	8.6	7.0	5.0	4.0	2.9

We conclude from this simulation study that the performance of the method is completely in agreement with the theoretical conclusions and the procedure demonstrates quite reasonable performance in all the examples including regular and severely ill-posed problems and for different configurations of the signal and different noise levels.

5. Appendix: Proof of Theorems 3.1 and 3.3. Define for every $k \le K$ the random set

$$\mathcal{A}_k = \bigcap_{j=1}^{k-1} \Big\{ \max_{j<l\le k} v_j^{-1}(\widetilde{\theta}_j - \widetilde{\theta}_l)^2 \le \mathfrak{z}_j \Big\}. \tag{5.1}$$

Note first that $\widehat{\theta}_k = \widetilde{\theta}_k$ on $\mathcal{A}_k$ for all $k \le K$.

Therefore, it remains to bound the risk of $\widehat{\theta}_k$ on the complement $\overline{\mathcal{A}}_k$ of $\mathcal{A}_k$. Define $\mathcal{B}_{k-1} = \mathcal{A}_{k-1} \backslash \mathcal{A}_k$. By definition $\widehat{k} = \min\{\widehat{k}, k\} = k-1$ on $\mathcal{B}_{k-1}$ and $\overline{\mathcal{A}}_k = \bigcup_{l<k} \mathcal{B}_l$. First we bound the probability $\boldsymbol{P}_0(\mathcal{B}_l)$. Assumption (MD) yields for every $l < k$

$$\begin{aligned} v_l^{-1}(\widetilde{\theta}_l - \widetilde{\theta}_k)^2 &\le 2v_l^{-1}\{(\widetilde{\theta}_l - \theta)^2 + (\widetilde{\theta}_k - \theta)^2\} \\ &\le 2\{v_l^{-1}(\widetilde{\theta}_l - \theta)^2 + v_k^{-1}(\widetilde{\theta}_k - \theta)^2\}. \end{aligned}$$

Therefore, by Theorem 1.1, for all $\lambda < 1/2$

$$\boldsymbol{P}_0(\mathcal{B}_l) \le \sum_{j=1}^{l-1} \boldsymbol{P}_0\big(v_j^{-1}(\widetilde{\theta}_j - \widetilde{\theta}_l)^2 > \mathfrak{z}_j\big) \le 2(1-2\lambda)^{-1/2} \sum_{j=1}^{l-1} \mathrm{e}^{-\lambda \mathfrak{z}_j/4}.$$

Similarly for $l < k$

$$\begin{aligned} \boldsymbol{E}_0 |v_k^{-1}(\widetilde{\theta}_l - \widetilde{\theta}_k)^2|^r &\le 2^{(r-1)_+}\{\boldsymbol{E}_0 |v_k^{-1}(\widetilde{\theta}_l - \theta)^2|^r + \boldsymbol{E}_0 |v_k^{-1}(\widetilde{\theta}_k - \theta)^2|^r\} \\ &\le 2^{(r-1)_+}\Big\{\frac{v_l^r}{v_k^r} \boldsymbol{E}_0 |v_l^{-1}(\widetilde{\theta}_l - \theta)^2|^r + \boldsymbol{E}_0 |v_k^{-1}(\widetilde{\theta}_k - \theta)^2|^r\Big\} \\ &\le 2^{r\vee 1} \mathfrak{c}_r v_l^r / v_k^r. \end{aligned}$$

Now we employ the obvious representation

$$v_k^{-1}\big(\widetilde{\theta}_k - \widehat{\theta}_k\big)^2 = \sum_{l=1}^{k-1} v_k^{-1}\big(\widetilde{\theta}_k - \widetilde{\theta}_l\big)^2 \mathbf{1}(\mathcal{B}_l).$$

Therefore, for every r and $\lambda < 1/2$ by the Cauchy-Schwartz inequality

$$\begin{aligned} \boldsymbol{E}_0 |v_k^{-1}\big(\widetilde{\theta}_k - \widehat{\theta}_k\big)^2|^r &= \sum_{l=1}^{k-1} \boldsymbol{E}_0 |v_k^{-1}\big(\widetilde{\theta}_k - \widetilde{\theta}_l\big)^2|^r \mathbf{1}(\mathcal{B}_l) \\ &\le \sum_{l=1}^{k-1} \boldsymbol{E}_0^{1/2} |v_k^{-1}\big(\widetilde{\theta}_k - \widetilde{\theta}_l\big)^2|^{2r} \boldsymbol{P}_0^{1/2}(\mathcal{B}_l) \\ &\le 2^{r\vee 1} \mathfrak{c}_{2r}^{1/2} (1-2\lambda)^{-1/4} \sum_{l=1}^{k-1} \frac{v_l^r}{v_k^r} \Big(\sum_{j=1}^{l-1} \mathrm{e}^{-\lambda \mathfrak{z}_j/4} \Big)^{1/2}. \end{aligned}$$

It remains to check that the choice $\mathfrak{z}_j = a_0 + a_1 \log \alpha^{-1} + a_2 r \log(v_j/v_K)$ with properly selected a_0, a_1 and a_2 provide under condition (MD) the required bound $\boldsymbol{E}_0 |v_k^{-1}(\widetilde{\theta}_k - \widehat{\theta}_k)^2|^r \le \alpha \mathfrak{c}_r$ and Theorem 3.1 follows.

Now we turn to the proof of Theorem 3.3. We use again the decomposition

$$\begin{aligned} \boldsymbol{E}_0 |v_K^{-1}(\widehat{\theta} - \widetilde{\theta}_K)^2|^r &= \sum_{k=1}^{K-1} \boldsymbol{E}_0 |v_K^{-1}(\widetilde{\theta}_k - \widetilde{\theta}_K)^2|^r \mathbf{1}(\widehat{k} = k) \\ &\ge \boldsymbol{E}_0 |v_K^{-1}(\widetilde{\theta}_k - \widetilde{\theta}_K)^2|^r \mathbf{1}(\widehat{k} = k) \end{aligned}$$

for any $k < K$. The definition of $\widehat{k}$ implies that

$$\mathbf{1}(\widehat{k} = k) \ge \mathbf{1}(v_k^{-1}(\widetilde{\theta}_{k+1} - \widetilde{\theta}_k)^2 > \mathfrak{z}_k)$$

To simplify the presentation, we assume that $\widetilde{\theta}_K = 0$. The general case can be reduced to this one.

Let $\mathcal{F}_k$ be the σ-field generated by $\widetilde{\theta}_k$. The estimate $\widetilde{\theta}_{k+1} - \widetilde{\theta}_k$ can be decomposed as $\widetilde{\theta}_{k+1,k} + \rho \widetilde{\theta}_k$ where ρ is the correlation coefficient between $\widetilde{\theta}_k$ and $\widetilde{\theta}_{k+1} - \widetilde{\theta}_k$:

$$\rho = \frac{\boldsymbol{E}_0\{\widetilde{\theta}_k(\widetilde{\theta}_{k+1} - \widetilde{\theta}_k)\}}{\{v_k \operatorname{Var}(\widetilde{\theta}_{k+1} - \widetilde{\theta}_k)\}^{1/2}}$$

and $\widetilde{\theta}_{k+1,k} = \widetilde{\theta}_{k+1} - \widetilde{\theta}_k$ is independent of $\mathcal{F}_k$. Note that $|\widetilde{\theta}_{k+1} - \widetilde{\theta}_k|^2 = |\widetilde{\theta}_{k+1,k}|^2 + \rho^2 |\widetilde{\theta}_k|^2$ and hence,

$$\begin{aligned} \mathbf{1}(v_k^{-1}|\widetilde{\theta}_{k+1} - \widetilde{\theta}_k|^2 > \mathfrak{z}_k) &= \mathbf{1}(v_k^{-1}|\widetilde{\theta}_{k+1,k}|^2 + v_k^{-1}\rho^2|\widetilde{\theta}_k|^2 > \mathfrak{z}_k) \\ &\ge \mathbf{1}(v_k^{-1}|\widetilde{\theta}_{k+1,k}|^2 > \mathfrak{z}_k). \end{aligned}$$

The condition (MD) ensures that $v_{k+1,k} := \operatorname{Var}(\widetilde{\theta}_{k+1,k}) \ge v_k / C_{\mathfrak{u}}$ with some constant $C_{\mathfrak{u}} > 0$ depending on $\mathfrak{u}$ only. This yields that

$$\boldsymbol{P}_0(v_k^{-1}|\widetilde{\theta}_{k+1,k}|^2 > \mathfrak{z}_k) \ge \boldsymbol{P}_0(v_{k+1,k}^{-1}|\widetilde{\theta}_{k+1,k}|^2 > \mathfrak{z}_k C_{\mathfrak{u}}) = Q(\mathfrak{z}_k C_{\mathfrak{u}})$$

with $Q(z) = \boldsymbol{P}(|\xi|^2 > z)$ and standard normal ξ. Therefore,

$$\begin{aligned} &\boldsymbol{E}_0 |\widetilde{\theta}_k|^{2r} \mathbf{1}(v_k^{-1}|\widetilde{\theta}_{k+1} - \widetilde{\theta}_k|^2 > \mathfrak{z}_k) \\ &\ge \boldsymbol{E}_0 |\widetilde{\theta}_k|^{2r} \mathbf{1}(v_k^{-1}|\widetilde{\theta}_{k+1,k}|^2 > \mathfrak{z}_k) \\ &= \boldsymbol{E}_0 \Big[|\widetilde{\theta}_k|^{2r} \boldsymbol{E}_0\{\mathbf{1}(|\widetilde{\theta}_{k+1,k}|^2 > v_k \mathfrak{z}_k) \mid \mathcal{F}_k\} \Big] \\ &\ge Q(\mathfrak{z}_k C_{\mathfrak{u}}) \boldsymbol{E}_0 |\widetilde{\theta}_k|^{2r}. \end{aligned}$$

Next we have $\boldsymbol{E}_0|\widetilde{\theta}_k|^{2r} = \mathfrak{c}_r v_k^r$. Then

$$\alpha \mathfrak{c}_r \geq \boldsymbol{E}_0|v_K^{-1}\widehat{\theta}^2|^r \geq \mathfrak{c}_r\, Q(\mathfrak{z}_k C_{\mathrm{u}})(v_k v_K^{-1})^r.$$

The inequality

$$Q(z) \geq \frac{e^{-z/2}}{2\pi}(1/\sqrt{z} - 1/z^{3/2}),$$

and the usual upper bounds for the logarithm yields for some constant $C > 0$,

$$\log \alpha^{-1} + r \log \frac{v_k}{v_K} \leq C\, C_{\mathrm{u}}\, z$$

References.

[1] Y. Aït-Sahalia, Y. and Lo, A.W. (1998). Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices *The Journal of Finance* **53** (2), 499-?547.

[2] Bauer, F. (2007). Some Considerations Concerning Regularization and Parameter Choice Algorithms. *Inverse Problems* **23**, pp. 837–858 (2007).

[3] Birgé, L.; Massart, P. (1993). Rate of convergence for minimum contrast estimators. *Probab.Theory and Related Fields* **97**, 113–150.

[4] Birgé, L.; Massart, P. (1998). Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli* **4**, No.3, 329–375 (1998).

[5] Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. Henri Poincaré, Probab. Stat.* **42**, No. 3, 273–325 (2006).

[6] Cai, T. T. and Peter Hall, P. (2006). Prediction in functional linear regression *Ann. Statist.* **34**, Number 5, 2159–2179.

[7] Cavalier L. (2001). On the problem of local adaptive estimation in tomography. *Bernoulli* **7**, 63–78.

[8] Cavalier L., Golubev, G.K., Picard, D., Tsybakov, A. (2002). Oracle inequalities for inverse problems. *Annals of Statistics*, **30**, n.3, 843-874.

[9] Cavalier L., Hengartner N.W. (2005). Adaptive estimation for inverse problems with noisy operators. *Inverse Problems* **21**, 1345–1361.

[10] Cavalier L., Tsybakov, A. (2001). Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Mathematical Methods of Statistics*, **10**, 247-282.

[11] Cavalier L., Tsybakov, A. (2002). Sharp adaptation for inverse problems with random noise. *Probability Theory and Related Fields*, **123**, 323-354.

[12] Goldenshluger A. (1999). On pointwise adaptive nonparametric deconvolution. *Bernoulli* **5**, 907-925.

[13] Goldenshluger A., Pereverzev S. (2000). Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations. *Probab. Theory and Related Fields* **118**, 169-186.

[14] Goldenshluger A., Pereverzev S. (2003). On adaptive inverse estimation of linear functionals in Hilbert scales. *Bernoulli* **9**(5), 783–807.

[15] Golubev G.K. (2004). The Method of Risk Envelope in Estimation of Linear Functionals. *Problems of Information Transmission*, **40**, No. 1, 2004, pp. 53–65. Translated from *Problemy Peredachi Informatsii*, No. 1, 2004, pp. 58–72.

[16] Juditsky, A., Rigollet, P., and Tsybakov, A. (2006). Learning by mirror averaging. *Annals of Statistics*, to appear.

[17] Lepskii, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35**, No. 3, 454–466. Translated from *Teor. Veroyatnost. i Primenen.* 35 (1990), no. 3, 459–470.

[18] Mathé P., Pereverzev S. V. (2001) Optimal discretization of inverse problems in Hilbert scales. Regularization and self-regularization of projection methods, *SIAM J. Numer. Anal.* **38**, No. 6, pp. 1999–2021.

[19] Tsybakov A.B.(2000) Adaptive estimation for inverse problems: a logarithmic effect in L_2 . *C.R. Acad. Sci. Paris*, ser. 1, t.330, 835-840.

Weierstrass-Institute and
Humboldt University Berlin,
Mohrenstr. 39, 10117 Berlin, Germany
E-mail: spokoiny@wias-berlin.de

Modal'X and Université
Européenne de Bretagne,
Campus de Ker Lann,
Rue Blaise Pascal-BP 37203,
35172 BRUZ Cedex, France
E-mail: celine.vial@univ-rennes1.fr

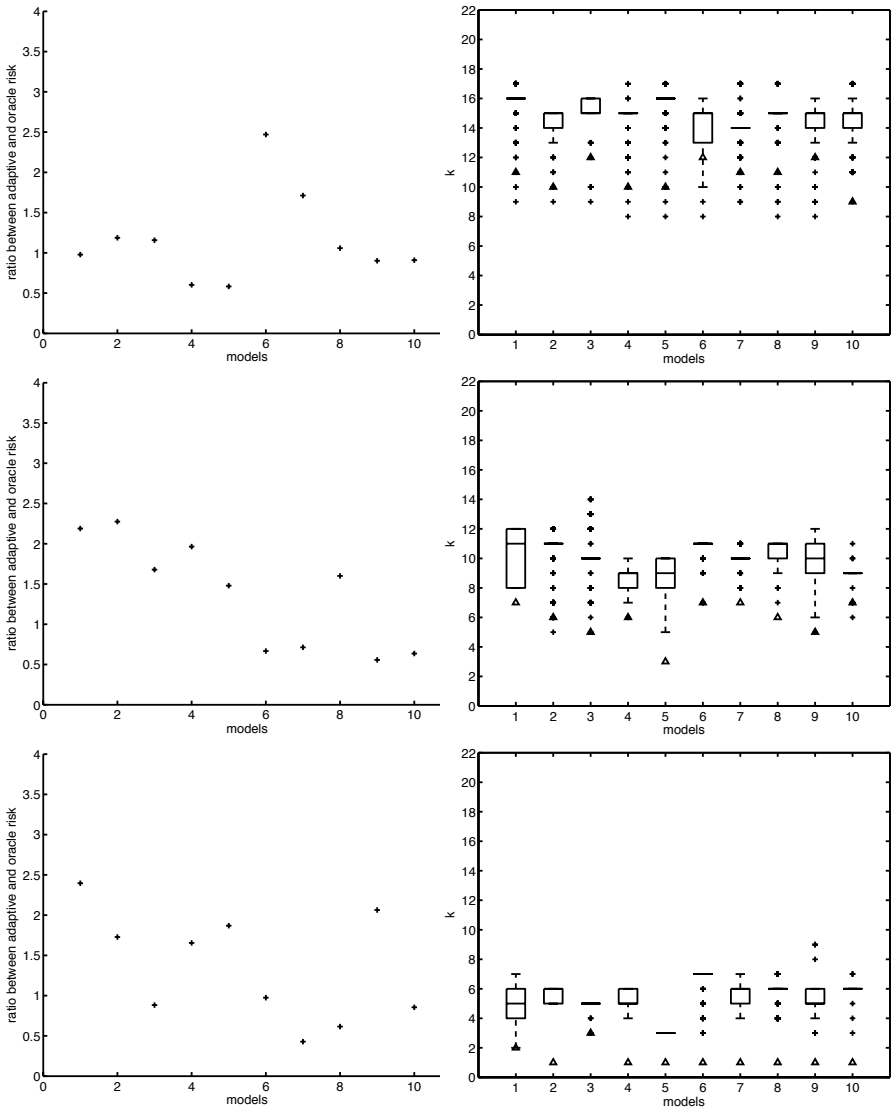
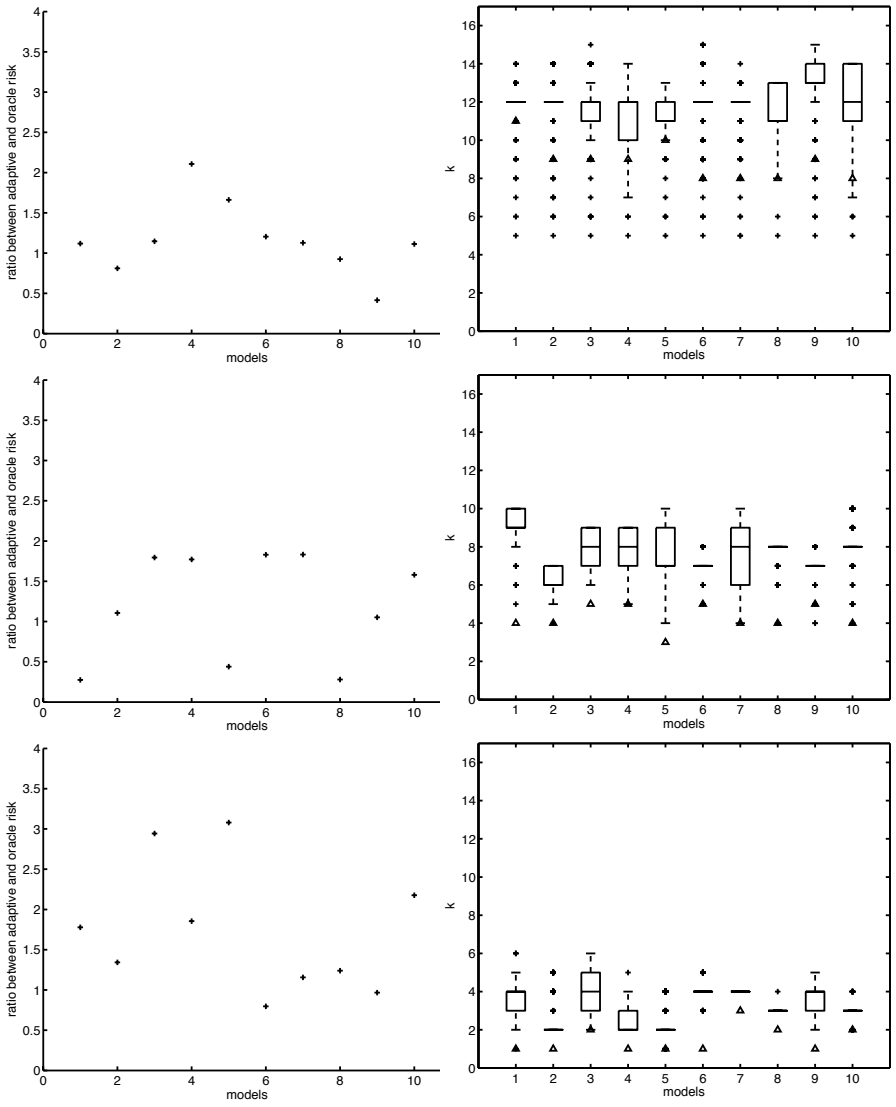


FIG 1. *The first line of 2 graphics are results in the case of the error level* $\delta = 10^{-4}$ *, the second for* $\delta = 10^{-5}$ *and the third for* $\delta = 10^{-6}$ *. The left panel (a) draws the ratio of the adaptive risk* $\boldsymbol{E}\left|\widehat{\theta} - \theta\right|$ *divided by the "oracle" risk* $\boldsymbol{E}\left|\widetilde{\theta}_{k^*} - \theta\right|$ *as function of the model. The right panel (b) draws for each of the ten models model the box plot of the adaptive estimate* $\widehat{k}$ *after 500 iterations and the triangle corresponding to the "oracle " estimate* k^* *.*



(a) (b)

FIG 2. *The first line of 2 graphics are results for an error level* $\delta = 10^{-4}$ *, the second for* $\delta = 10^{-5}$ *and the third for* $\delta = 10^{-6}$ *. The left panel (a) draws the ratio of the adaptive risk* $\boldsymbol{E}\left|\widehat{\theta} - \theta\right|$ *devided by the "oracle" risk* $\boldsymbol{E}\left|\widetilde{\theta}_{k^*} - \theta\right|$ *as function of the model. The right panel (b) draws for each model the boxplot of the adaptive estimate* $\widehat{k}$ *using* 500 *iterations and the triangle corresponding to the "oracle " value* k^* *.*

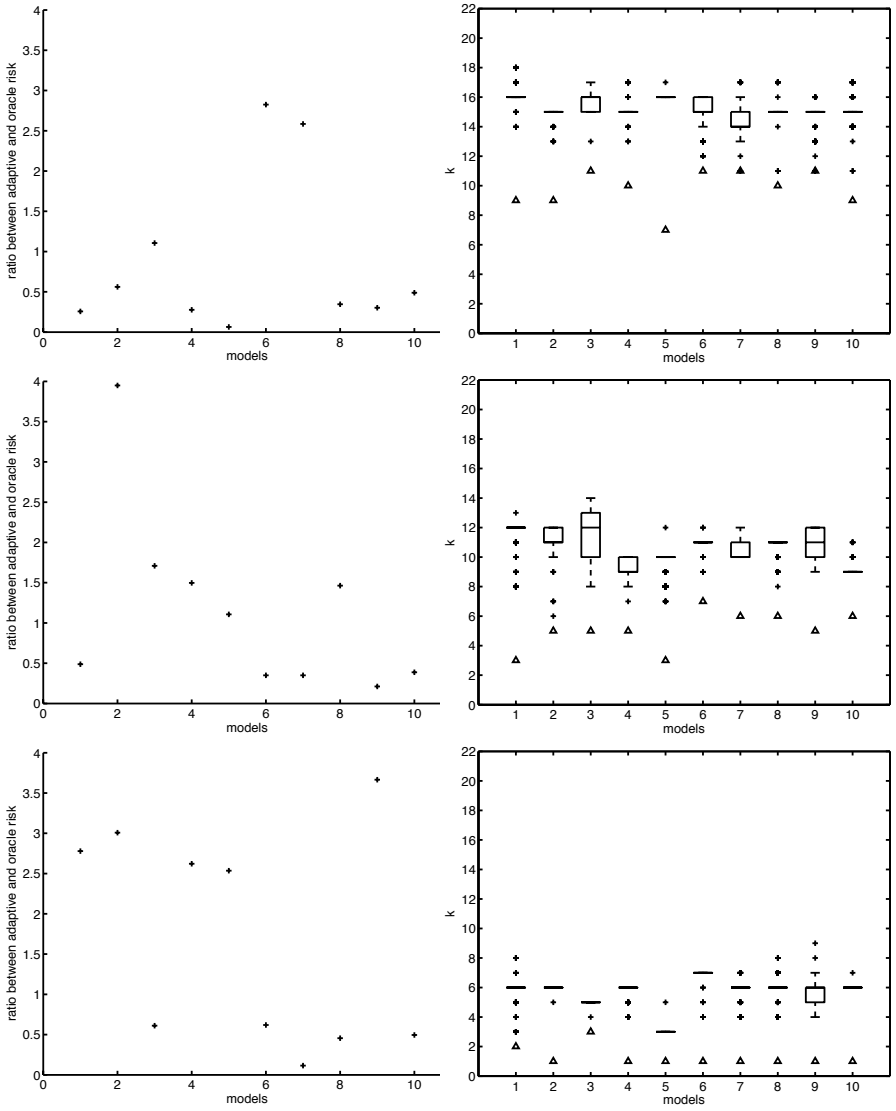


FIG 3. *The first line of 2 graphics are results for an error level* $\delta = 10^{-4}$ *, the second for* $\delta = 10^{-5}$ *and the third for* $\delta = 10^{-6}$ *. The left panel (a) draws the ratio of the adaptive risk* $\boldsymbol{E}\left|\widehat{\theta} - \theta\right|^2$ *devided by the "oracle" risk* $\boldsymbol{E}\left|\widetilde{\theta}_{k^*} - \theta\right|^2$ *as function of the model. The right panel (b) draws for each model the boxplot of the adaptive estimate* $\widehat{k}$ *using* 500 *iterations and the triangle corresponding to the "oracle " value* k^* *.*

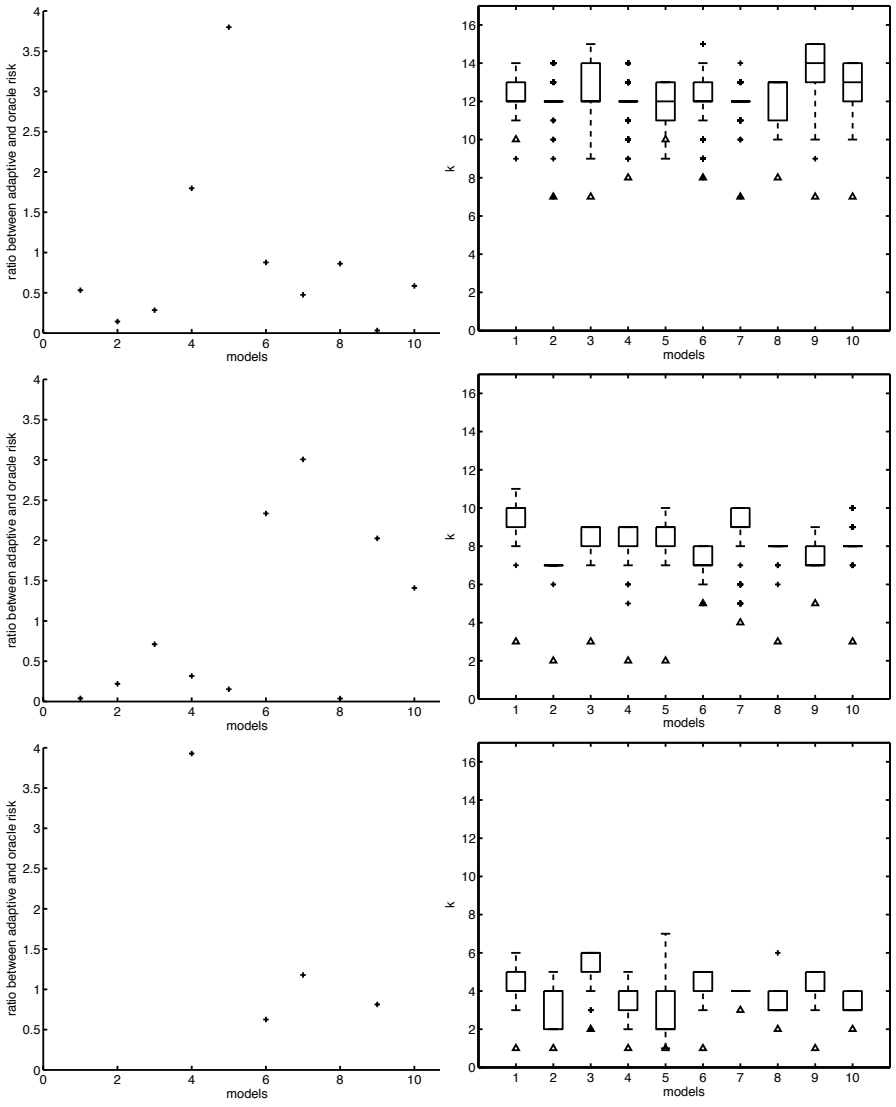


FIG 4. *The first line of 2 graphics are results for an error level* $\delta = 10^{-4}$ *, the second for* $\delta = 10^{-5}$ *and the third for* $\delta = 10^{-6}$ *. The left panel (a) draws the ratio of the adaptive risk* $\boldsymbol{E}\left|\widehat{\theta} - \theta\right|^2$ *devided by the "oracle" risk* $\boldsymbol{E}\left|\widetilde{\theta}_{k^*} - \theta\right|^2$ *as function of the model. The right panel (b) draws for each model the boxplot of the adaptive estimate* $\widehat{k}$ *using* 500 *iterations and the triangle corresponding to the "oracle " value* k^* *.*