

# Statistical Inference for the Optimal Approximating Model

Angelika Rohde

*Universität Hamburg  
Department Mathematik  
Bundesstraße 55  
D-20146 Hamburg  
Germany*

*e-mail: angelika.rohde@math.uni-hamburg.de*

and

Lutz Dümbgen

*Universität Bern  
Institut für Mathematische Statistik  
und Versicherungslehre  
Sidlerstrasse 5  
CH-3012 Bern  
Switzerland*

*e-mail: duembgen@stat.unibe.ch*

*Universität Hamburg and Universität Bern*

**Abstract:** In the setting of high-dimensional linear models with Gaussian noise, we investigate the possibility of confidence statements connected to model selection. Although there exist numerous procedures for adaptive (point) estimation, the construction of adaptive confidence regions is severely limited (cf. Li, 1989). The present paper sheds new light on this gap. We develop exact and adaptive confidence regions for the best approximating model in terms of risk. One of our constructions is based on a multiscale procedure and a particular coupling argument. Utilizing exponential inequalities for noncentral  $\chi^2$ -distributions, we show that the risk and quadratic loss of all models within our confidence region are uniformly bounded by the minimal risk times a factor close to one.

**AMS 2000 subject classifications:** 62G15, 62G20.

**Keywords and phrases:** Adaptivity, confidence regions, coupling, exponential inequality, model selection, multiscale inference, risk optimality.

## 1. Introduction

When dealing with a high dimensional observation vector, the natural question arises whether the data generating process can be approximated by a model of substantially lower dimension. Typically the models under consideration are characterized by the non-zero components of some parameter vector, and es-

pecially the presence of some approximately sparse parametrization found recently substantial interest in the literature. Sometimes consistent estimation of the so-called sparsity pattern (the locations of the non-zero components, i.e. the true model) is one of the central goals. However, consistently estimating the true model requires the rather idealistic situation that each component is either equal to zero or has sufficiently large modulus: A tiny perturbation of the parameter vector may result in the biggest model, so the question about the true model does not seem to be adequate in general. Instead of focussing on the true model one could aim for parsimonious ones which still contain the essential information and are easier to interpret. However there may exist several and quite different models which explain the data comparably well. This leads to the question which models are definitely inferior to others with a given confidence. The present paper is concerned with confidence regions for those approximating models which are optimal in terms of risk.

Suppose that we observe a random vector  $X_n = (X_{in})_{i=1}^n$  with distribution  $\mathcal{N}_n(\theta_n, \sigma^2 I_n)$ , where the mean vector  $\theta_n$  is unknown while the noise level is assumed to be known for the moment. Often the signal  $\theta_n$  represents coefficients of an unknown smooth function with respect to a given orthonormal basis of functions. There is a vast amount of literature on point estimation of  $\theta_n$ . For a given estimator  $\hat{\theta}_n = \hat{\theta}_n(X_n, \hat{\sigma}_n)$  for  $\theta_n$ , let

$$L(\hat{\theta}_n, \theta_n) := \|\hat{\theta}_n - \theta_n\|^2 \quad \text{and} \quad R(\hat{\theta}_n, \theta_n) := \mathbb{E}L(\hat{\theta}_n, \theta_n)$$

be its quadratic loss and the corresponding risk, respectively. Here  $\|\cdot\|$  denotes the standard Euclidean norm of vectors. Various adaptivity results are known for this setting, often in terms of oracle inequalities. A typical result reads as follows: Let  $(\check{\theta}_n^{(c)})_{c \in \mathcal{C}_n}$  be a family of candidate estimators  $\check{\theta}_n^{(c)} = \check{\theta}_n^{(c)}(X_n)$  for  $\theta_n$ . Then there exist estimators  $\hat{\theta}_n$  and constants  $A_n = 1 + o(1)$ ,  $B_n = O(\log(n)^\gamma)$  with  $\gamma \geq 0$  such that for arbitrary  $\theta_n$  in a certain set  $\Theta_n \subset \mathbb{R}^n$ ,

$$R(\hat{\theta}_n, \theta_n) \leq A_n \inf_{c \in \mathcal{C}_n} R(\check{\theta}_n^{(c)}, \theta_n) + B_n \sigma^2.$$

Results of this type are provided, for instance, by Polyak and Tsybakov (1991) and Donoho and Johnstone (1994, 1995, 1998), in the framework of Gaussian model selection by Birgé and Massart (2001). The latter article copes in partic-

ular with the fact that a model is not necessarily true. Further results of this type, partly in different settings, have been provided by Stone (1984), Lepski et al. (1997), Efremovich (1998) and Cai (1999, 2002), to mention just a few.

By way of contrast, when aiming at adaptive confidence sets one faces severe limitations. Here is a result of Li (1989), slightly rephrased: Suppose that  $\Theta_n$  contains a closed Euclidean ball  $B(\theta_n^o, cn^{1/4})$  around some vector  $\theta_n^o \in \mathbb{R}^n$  with radius  $cn^{1/4} > 0$ . Let  $\hat{D}_n = \hat{D}_n(X_n) \subset \Theta_n$  be a  $(1 - \alpha)$ -confidence set for  $\theta_n \in \Theta_n$ . Such a confidence set may be used as a test of the (Bayesian) null hypothesis that  $\theta_n$  is uniformly distributed on the sphere  $\partial B(\theta_n^o, cn^{1/4})$  versus the alternative that  $\theta_n = \theta_n^o$ : We reject this null hypothesis at level  $\alpha$  if  $\|\eta - \theta_n^o\| < cn^{1/4}$  for all  $\eta \in \hat{D}_n$ . Since this test cannot have larger power than the corresponding Neyman-Pearson test,

$$\begin{aligned} \mathbb{P}_{\theta_n^o} \left( \sup_{\eta \in \hat{D}_n} \|\eta - \theta_n^o\|_n < cn^{1/4} \right) &\leq \mathbb{P} \left( S_n^2 \leq \chi_{n;\alpha}^2 (c^2 n^{1/2} / \sigma^2) \right) \\ &= \Phi \left( \Phi^{-1}(\alpha) + 2^{-1/2} c^2 / \sigma^2 \right) + o(1), \end{aligned}$$

where  $S_n^2 \sim \chi_n^2$  and  $\chi_{n;\alpha}^2(\delta^2)$  stands for the  $\alpha$ -quantile of the noncentral chi-squared distribution with  $n$  degrees of freedom and noncentrality parameter  $\delta^2$ . Throughout this paper, asymptotic statements refer to  $n \rightarrow \infty$ . The previous inequality entails that no reasonable confidence set has a diameter of order  $o_p(n^{1/4})$  uniformly over the parameter space  $\Theta_n$ , as long as the latter is sufficiently large. Despite these limitations, there is some literature on confidence sets in the present or similar settings; see for instance Beran (1996, 2000), Beran and Dümbgen (1998) and Genovese and Wassermann (2005).

Improving the rate of  $O_p(n^{1/4})$  is only possible via additional constraints on  $\theta_n$ , i.e. considering substantially smaller sets  $\Theta_n$ . For instance, Baraud (2004) developed nonasymptotic confidence regions which perform well on finitely many linear subspaces. Juditsky and Lambert-Lacroix (2003) develop adaptive  $L_2$ -confidence balls for a regression function in fixed design Gaussian regression via unbiased risk estimates within the scale of Besov spaces if it is known a priori that the function belongs to a certain Besov ball. Robins and van der Vaart (2006) construct confidence balls via sample splitting which adapt to some extent to the unknown ‘‘smoothness’’ of  $\theta_n$ . In their context,  $\Theta_n$  corresponds to

a Sobolev smoothness class with given parameter  $(\beta, L)$ . However, adaptation in this context is possible only within a range  $[\beta, 2\beta]$ . Independently, Cai and Low (2006) treat the same problem in the special case of the Gaussian white noise model, obtaining the same kind of adaptivity in the broader scale of Besov bodies. Other possible constraints on  $\theta_n$  are so-called shape constraints; see for instance Cai and Low (2007), Dümbgen (2003) or Hengartner and Stark (1995). New input to the related problem in sup-norm loss has come very recently by Giné and Nickl (2010) who demonstrate in the context of density estimation that honest confidence bands can be achieved over Hölder balls if a set of only first Baire category is removed, see also Hoffmann and Nickl (2011).

The motivation of our work is twofold. First of all, the natural question arises whether one can bridge the gap mentioned above between point estimators and confidence sets. More precisely, we would like to understand profoundly the possibility of adaptation for point estimators in terms of some confidence region for the set of all optimal candidate estimators  $\check{\theta}_n^{(c)}$ . That means, we want to construct a confidence region  $\hat{\mathcal{K}}_{n,\alpha} = \hat{\mathcal{K}}_{n,\alpha}(X_n, \hat{\sigma}_n) \subset \mathcal{C}_n$  for the set

$$\begin{aligned} \mathcal{K}_n(\theta_n) &:= \operatorname{Arg\,min}_{c \in \mathcal{C}_n} R(\check{\theta}_n^{(c)}) \\ &= \left\{ c \in \mathcal{C}_n : R(\check{\theta}_n^{(c)}, \theta_n) \leq R(\check{\theta}_n^{(c')}, \theta_n) \text{ for all } c' \in \mathcal{C}_n \right\} \end{aligned}$$

such that for arbitrary  $\theta_n \in \mathbb{R}^n$ ,

$$\mathbb{P}_{\theta_n} \left( \mathcal{K}_n(\theta_n) \subset \hat{\mathcal{K}}_{n,\alpha} \right) \geq 1 - \alpha \quad (1)$$

and

$$\left. \begin{array}{l} \max_{c \in \hat{\mathcal{K}}_{n,\alpha}} R(\check{\theta}_n^{(c)}, \theta_n) \\ \max_{c \in \hat{\mathcal{K}}_{n,\alpha}} L(\check{\theta}_n^{(c)}, \theta_n) \end{array} \right\} = O_p(A_n) \min_{c \in \mathcal{C}_n} R(\check{\theta}_n^{(c)}, \theta_n) + O_p(B_n) \sigma^2. \quad (2)$$

Solving this problem means that statistical inference about differences in the performance of estimators is possible, although inference about their risk and loss is severely limited. Our second motivation is that in some settings, selecting estimators out of a class of competing estimators entails estimating implicitly an unknown regularity, smoothness class or model for the underlying signal  $\theta_n$ , and the statistician may be interested in drawing conclusions about the model or the

data generating process itself rather than about the specific signal. Computing a confidence region for optimal estimators is particularly suitable in situations in which several good candidate estimators fit the data quite well although they look different. Here it is important not to overinterpret a single fit. This aspect of exploring various candidate estimators is not covered by the usual theory of point estimation. For a good point estimator it is sufficient to pick a candidate estimator the risk of which is close to  $\min_{c \in \mathcal{C}_n} R(\check{\theta}_n^{(c)}, \theta_n)$ . This is substantially easier than trying to cover a really optimal candidate estimator. Note also that our confidence region  $\hat{\mathcal{K}}_{n,\alpha}$  is even required to cover the whole set  $\mathcal{K}_n(\theta_n)$  rather than just some element of it, with probability at least  $1 - \alpha$ ; see also the remark at the end of Section 3.

The remainder of this paper is organized as follows. In Section 3 we develop and analyze an explicit confidence region  $\hat{\mathcal{K}}_{n,\alpha}$  related to  $\mathcal{C}_n := \{0, 1, \dots, n\}$  with candidate estimators

$$\check{\theta}_n^{(k)} := \left(1_{\{i \leq k\}} X_{in}\right)_{i=1}^n.$$

These correspond to a standard nested sequence of approximating models. For this purely data-dependent set  $\hat{\mathcal{K}}_{n,\alpha}$  we shall prove the following main result.

**Theorem 1.** *Let  $(\theta_n)_{n \in \mathbb{N}}$  be arbitrary. Then*

$$\mathbb{P}_{\theta_n} \left( \mathcal{K}_n(\theta_n) \not\subset \hat{\mathcal{K}}_{n,\alpha} \right) \leq \alpha,$$

and  $\hat{\mathcal{K}}_{n,\alpha}$  satisfies the oracle inequality

$$\begin{aligned} \max_{\check{\theta}_n^{(k)} \in \hat{\mathcal{K}}_{n,\alpha}} R_n(\check{\theta}_n^{(k)}, \theta_n) &\leq \min_{j \in \mathcal{C}_n} R_n(\check{\theta}_n^{(j)}, \theta_n) \\ &+ (4\sqrt{3} + o_p(1)) \sqrt{\sigma^2 \log(n) \min_{j \in \mathcal{C}_n} R_n(\check{\theta}_n^{(j)}, \theta_n)} \\ &+ O_p(\sigma^2 \log n). \end{aligned}$$

Note that this statement implies and is more precise than (2), where  $B_n = \log n$ . Since our result is not about the existence only but contains additionally an explicit construction of the set  $\hat{\mathcal{K}}_{n,\alpha}$  which is rather involved, the mathematical techniques of our approach are first described in a simple toy model in Section 2 for the reader's convenience. Section 4 discusses richer and rather

general families of candidate estimators. In Section 5 we discuss briefly the case of unknown  $\sigma$  and explain that the main results remain valid under moderate regularity assumptions on an estimator  $\hat{\sigma}_n$ . For a more detailed treatment of this case we refer to the technical report of Rohde and Dümbgen (2009). All proofs and auxiliary results are deferred to Section 6.

## 2. A toy problem

Suppose we observe a stochastic process  $Y = (Y(t))_{t \in [0,1]}$ , where

$$Y(t) = F(t) + W(t), \quad t \in [0, 1],$$

with an unknown fixed continuous function  $F$  on  $[0, 1]$  and a Brownian motion  $W = (W(t))_{t \in [0,1]}$ . We are interested in the set

$$\mathcal{S}(F) := \underset{t \in [0,1]}{\text{Arg min}} F(t).$$

Precisely, we want to construct a  $(1 - \alpha)$ -confidence region  $\hat{\mathcal{S}}_\alpha = \hat{\mathcal{S}}_\alpha(Y) \subset [0, 1]$  for  $\mathcal{S}(F)$  in the sense that

$$P(\mathcal{S}(F) \subset \hat{\mathcal{S}}_\alpha) \geq 1 - \alpha, \tag{3}$$

regardless of  $F$ . To construct such a confidence set we regard  $Y(s) - Y(t)$  for arbitrary different  $s, t \in [0, 1]$  as a test statistic for the null hypothesis that  $s \in \mathcal{S}(F)$ , i.e. large values of  $Y(s) - Y(t)$  give evidence for  $s \notin \mathcal{S}(F)$ .

A first and naive proposal is the set

$$\hat{\mathcal{S}}_\alpha^{\text{naive}} := \left\{ s \in [0, 1] : Y(s) \leq \min_{[0,1]} Y + \kappa_\alpha^{\text{naive}} \right\}$$

with  $\kappa_\alpha^{\text{naive}}$  denoting the  $(1 - \alpha)$ -quantile of  $\max_{[0,1]} W - \min_{[0,1]} W$ . Here is a refined method based on results of Dümbgen and Spokoiny (2001): Let  $\kappa_\alpha$  be the  $(1 - \alpha)$ -quantile of

$$\sup_{s,t \in [0,1] : s \neq t} \left( \frac{|W(s) - W(t)|}{\sqrt{|s - t|}} - \Gamma(|s - t|) \right), \tag{4}$$

where

$$\Gamma(u) := \sqrt{2 \log(e/u)} \quad \text{for } 0 < u \leq 1.$$

Then constraint (3) is satisfied by the confidence region  $\hat{\mathcal{S}}_\alpha$  which consists of all  $s \in [0, 1]$  such that

$$Y(s) \leq Y(t) + \sqrt{|s-t|}(\Gamma(|s-t|) + \kappa_\alpha) \text{ for all } t \in [0, 1].$$

To illustrate the power of this method, consider for instance a sequence of functions  $F = F_n = c_n F_o$  with positive constants  $c_n \rightarrow \infty$  and a fixed continuous function  $F_o$  with unique minimizer  $s_o$ . Suppose that

$$\lim_{t \rightarrow s_o} \frac{F_o(t) - F_o(s_o)}{|t - s_o|^\gamma} = 1$$

for some  $\gamma > 1/2$ . Then the naive confidence region satisfies only

$$\max_{t \in \hat{\mathcal{S}}_\alpha^{\text{naive}}} |t - s_o| = O_p(c_n^{-1/\gamma}), \tag{5}$$

whereas

$$\max_{t \in \hat{\mathcal{S}}_\alpha} |t - s_o| = O_p\left(\log(c_n)^{1/(2\gamma-1)} c_n^{-1/(\gamma-1/2)}\right). \tag{6}$$

### 3. Confidence regions for nested approximating models

In this section we develop the confidence regions  $\hat{\mathcal{K}}_{n,\alpha}$  in detail. As in the introduction let  $X_n = \theta_n + \epsilon_n$  denote the  $n$ -dimensional observation vector with  $\theta_n \in \mathbb{R}^n$  and  $\epsilon_n \sim \mathcal{N}_n(0, \sigma^2 I_n)$ . For any candidate estimator  $\check{\theta}_n^{(k)} = (1\{i \leq k\} X_{in})_{i=1}^n$  the loss is given by

$$L_n(k) := L(\check{\theta}_n^{(k)}, \theta_n) = \sum_{i=k+1}^n \theta_{in}^2 + \sum_{i=1}^k (X_{in} - \theta_{in})^2$$

with corresponding risk

$$R_n(k) := R(\check{\theta}_n^{(k)}, \theta_n) = \sum_{i=k+1}^n \theta_{in}^2 + k\sigma^2.$$

Model selection usually aims at estimating a candidate estimator which is optimal in terms of risk. Since the risk depends on the unknown signal and therefore is not available, the selection procedure minimizes an unbiased risk estimator instead. In the sequel, the bias-corrected risk estimator for the candidate  $\check{\theta}_n^{(k)}$  is defined as

$$\hat{R}_n(k) := \sum_{i=k+1}^n (X_{in}^2 - \sigma^2) + k\sigma^2.$$

Important for our analysis is the behavior of the centered and rescaled difference process  $D_n = (D_n(j, k))_{0 \leq j < k \leq n}$  with

$$\begin{aligned} D_n(j, k) &:= \frac{\hat{R}_n(j) - \hat{R}_n(k) - R_n(j) + R_n(k)}{\sigma^2 \sqrt{4\|\theta_n/\sigma\|^2 + 2n}} \\ &= \frac{\sum_{i=j+1}^k (X_{in}^2 - \sigma^2 - \theta_{in}^2)}{\sigma^2 \sqrt{4\|\theta_n/\sigma\|^2 + 2n}} \\ &= \frac{1}{\sqrt{4\|\theta_n/\sigma\|^2 + 2n}} \sum_{i=j+1}^k (2(\theta_{in}/\sigma)(\epsilon_{in}/\sigma) + (\epsilon_{in}/\sigma)^2 - 1). \end{aligned}$$

Hence the process  $D_n$  consists of partial sums of the independent and centered, but in general not identically distributed random variables  $2(\theta_{in}/\sigma)(\epsilon_{in}/\sigma) + (\epsilon_{in}/\sigma)^2 - 1$ . The standard deviation of  $D_n(j, k)$  is given by

$$\tau_n(j, k) := \frac{1}{\sqrt{4\|\theta_n/\sigma\|^2 + 2n}} \left( \sum_{i=j+1}^k (4\theta_{in}^2/\sigma^2 + 2) \right)^{1/2}.$$

Note that  $\tau_n(0, n) = 1$  by construction. To imitate the more powerful confidence region of Section 2 based on the multiscale approach, one needs a refined analysis of the increment process  $D_n$ . Since this process does not have subgaussian tails, the standardization is more involved than the correction in (4).

**Theorem 2.** *Define  $\Gamma_n(j, k) := \Gamma(\tau_n(j, k)^2)$  for  $0 \leq j < k \leq n$ . Then*

$$\sup_{0 \leq j < k \leq n} \frac{|D_n(j, k)|}{\tau_n(j, k)} \leq \sqrt{32} \log n + O_p(1),$$

and for any fixed  $c > 2$ ,

$$d_n := \max_{0 \leq j < k \leq n} \left( \frac{|D_n(j, k)|}{\tau_n(j, k)} - \Gamma_n(j, k) - \frac{c \cdot \Gamma_n(j, k)^2}{\sqrt{4\|\theta_n/\sigma\|^2 + 2n} \tau_n(j, k)} \right)^+$$

is bounded in probability. In the special case of  $\theta_n$  having components  $\pm\sigma$ , the random variable  $d_n$  converges in distribution to the random variable in (4).

The limiting distribution is closely related to Lévy's modulus of continuity of Brownian motion, and this indicates that the additive correction term in the definition of  $d_n$  cannot be chosen essentially smaller. It will play a crucial role for the efficiency of the confidence region.

As shown by Rohde and Dümbgen (2009), convergence in distribution of  $d_n$  holds under much weaker assumptions on the signal-to-noise vector  $\theta_n/\sigma$ .



However, to utilize this fact for inference on the set  $\mathcal{K}_n(\theta_n)$ , we are facing the problem that the auxiliary function  $\tau_n(\cdot, \cdot)$  depends on the unknown signal-to-noise vector  $\theta_n/\sigma$ . In fact, knowing  $\tau_n$  would imply knowledge of  $\mathcal{K}_n(\theta_n)$  already. One could try to estimate the variances  $\tau_n(j, k)^2$ ,  $j < k$ , by

$$\hat{\tau}_n(j, k)^2 := \left\{ \sum_{i=1}^n \left( 4(X_{in}^2/\sigma^2 - 1) + 2 \right) \right\}^{-1} \sum_{i=j+1}^k \left( 4(X_{in}^2/\sigma^2 - 1) + 2 \right).$$

However, using such an estimator does not seem to work since

$$\sup_{0 \leq j < k \leq n} \left| \frac{\hat{\tau}_n(j, k)}{\tau_n(j, k)} - 1 \right| \not\rightarrow_p 0$$

as  $n$  goes to infinity. This can be verified by noting that the (rescaled) numerator of  $(\hat{\tau}_n(j, k)^2)_{0 \leq j < k \leq n}$  is essentially, up to centering, of the same structure as the rescaled difference process  $D_n$  itself. These difficulties may be overcome with a trick described next.

### ***The least favourable case of constant risk***

The problem of estimating the set  $\text{Arg min}_k R_n(k)$  can be cast into our toy model where  $Y(t)$ ,  $F(t)$  and  $W(t)$  correspond to  $\hat{R}_n(k)$ ,  $R_n(k)$  and the difference  $\hat{R}_n(k) - R_n(k)$ , respectively. One may expect that the more distinctive the global minima are, the easier it is to identify their location. Hence the case of constant risks appears to be least favourable, corresponding to a signal

$$\theta_n^* := (\pm\sigma)_{i=1}^n,$$

In this situation, each candidate estimator  $\check{\theta}_n^{(k)}$  has the same risk of  $n\sigma^2$ .

A related consideration leading to an explicit procedure is as follows: For fixed indices  $0 \leq j < k \leq n$ ,

$$R_n(j) - R_n(k) = \sum_{i=j+1}^k \theta_{in}^2 - (k-j)\sigma^2,$$

and the test statistic

$$T_{jkn} := \sum_{i=j+1}^k X_{in}^2/\sigma^2 = 2(k-j) - (\hat{R}_n(k) - \hat{R}_n(j))/\sigma^2$$

has a noncentral  $\chi^2$  distribution

$$\chi_{k-j}^2 \left( \sum_{i=j+1}^k \theta_{in}^2 / \sigma^2 \right) = \chi_{k-j}^2 \left( k - j + (R_n(j) - R_n(k)) / \sigma^2 \right).$$

Thus large or small values of  $T_{jkn}$  give evidence for  $R_n(j)$  being larger or smaller, respectively, than  $R_n(k)$ . Precisely,

$$\mathcal{L}_{\theta_n}(T_{jkn}) \begin{cases} \leq_{\text{st.}} \mathcal{L}_{\theta_n^*}(T_{jkn}) & \text{whenever } j \in \mathcal{K}_n(\theta_n), \\ \geq_{\text{st.}} \mathcal{L}_{\theta_n^*}(T_{jkn}) & \text{whenever } k \in \mathcal{K}_n(\theta_n). \end{cases}$$

Via a suitable construction involving Poisson mixtures of central  $\chi^2$ -distributed random variables, this pointwise stochastic ordering can be extended to a coupling for the whole process  $(T_{jkn})_{0 \leq j < k \leq n}$ :

**Proposition 3** (Coupling). *For any  $\theta_n \in \mathbb{R}^n$  there exists a probability space with random variables  $(\tilde{T}_{jkn})_{0 \leq j < k \leq n}$  and  $(\tilde{T}_{jkn}^*)_{0 \leq j < k \leq n}$  such that*

$$\begin{aligned} \mathcal{L}\left((\tilde{T}_{jkn})_{0 \leq j < k \leq n}\right) &= \mathcal{L}_{\theta_n}\left((T_{jkn})_{0 \leq j < k \leq n}\right), \\ \mathcal{L}\left((\tilde{T}_{jkn}^*)_{0 \leq j < k \leq n}\right) &= \mathcal{L}_{\theta_n^*}\left((T_{jkn})_{0 \leq j < k \leq n}\right), \end{aligned}$$

and for arbitrary indices  $0 \leq j < k \leq n$ ,

$$\tilde{T}_{jkn} \begin{cases} \leq \tilde{T}_{jkn}^* & \text{whenever } j \in \mathcal{K}_n(\theta_n), \\ \geq \tilde{T}_{jkn}^* & \text{whenever } k \in \mathcal{K}_n(\theta_n). \end{cases}$$

By means of Proposition 3 we can define a confidence set for  $\mathcal{K}_n(\theta_n)$ , based on the least favourable case  $\theta_n = \theta_n^*$ . Let  $\kappa_{n,\alpha}$  denote the  $(1 - \alpha)$ -quantile of  $\mathcal{L}_{\theta_n^*}(d_n)$ , where for simplicity  $c := 3$  in the definition of  $d_n$ . Note also that  $\tau_n(j, k)^2 = (k - j)/n$  in case of  $\theta_n = \theta_n^*$ . Motivated by Theorem 2, we define

$$\begin{aligned} \hat{\mathcal{K}}_{n,\alpha} &:= \{j : \hat{R}_n(j) \leq \hat{R}_n(k) + \sigma^2 c_{jkn} \text{ for all } k \neq j\} \\ &= \{j : T_{ijn} \geq 2(j - i) - c_{ijn} \text{ for all } i < j, \\ &\quad T_{jkn} \leq 2(k - j) + c_{jkn} \text{ for all } k > j\} \end{aligned} \tag{7}$$

with

$$c_{jkn} = c_{jkn,\alpha} := \sqrt{6|k - j|} \left( \Gamma\left(\frac{|k - j|}{n}\right) + \kappa_{n,\alpha} \right) + 3\Gamma\left(\frac{|k - j|}{n}\right)^2.$$

With this construction we obtain an extended version of Theorem 1 from the introduction:

**Theorem 4.** *Let  $(\theta_n)_{n \in \mathbb{N}}$  be arbitrary. With  $\hat{\mathcal{K}}_{n,\alpha}$  as defined above,*

$$\mathbb{P}_{\theta_n}(\mathcal{K}_n(\theta_n) \not\subset \hat{\mathcal{K}}_{n,\alpha}) \leq \alpha.$$

*The critical values  $\kappa_{n,\alpha}$  converge to  $\kappa_\alpha$  introduced in Section 2, and the confidence regions  $\hat{\mathcal{K}}_{n,\alpha}$  satisfy the oracle inequalities*

$$\begin{aligned} \max_{k \in \hat{\mathcal{K}}_{n,\alpha}} R_n(k) &\leq \min_{j \in \mathcal{C}_n} R_n(j) + (4\sqrt{3} + o_p(1)) \sqrt{\sigma^2 \log(n) \min_{j \in \mathcal{C}_n} R_n(j)} \\ &\quad + O_p(\sigma^2 \log n) \end{aligned} \quad (8)$$

and

$$\max_{k \in \hat{\mathcal{K}}_{n,\alpha}} \sqrt{L_n(k)} \leq \min_{j \in \mathcal{C}_n} \sqrt{L_n(j)} + O_p(\sqrt{\sigma^2 \log n}). \quad (9)$$

The upper bounds in this theorem are of the form

$$\sqrt{\rho_n} \left( 1 + O_p(\sqrt{\sigma^2 \log(n)/\rho_n}) \right)$$

with  $\rho_n$  denoting minimal risk or minimal loss. Thus the maximal risk (loss) over  $\hat{\mathcal{K}}_{n,\alpha}$  exceeds the minimal risk (loss) only by a factor close to one, provided that the minimal risk (loss) is substantially larger than  $\sigma^2 \log n$ .

**Remark (Dependence on  $\alpha$ ).** The proof reveals a refined version of the bounds in Theorem 4 in case of signals  $\theta_n$  such that

$$\left( \min_{j \in \mathcal{C}_n} R_n(j) \right)^{-1} = O(\log(n)^{-3}).$$

Let  $0 < \alpha(n) \rightarrow 0$  such that  $\kappa_{n,\alpha(n)}^6 = O(\min_{j \in \mathcal{C}_n} R_n(j))$ . Then

$$\begin{aligned} \max_{k \in \hat{\mathcal{K}}_{n,\alpha}} R_n(k) &\leq \min_{j \in \mathcal{C}_n} R_n(j) \\ &\quad + \left( 4\sqrt{3}\sqrt{\log n} + 2\sqrt{6}\kappa_{n,\alpha} + O_p(1) \right) \sqrt{\sigma^2 \min_{j \in \mathcal{C}_n} R_n(j)} \end{aligned}$$

uniformly in  $\alpha \geq \alpha(n)$ .

**Remark (Point estimation versus confidence regions).** As stated in the introduction, the construction of a confidence region for  $\mathcal{K}_n(\theta_n)$  is more

ambitious than the construction of an adaptive point estimator for  $\theta_n$ . To see this, suppose that the true signal vector  $\theta_n$  satisfies

$$|\theta_{in}| \begin{cases} \gg \sigma & \text{for } i \leq j_n \\ \in [\sigma - c_n, \sigma + c_n] & \text{for } j_n < i \leq k_n \\ \ll \sigma & \text{for } i > k_n \end{cases}$$

with indices  $1 \leq j_n < k_n \leq n$  such that  $k_n - j_n \rightarrow \infty$  and arbitrarily small constants  $c_n > 0$  tending to zero. Constructing an almost optimal point estimator (based on the given candidates) requires to pick a candidate estimator  $\check{\theta}_n^{(k)}$  with  $j_n \leq k \leq k_n$ . However, depending on the precise values of  $|\theta_{in}|$  for  $j_n < i \leq k_n$ , the set  $\mathcal{K}_n(\theta_n)$  may be any given nonvoid subset of  $\{j_n, \dots, k_n\}$ , see also the proof of Proposition 3 and Figure 1. Hence it may happen with asymptotically positive probability that the point estimator uses a candidate  $\check{\theta}_n^{(k)}$  with  $k \notin \mathcal{K}_n(\theta_n)$ . By way of contrast, if  $c_n$  is small, the confidence region  $\hat{\mathcal{K}}_{n,\alpha}$  will contain  $\{j_n, \dots, k_n\}$  with probability close to or higher than  $1 - \alpha$  and thus indicate that there are many candidate estimators of comparable quality.

#### 4. Confidence sets in case of larger families of candidates

The previous result relies strongly on the assumption of nested models. It is possible to obtain confidence sets for the optimal approximating models in a more general setting, albeit the resulting oracle property is not as strong as in the nested case. In particular, we can no longer rely on a coupling result but need a different construction.

Let  $\mathcal{C}_n$  be a family of index sets  $C \subset \{1, 2, \dots, n\}$  with candidate estimators

$$\check{\theta}^{(C)} := \left(1_{\{i \in C\}} X_{in}\right)_{i=1}^n$$

and corresponding risks

$$R_n(C) := R(\check{\theta}^{(C)}, \theta_n) = \sum_{i \notin C} \theta_{in}^2 + |C| \sigma^2,$$

where  $|S|$  denotes the cardinality of a set  $S$ . For two index sets  $C$  and  $D$ ,

$$\sigma^{-2}(R_n(D) - R_n(C)) = \delta_n^2(C \setminus D) - \delta_n^2(D \setminus C) + |D| - |C|$$

with the auxiliary quantities

$$\delta_n^2(J) := \sum_{i \in J} \theta_{in}^2 / \sigma^2, \quad J \subset \{1, 2, \dots, n\}.$$

Hence we aim at simultaneous  $(1 - \alpha)$ -confidence intervals for these noncentrality parameters  $\delta_n(J)$ , where  $J \in \mathcal{M}_n := \{D \setminus C : C, D \in \mathcal{C}_n\}$ . To this end we utilize the fact that

$$T_n(J) := \frac{1}{\sigma^2} \sum_{i \in J} X_{in}^2$$

has a  $\chi_{|J|}^2(\delta_n^2(J))$ -distribution. We denote the distribution function of  $\chi_k^2(\delta^2)$  by  $F_k(\cdot | \delta^2)$ . Now let  $M_n := |\mathcal{M}_n| - 1 \leq |\mathcal{C}_n|(|\mathcal{C}_n| - 1)$ , the number of nonvoid index sets  $J \in \mathcal{M}_n$ . Then with probability at least  $1 - \alpha$ ,

$$\alpha / (2M_n) \leq F_{|J|}(T_n(J) | \delta_n^2(J)) \leq 1 - \alpha / (2M_n) \quad \text{for } \emptyset \neq J \in \mathcal{M}_n. \quad (10)$$

Since  $F_{|J|}(T_n(J) | \delta^2)$  is strictly decreasing in  $\delta^2$  with limit 0 as  $\delta^2 \rightarrow \infty$ , (10) entails the simultaneous  $(1 - \alpha)$ -confidence intervals  $[\hat{\delta}_{n,\alpha,l}^2(J), \hat{\delta}_{n,\alpha,u}^2(J)]$  for all parameters  $\delta_n^2(J)$  as follows: We set  $\hat{\delta}_{n,\alpha,l}^2(\emptyset) := \hat{\delta}_{n,\alpha,u}^2(\emptyset) := 0$ , while for nonvoid  $J$ ,

$$\hat{\delta}_{n,\alpha,l}^2(J) := \min \left\{ \delta^2 \geq 0 : F_{|J|}(T_n(J) | \delta^2) \leq 1 - \alpha / (2M_n) \right\}, \quad (11)$$

$$\hat{\delta}_{n,\alpha,u}^2(J) := \max \left\{ \delta^2 \geq 0 : F_{|J|}(T_n(J) | \delta^2) \geq \alpha / (2M_n) \right\}. \quad (12)$$

By means of these bounds, we may claim with confidence  $1 - \alpha$  that for arbitrary  $C, D \in \mathcal{C}_n$  the normalized difference  $(n/\sigma^2)(R_n(D) - R_n(C))$  is at most  $\hat{\delta}_{n,\alpha,u}^2(C \setminus D) - \hat{\delta}_{n,\alpha,l}^2(D \setminus C) + |D| - |C|$ . Thus a  $(1 - \alpha)$ -confidence set for  $\mathcal{K}_n(\theta_n) = \text{Arg min}_{C \in \mathcal{C}_n} R_n(C)$  is given by

$$\hat{\mathcal{K}}_{n,\alpha} := \left\{ C \in \mathcal{C}_n : \hat{\delta}_{n,\alpha,u}^2(C \setminus D) - \hat{\delta}_{n,\alpha,l}^2(D \setminus C) + |D| - |C| \geq 0 \text{ for all } D \in \mathcal{C}_n \right\}.$$

These confidence sets  $\hat{\mathcal{K}}_{n,\alpha}$  satisfy the following oracle inequalities:

**Theorem 5.** *Let  $(\theta_n)_{n \in \mathbb{N}}$  be arbitrary, and suppose that  $\log |\mathcal{C}_n| = o(n)$ . Then*

$$\max_{C \in \hat{\mathcal{K}}_{n,\alpha}} \sqrt{R_n(C)} \leq \min_{D \in \mathcal{C}_n} \sqrt{R_n(D)} + O_p(\sqrt{\sigma^2 \log |\mathcal{C}_n|})$$

and

$$\max_{C \in \hat{\mathcal{K}}_{n,\alpha}} \sqrt{L_n(C)} \leq \min_{D \in \mathcal{C}_n} \sqrt{L_n(D)} + O_p(\sqrt{\sigma^2 \log |\mathcal{C}_n|}).$$

The upper bounds in this theorem are of the form

$$\sqrt{\rho_n} \left( 1 + O_p \left( \sqrt{\sigma^2 \log(|\mathcal{C}_n|) / \rho_n} \right) \right)$$

with  $\rho_n$  denoting minimal risk or minimal loss. This is analogous to the setting of nested models, where  $\log n$  is replaced with  $\log |\mathcal{C}_n|$ . Again, the maximal risk (loss) over  $\hat{\mathcal{K}}_{n,\alpha}$  exceeds the minimal risk (loss) only by a factor close to one, provided that the minimal risk (loss) is substantially larger than  $\sigma^2 \log |\mathcal{C}_n|$ .

**Remark (Suboptimality in case of nested models).** In case of nested models, the general construction in this section is suboptimal. For if one follows the proof carefully and uses  $\sigma^2 \log |\mathcal{C}_n| = 2\sigma^2 \log n + O(1)$  in this special setting, one obtains the refined inequality

$$\begin{aligned} \max_{k \in \hat{\mathcal{K}}_{n,\alpha}} R_n(k) &\leq \min_{j \in \mathcal{C}_n} R_n(j) + (4\sqrt{8} + o_p(1)) \sqrt{\sigma^2 \log(n) \min_{j \in \mathcal{C}_n} R_n(j)} \\ &\quad + O_p(\sigma^2 \log n), \end{aligned}$$

so the multiplier of the term  $\sqrt{\min_j R_n(j)}$  is larger than the one in Theorem 4. The intrinsic reason is that the general procedure does not assume any structure of the family of candidate estimators. Hence advanced multiscale theory is not applicable.

## 5. The impact of estimating the noise level

We discuss briefly the extension of our results to the case of unknown noise variance. It is assumed subsequently that a variance estimator  $\hat{\sigma}_n^2$  satisfying the subsequent condition (A) is available.

(A)  $\hat{\sigma}_n^2$  and  $X_n$  are stochastically independent with

$$\frac{m\hat{\sigma}_n^2}{\sigma^2} \sim \chi_m^2,$$

where  $m = m_n \geq 1$  satisfies

$$\beta_n^2 := \frac{2n}{m_n} = O(1).$$

**Example.** Suppose that we observe  $Y = M\eta + \delta$  with given design matrix  $M \in \mathbb{R}^{(n+m) \times n}$  of rank  $n$ , unknown parameter vector  $\eta \in \mathbb{R}^n$  and unobserved error vector  $\delta \sim \mathcal{N}_{n+m}(0, \sigma^2 I_{n+m})$ . Then the previous assumptions are satisfied by  $X_n := (M^\top M)^{1/2} \hat{\eta}$  with the least squares estimator  $\hat{\eta} := (M^\top M)^{-1} M^\top Y$  and  $\hat{\sigma}_n^2 := \|Y - M\hat{\eta}\|^2/m$ , where  $\theta_n := (M^\top M)^{1/2} \eta$ .

Assumption (A) implies the following weaker condition:

(A')  $\hat{\sigma}_n^2$  and  $X_n$  are stochastically independent such that for constants  $0 < \beta_n = O(1)$ ,

$$\sqrt{n}(\hat{\sigma}_n^2/\sigma^2 - 1)/\beta_n \rightarrow_{\mathcal{L}} \mathcal{N}(0, 1).$$

This condition covers, for instance, estimators of  $\sigma$  used in connection with wavelets. There  $\sigma$  is estimated by the median of some very high frequency wavelet coefficients divided by the normal quantile  $\Phi^{-1}(3/4)$ , whereas the signal  $\theta_n$  corresponds only to the other wavelet coefficients.

**Nested models.** In the setting of Section 3, the modified bias-corrected risk estimator for the candidate  $\check{\theta}_n^{(k)}$  is redefined as

$$\hat{R}_n(k) := \sum_{i=k+1}^n (X_{in}^2 - \hat{\sigma}_n^2) + k\hat{\sigma}_n^2,$$

and we consider  $T_{jkn} := \sum_{i=j+1}^k X_{in}^2/\hat{\sigma}_n^2$ . Now

$$\begin{aligned} \hat{D}_n(j, k) &:= \frac{\hat{R}_n(j) - \hat{R}_n(k) - R_n(j) + R_n(k)}{\hat{\sigma}_n^2 \sqrt{4\|\theta_n/\sigma\|^2 + 2n}} \\ &= \frac{\sigma^2}{\hat{\sigma}_n^2} (D_n(j, k) + V_n(j, k)), \end{aligned}$$

where  $D_n(j, k)$  is defined as in Section 3, while

$$V_n(j, k) := \frac{2(k-j)(1 - \hat{\sigma}_n^2/\sigma^2)}{\sqrt{4\|\theta_n/\sigma\|^2 + 2n}}.$$

Since  $\hat{\sigma}_n^2/\sigma^2 = 1 + O_p(n^{-1/2})$ , the processes  $\hat{D}_n$  and  $D_n$  behave similarly on small scales (i.e. for arguments  $(j, k)$  with  $|k-j|/n$  being small). Nevertheless the contribution of  $V_n$  is non-negligible asymptotically, unless  $\beta_n \rightarrow 0$ .

The confidence region  $\hat{\mathcal{K}}_{n,\alpha}$  is defined as before in (7) with the new versions of  $\hat{R}_n$  and  $T_{jkn}$ ,  $\sigma^2$  replaced with  $\hat{\sigma}_n^2$ , and the quantile  $\kappa_{n,\alpha}$  in the definition of

$c_{jkn}$  has to be redefined to be the  $(1 - \alpha)$ -quantile of  $\mathcal{L}_{\theta_n^*}(\hat{d}_n)$ . Here  $\hat{d}_n$  is defined as  $d_n$  with  $\hat{D}_n$  in place of  $D_n$ . Note that  $\hat{D}_n$  involves the process  $D_n$  and the ratio  $S_n^2 := (\hat{\sigma}_n/\sigma)^2$ . The latter random variable is known to be independent of  $X_n$  and to have distribution  $\chi_m^2$  under (A). In case of the weaker assumption (A'), one may replace  $S_n^2$  with a random variable with distribution  $\chi_m^2$ , where  $m := \lceil 2n/\beta_n^2 \rceil$ .

With these modifications, Theorem 4 remains true under (A) or (A'). The only modification is that  $\kappa_{n,\alpha} \not\rightarrow \kappa_\alpha$  in general, but still  $\kappa_{n,\alpha} = O(1)$ .

**General candidate families.** In the setting of Section 4, one could replace  $T_n(J)$  with  $\sum_{i \in J} X_{in}^2 / \hat{\sigma}_n^2$  which has a non-central  $F$  distribution under (A). However, this approach might be very conservative because it ignores the fact that all test statistics involve one and the same denominator  $\hat{\sigma}_n^2$ . Here is an alternative proposal: Let  $\alpha' := 1 - (1 - \alpha)^{1/2}$ . It follows from Assumption (A) that with probability  $1 - \alpha'$ ,

$$\tau_{n,\alpha,l} := \frac{m}{\chi_{m;1-\alpha'/2}} \leq \frac{\hat{\sigma}_n^2}{\sigma^2} \leq \tau_{n,\alpha,u} := \frac{m}{\chi_{m;\alpha'/2}}.$$

Under Assumption (A') this is true with asymptotic probability  $1 - \alpha'$ . Now we obtain simultaneous  $(1 - \alpha)$ -confidence bounds  $\hat{\delta}_{n,\alpha,l}^2(J)$  and  $\hat{\delta}_{n,\alpha,u}^2(J)$  as in (11) and (12) by replacing  $\alpha$  with  $\alpha'$  and  $T_n(J)$  with

$$\frac{\tau_{n,\alpha,l}}{\hat{\sigma}_n^2} \sum_{i \in J} X_{in}^2 \quad \text{and} \quad \frac{\tau_{n,\alpha,u}}{\hat{\sigma}_n^2} \sum_{i \in J} X_{in}^2,$$

respectively. The conclusions of Theorem 5 continue to hold, essentially because  $\tau_{n,\alpha,l}, \tau_{n,\alpha,u} = 1 + O(n^{-1/2})$  and  $(\hat{\sigma}_n/\sigma)^2 = 1 + O_p(n^{-1/2})$ .

## 6. Proofs

### 6.1. Proof of (5) and (6)

Note first that  $\min_{[0,1]} Y$  lies between  $F_n(s_o) + \min_{[0,1]} W$  and  $F_n(s_o) + W(s_o)$ . Hence for any  $\alpha' \in (0, 1)$ ,

$$\begin{aligned} \hat{S}_\alpha^{\text{naive}} &\subset \{s \in [0, 1] : F_n(s) + W(s) \leq F_n(s_o) + W(s_o) + \kappa_\alpha^{\text{naive}}\} \\ &\subset \{s \in [0, 1] : F_n(s) - F_n(s_o) \leq \kappa_{\alpha'}^{\text{naive}} + \kappa_\alpha^{\text{naive}}\} \\ &= \{s \in [0, 1] : F_o(s) - F_o(s_o) \leq c_n^{-1}(\kappa_{\alpha'}^{\text{naive}} + \kappa_\alpha^{\text{naive}})\} \end{aligned}$$



and

$$\begin{aligned} \hat{\mathcal{S}}_\alpha^{\text{naive}} &\supset \{s \in [0, 1] : F_n(s) + W(s) \leq F_n(s_o) + \min_{[0,1]} W + \kappa_\alpha^{\text{naive}}\} \\ &\supset \{s \in [0, 1] : F_n(s) - F_n(s_o) \leq \kappa_\alpha^{\text{naive}} - \kappa_{\alpha'}^{\text{naive}}\} \\ &= \{s \in [0, 1] : F_o(s) - F_o(s_o) \leq c_n^{-1}(\kappa_\alpha^{\text{naive}} - \kappa_{\alpha'}^{\text{naive}})\} \end{aligned}$$

with probability  $1 - \alpha'$ . Since  $\kappa_{\alpha'}^{\text{naive}} < \kappa_\alpha^{\text{naive}}$  if  $\alpha < \alpha' < 1$ , these considerations, combined with the expansion of  $F_o$  near  $s_o$ , show that the maximum of  $|s - s_o|$  over all  $s \in \hat{\mathcal{S}}_\alpha^{\text{naive}}$  is precisely of order  $O_p(c_n^{-1/\gamma})$ .

On the other hand, the confidence region  $\hat{\mathcal{S}}_\alpha$  is contained in the set of all  $s \in [0, 1]$  such that

$$F_n(s) + W(s) \leq F_n(s_o) + W(s_o) + \sqrt{|s - s_o|} \left( \sqrt{2 \log(e/|s - s_o|)} + \kappa_\alpha \right),$$

and this entails that

$$F_o(s) - F_o(s_o) \leq c_n^{-1} \sqrt{|s - s_o|} \left( \sqrt{2 \log(e/|s - s_o|)} + \kappa_\alpha + O_p(1) \right)$$

with  $O_p(1)$  not depending on  $s$ . Now the expansion of  $F_o$  near  $s_o$  entails claim (6).  $\square$

## 6.2. Exponential inequalities

An essential ingredient for our main results is an exponential inequality for quadratic functions of a Gaussian random vector. It extends inequalities of Dahlhaus and Polonik (2006) for quadratic forms and is of independent interest.

**Proposition 6.** *Let  $Z_1, \dots, Z_n$  be independent, standard Gaussian random variables. Furthermore, let  $\lambda_1, \dots, \lambda_n$  and  $\delta_1, \dots, \delta_n$  be real constants, and define  $\gamma^2 := \text{Var}\left(\sum_{i=1}^n \lambda_i (Z_i + \delta_i)^2\right) = \sum_{i=1}^n \lambda_i^2 (2 + 4\delta_i^2)$ . Then for arbitrary  $\eta \geq 0$  and  $\lambda_{\max} := \max(\lambda_1, \dots, \lambda_n, 0)$ ,*

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^n \lambda_i ((Z_i + \delta_i)^2 - (1 + \delta_i^2)) \geq \eta\gamma\right) &\leq \exp\left(-\frac{\eta^2}{2 + 4\eta\lambda_{\max}/\gamma}\right) \\ &\leq e^{1/4} \exp(-\eta/\sqrt{8}). \end{aligned}$$

Note that replacing  $\lambda_i$  in Proposition 6 with  $-\lambda_i$  yields twosided exponential inequalities. By means of Proposition 6 and elementary calculations one obtains exponential and related inequalities for noncentral  $\chi^2$  distributions:

**Corollary 7.** *For an integer  $n > 0$  and a constant  $\delta \geq 0$  let  $F_n(\cdot \mid \delta^2)$  be the distribution function of  $\chi_n^2(\delta^2)$ . Then for arbitrary  $r \geq 0$ ,*

$$F_n(n + \delta^2 + r \mid \delta^2) \geq 1 - \exp\left(-\frac{r^2}{4n + 8\delta^2 + 4r}\right), \quad (13)$$

$$F_n(n + \delta^2 - r \mid \delta^2) \leq \exp\left(-\frac{r^2}{4n + 8\delta^2}\right). \quad (14)$$

*In particular, for any  $u \in (0, 1/2)$ ,*

$$F_n^{-1}(1 - u \mid \delta^2) \leq n + \delta^2 + \sqrt{(4n + 8\delta^2) \log(u^{-1})} + 4 \log(u^{-1}), \quad (15)$$

$$F_n^{-1}(u \mid \delta^2) \geq n + \delta^2 - \sqrt{(4n + 8\delta^2) \log(u^{-1})}. \quad (16)$$

*Moreover, for any number  $\hat{\delta} \geq 0$ , the inequalities  $u \leq F_n(n + \hat{\delta}^2 \mid \delta^2) \leq 1 - u$  entail that*

$$\delta^2 - \hat{\delta}^2 \begin{cases} \leq +\sqrt{(4n + 8\hat{\delta}^2) \log(u^{-1})} + 8 \log(u^{-1}), \\ \geq -\sqrt{(4n + 8\hat{\delta}^2) \log(u^{-1})}. \end{cases} \quad (17)$$

Conclusion (17) follows from (13) and (14), applied to  $r = \hat{\delta}^2 - \delta^2$  and  $r = \delta^2 - \hat{\delta}^2$ , respectively.

**PROOF OF PROPOSITION 6.** Standard calculations show that for  $0 \leq t < (2\lambda_{\max})^{-1}$ ,

$$\mathbb{E} \exp\left(t \sum_{i=1}^n \lambda_i (Z_i + \delta_i)^2\right) = \exp\left(\frac{1}{2} \sum_{i=1}^n \left\{ \delta_i^2 \frac{2t\lambda_i}{1 - 2t\lambda_i} - \log(1 - 2t\lambda_i) \right\}\right).$$

Then for any such  $t$ ,

$$\begin{aligned} & \mathbb{P}\left(\sum_{i=1}^n \lambda_i ((Z_i + \delta_i)^2 - (1 + \delta_i^2)) \geq \eta\gamma\right) \\ & \leq \exp\left(-t\eta\gamma - t \sum_{i=1}^n \lambda_i (1 + \delta_i^2)\right) \cdot \mathbb{E} \exp\left(t \sum_{i=1}^n \lambda_i (Z_i + \delta_i)^2\right) \\ & = \exp\left(-t\eta\gamma + \frac{1}{2} \sum_{i=1}^n \left\{ \delta_i^2 \frac{4t^2 \lambda_i^2}{1 - 2t\lambda_i} - \log(1 - 2t\lambda_i) - 2t\lambda_i \right\}\right). \end{aligned} \quad (18)$$

Elementary considerations reveal that

$$-\log(1 - x) - x \leq \begin{cases} x^2/2 & \text{if } x \leq 0, \\ x^2/(2(1 - x)) & \text{if } x \geq 0. \end{cases}$$

Thus (18) is not greater than

$$\begin{aligned} & \exp\left(-t\eta\gamma + \frac{1}{2} \sum_{i=1}^n \left\{ \delta_i^2 \frac{4t^2\lambda_i^2}{1-2t\lambda_i} + \frac{2t^2\lambda_i^2}{1-2t\max(\lambda_i, 0)} \right\}\right) \\ & \leq \exp\left(-t\eta\gamma + \frac{\gamma^2 t^2/2}{1-2t\lambda_{\max}}\right). \end{aligned}$$

Setting

$$t := \frac{\eta}{\gamma + 2\eta\lambda_{\max}} \in [0, (2\lambda_{\max})^{-1}),$$

the preceding bound becomes

$$\mathbb{P}\left(\sum_{i=1}^n \lambda_i ((Z_i + \delta_i)^2 - (1 + \delta_i^2)) \geq \eta\gamma\right) \leq \exp\left(-\frac{\eta^2}{2 + 4\eta\lambda_{\max}/\gamma}\right).$$

Finally, since  $\gamma \geq \lambda_{\max}\sqrt{2}$ , the second asserted inequality follows from

$$\frac{\eta^2}{2 + 4\eta\lambda_{\max}/\gamma} \geq \frac{\eta^2}{2 + \sqrt{8}\eta} = \frac{\eta}{\sqrt{8}} - \frac{\eta}{\sqrt{8} + 4\eta} \geq \frac{\eta}{\sqrt{8}} - \frac{1}{4}. \quad \square$$

### 6.3. Proofs of the main results

Throughout this section we assume without loss of generality that  $\sigma = 1$ . Further let  $\mathcal{S}_n := \{0, 1, \dots, n\}$  and  $\mathcal{T}_n := \{(j, k) : 0 \leq j < k \leq n\}$ .

PROOF OF THEOREM 2. *Step I.* Let the metric  $\rho_n$  on  $\mathcal{T}_n$  be defined by

$$\rho_n((j, k), (j', k')) := \sqrt{\tau_n(j, j')^2 + \tau_n(k, k')^2}.$$

Later on we need bounds for the capacity numbers

$$D(u, \mathcal{T}', \rho_n) := \sup\{|\mathcal{T}_o| : \mathcal{T}_o \subset \mathcal{T}', \rho_n(s, t) > u \text{ for different } s, t \in \mathcal{T}_o\}$$

for certain  $u > 0$  and  $\mathcal{T}' \subset \mathcal{T}$ . Indeed the proof of Theorem 2.1 of Dümbgen and Spokoiny (2001) entails that

$$D(u\delta, \{t \in \mathcal{T}_n : \tau_n(t) \leq \delta\}, \rho_n) \leq 12u^{-4}\delta^{-2} \quad \text{for all } u, \delta \in (0, 1]. \quad (19)$$

Note that for fixed  $(j, k) \in \mathcal{T}_n$ ,  $\pm D_n(j, k)$  may be written as

$$\sum_{i=1}^n \lambda_i ((\epsilon_{in} + \theta_{in})^2 - (1 + \theta_{in}^2))$$

with

$$\lambda_i = \lambda_{in}(j, k) := \pm(4\|\theta_n\|^2 + 2n)^{-1/2} I_{(j,k]}(i),$$

so  $|\lambda_i| \leq (4\|\theta_n\|^2 + 2n)^{-1/2}$ . Hence it follows from Proposition 6 that

$$\mathbb{P}\left(|D_n(t)| \geq \tau_n(t)\eta\right) \leq 2 \exp\left(-\frac{\eta^2}{2 + 4\eta(4\|\theta_n\|^2 + 2n)^{1/2}/\tau_n(t)}\right)$$

for arbitrary  $t \in \mathcal{T}_n$  and  $\eta \geq 0$ . One may rewrite this exponential inequality as

$$\mathbb{P}\left(|D_n(t)| \geq \tau_n(t)G_n(\eta, \tau_n(t))\right) \leq 2 \exp(-\eta) \quad (20)$$

for arbitrary  $t \in \mathcal{T}_n$  and  $\eta \geq 0$ , where

$$G_n(\eta, \delta) := \sqrt{2\eta} + \frac{4\eta}{(4\|\theta_n\|^2 + 2n)^{1/2}\delta}.$$

The second exponential inequality in Proposition 6 entails that

$$\mathbb{P}\left(|D_n(t)| \geq \tau_n(t)\eta\right) \leq 2e^{1/4} \exp(-\eta/\sqrt{8}) \quad (21)$$

and

$$\mathbb{P}\left(|D_n(s) - D_n(t)| \geq \sqrt{8}\rho_n(s, t)\eta\right) \leq 2e^{1/4} \exp(-\eta) \quad (22)$$

for arbitrary  $s, t \in \mathcal{T}_n$  and  $\eta \geq 0$ .

Since  $|\mathcal{T}_n| \leq n^2/2$ , one can easily deduce from (21) that the maximum of  $|D_n|/\tau_n$  over  $\mathcal{T}_n$  exceeds  $\sqrt{32} \log n + \eta$  with probability at most  $e^{1/4} \exp(-\eta/\sqrt{8})$ .

Thus

$$\max_{t \in \mathcal{T}_n} \frac{|D_n(t)|}{\tau_n(t)} \leq \sqrt{32} \log n + O_p(1).$$

Utilizing (19) and (22), it follows from Theorem 7 and the subsequent Remark 3 in Dümbgen and Walther (2007) that

$$\lim_{\delta \downarrow 0} \sup_n \mathbb{P}\left(\sup_{s, t \in \mathcal{T}_n: \rho_n(s, t) \leq \delta} \frac{|D_n(s) - D_n(t)|}{\rho_n(s, t) \log(e/\rho_n(s, t))} > Q\right) = 0 \quad (23)$$

for a suitable constant  $Q > 0$ . Since  $D_n(j, k) = D_n(0, k) - D_n(0, j)$  and  $\tau_n(j, k) = \rho_n((0, j), (0, k))$ , this entails stochastic equicontinuity of  $D_n$  with respect to  $\rho_n$ .

For  $0 \leq \delta < \delta' \leq 1$  define

$$S_n(\delta, \delta') := \sup_{t \in \mathcal{T}_n: \delta < \tau_n(t) \leq \delta'} \left( \frac{|D_n(t)|}{\tau_n(t)} - \Gamma_n(t) - \frac{c \cdot \Gamma_n(t)^2}{\tau_n(t)(4\|\theta_n\|^2 + 2n)^{1/2}} \right)^+$$

with a constant  $c > 0$  to be specified later. Recall that  $\Gamma_n(t)$  equals  $\Gamma(\tau_n(t)^2) = (2 \log(e/\tau_n(t)^2))^{1/2}$ . Starting from (19), (20) and (23), Theorem 8 of Dümbgen and Walther (2007) and its subsequent remark imply that

$$S_n(0, \delta) \rightarrow_p 0 \quad \text{as } n \rightarrow \infty \text{ and } \delta \searrow 0, \quad (24)$$

provided that  $c > 2$ . On the other hand, (19), (21) and (23) entail that

$$S_n(\delta, 1) = O_p(1) \quad \text{for any fixed } \delta > 0. \quad (25)$$

In particular,  $d_n = S_n(0, 1) = O_p(1)$ .

*Step II.* In case of  $\theta_n = (\pm\sigma)_{i=1}^n$ , the process  $(D_n(j, k))_{0 \leq j < k \leq n}$  has the same distribution as  $(W_n(k/n) - W_n(j/n))_{0 \leq j < k \leq n}$  where

$$W_n(t) := \frac{1}{\sqrt{6n}} \sum_{i=1}^{\lfloor nt \rfloor} (\epsilon_{in} + \epsilon_{in}^2 - 1)$$

for  $t \in [0, 1]$  with  $\sum_{i=1}^0 \cdots := 0$ . Moreover,  $\tau_n(j, k)^2 = |k - j|/n$  and  $d_n$  has the same distribution as

$$\max_{0 \leq j < k \leq n} \left( \frac{|W_n(k/n) - W_n(j/n)|}{\tau_n(j, k)} - \Gamma(\tau_n(j, k)^2) - \frac{c \cdot \Gamma(\tau_n(j, k)^2)}{\sqrt{6n} \tau_n(j, k)} \right)^+.$$

According to Donsker's theorem, the process  $(W_n(t))_{t \in [0, 1]}$  converges in distribution to Brownian motion  $W$  on  $[0, 1]$ . Consequently, if we define

$$\Sigma(\delta, \delta') := \sup_{s, t \in [0, 1] : \delta < |s - t| \leq \delta'} \left( \frac{|W(s) - W(t)|}{\sqrt{|s - t|}} - \Gamma_n(|s - t|) \right)^+$$

for  $0 \leq \delta < \delta' \leq 1$ , then

$$S_n(\delta, 1) \rightarrow_{\mathcal{L}} \Sigma(\delta, 1)$$

for any fixed  $\delta \in (0, 1]$ . Moreover, we have seen in (24) that  $S_n(0, \delta) \rightarrow_p 0$  as  $n \rightarrow \infty$  and  $\delta \searrow 0$ . With similar arguments one can show that  $\Sigma(0, \delta) \rightarrow_p 0$  as  $\delta \searrow 0$ . These findings imply that  $d_n = S_n(0, 1)$  converges in distribution to  $\Sigma(0, 1)$  as  $n \rightarrow \infty$ .  $\square$

**PROOF OF PROPOSITION 3.** The main ingredient is a well-known representation of noncentral  $\chi^2$  distributions as Poisson mixtures of central  $\chi^2$  distributions. Precisely,

$$\chi_k^2(\delta^2) = \sum_{j=0}^{\infty} e^{-\delta^2/2} \frac{(\delta^2/2)^j}{j!} \cdot \chi_{k+2j}^2,$$

as can be proved via Laplace transforms. Now we define ‘time points’

$$t_{kn} := \sum_{i=1}^k \theta_{in}^2 \quad \text{and} \quad t_{kn}^* := t_{j(n)n} + k - j(n)$$

with  $j(n)$  any fixed index in  $\mathcal{K}_n(\theta_n)$ . This construction entails that  $t_{kn}^* \geq t_{kn}$  with equality if, and only if,  $k \in \mathcal{K}_n(\theta_n)$ .

Figure 1 illustrates this construction. It shows the time points  $t_{kn}$  (crosses) and  $t_{kn}^*$  (dots and line) versus  $k$  for a hypothetical signal  $\theta_n \in \mathbb{R}^{40}$ . Note that in this example,  $\mathcal{K}_n(\theta_n)$  is given by  $\{10, 11, 20, 21\}$ .

Let  $\Pi$ ,  $G_1$ ,  $G_2$ ,  $\dots$ ,  $G_n$  and  $Z_1$ ,  $Z_2$ ,  $Z_3$ ,  $\dots$  be stochastically independent random variables, where  $\Pi = (\Pi(t))_{t \geq 0}$  is a standard Poisson process, and  $G_i$  and  $Z_j$  are standard Gaussian random variables. Then one can easily verify that

$$\begin{aligned} \tilde{T}_{jkn} &:= \sum_{i=j+1}^k G_i^2 + \sum_{s=2\Pi(t_{jn}/2)+1}^{2\Pi(t_{kn}/2)} Z_s^2, \\ \tilde{T}_{jkn}^* &:= \sum_{i=j+1}^k G_i^2 + \sum_{s=2\Pi(t_{jn}^*/2)+1}^{2\Pi(t_{kn}^*/2)} Z_s^2 \end{aligned}$$

define random variables  $(\tilde{T}_{jkn})_{0 \leq j < k \leq n}$  and  $(\tilde{T}_{jkn}^*)_{0 \leq j < k \leq n}$  with the desired properties.  $\square$

In the proofs of Theorems 4 and 5 we utilize repeatedly two elementary inequalities:

**Lemma 8.** *Let  $a, b, c$  be nonnegative constants.*

(i) *Suppose that  $0 \leq x \leq y \leq x + \sqrt{b(x+y)} + c$ . Then*

$$y \leq x + \sqrt{2bx} + b + \sqrt{bc} + c \leq x + \sqrt{2bx} + (3/2)(b+c).$$

(ii) *For  $x \geq 0$  define  $h(x) := x + \sqrt{a+bx} + c$ . Then*

$$h(h(x)) \leq x + 2\sqrt{a+bx} + b/2 + \sqrt{bc} + 2c.$$

**PROOF OF THEOREM 4.** The definition of  $\hat{\mathcal{K}}_{n,\alpha}$  and Proposition 3 together entail that  $\hat{\mathcal{K}}_{n,\alpha}$  contains  $\mathcal{K}_n(\theta_n)$  with probability at least  $1 - \alpha$ . The assertion about  $\kappa_{n,\alpha}$  is an immediate consequence of Theorem 2.

Now we verify the oracle inequalities (8) and (9). Let  $\gamma_n := (4\|\theta_n\|^2 + 2n)^{1/2} \times \tau_n$ . With  $\gamma_n^*$  we denote the function  $\gamma_n$  on  $\mathcal{T}_n$  corresponding to  $\theta_n^*$ . Throughout

FIG 1. *Construction of the coupling.*

this proof we use the shorthand notation  $M_n(\ell, k) := M_n(\ell) - M_n(k)$  for  $M_n = \hat{R}_n, R_n, \hat{L}_n, L_n$  and arbitrary indices  $\ell, k \in \mathcal{C}_n$ . Furthermore,  $\gamma_n^*(\ell, k) := \gamma_n^*(k, \ell)$  if  $\ell > k$ , and  $\gamma_n^*(k, k) := 0$ .

In the subsequent arguments,  $k_n := \min(\mathcal{K}_n(\theta_n))$ , while  $j$  stands for a generic index in  $\hat{\mathcal{K}}_{n,\alpha}$ . The definition of the set  $\hat{\mathcal{K}}_{n,\alpha}$  entails that

$$\hat{R}_n(j, k_n) \leq \gamma_n^*(j, k_n) \left( \Gamma \left( \frac{|j - k_n|}{n} \right) + \kappa_{n,\alpha} \right) + O(\log n). \quad (26)$$

Combining this with the equation  $R_n(j, k_n) = \hat{R}_n(j, k_n) - D_n(j, k_n)$  yields

$$R_n(j, k_n) \leq \gamma_n^*(j, k_n) \left( \Gamma \left( \frac{j - k_n}{n} \right) + \kappa_{n,\alpha} \right) + O_p(\log n) + |D_n(j, k_n)|. \quad (27)$$

Since  $\gamma_n^*(j, k_n)^2 \leq 6n$  and  $\max_{t \in \mathcal{T}_n} |D_n(t)| / \gamma_n(t) = O_p(\log n)$ , (27) yields

$$R_n(j, k_n) \leq \sqrt{12n} + \sqrt{6n} \kappa_{n,\alpha} + O_p(\log n) \gamma_n(j, k_n).$$

But elementary calculations yield

$$\gamma_n(j, k_n)^2 = \gamma_n^*(j, k_n)^2 + \text{sign}(k_n - j) R_n(j, k_n) \leq 6n + R_n(j, k_n). \quad (28)$$

Hence we may conclude that

$$R_n(j, k_n) \leq O_p(\log n) \sqrt{R_n(j, k_n)} + O_p(\sqrt{n}(\log n + \kappa_{n,\alpha})),$$

and Lemma 8 (i), applied to  $x = 0$  and  $y = R_n(j, k_n)$ , yields

$$\max_{j \in \hat{\mathcal{K}}_{n,\alpha}} R_n(j, k_n) \leq O_p(\sqrt{n}(\log n + \kappa_{n,\alpha})). \quad (29)$$

This preliminary result allows us to restrict our attention to indices  $j$  in a certain subset of  $\mathcal{C}_n$ : Since  $0 \leq R_n(n, k_n) = n - k_n - \sum_{i=k_n+1}^n \theta_{in}^2$ ,

$$\sum_{i=k_n+1}^n \theta_{in}^2 \leq n - k_n.$$

On the other hand, in case of  $j < k_n$ ,  $R_n(j, k_n) = \sum_{i=j+1}^{k_n} \theta_{in}^2 - (k_n - j)$ , so

$$\sum_{i=j+1}^n \theta_{in}^2 \leq n + O_p(\sqrt{n}(\log n + \kappa_{n,\alpha})).$$

Thus if  $j_n$  denotes the smallest index  $j \in \mathcal{C}_n$  such that  $\sum_{i=j+1}^n \theta_{in}^2 \leq 2n$ , then  $k_n \geq j_n$ , and  $\hat{\mathcal{K}}_{n,\alpha} \subset \{j_n, \dots, n\}$  with asymptotic probability one, uniformly in  $\alpha \geq \alpha(n)$ . This allows us to restrict our attention to indices  $j$  in  $\{j_n, \dots, n\} \cap \hat{\mathcal{K}}_{n,\alpha}$ . For any  $\ell \geq j_n$ ,  $D_n(\ell, k_n)$  involves only the restricted signal vector  $(\theta_{in})_{i=j_n+1}^n$ , and the proof of Theorem 2 entails that

$$\max_{j_n \leq \ell \leq n} \left( \frac{|D_n(\ell, k_n)|}{\gamma_n(\ell, k_n)} - \sqrt{2 \log n} - \frac{2c \log n}{\gamma_n(\ell, k_n)} \right)^+ = O_p(1).$$

Thus we may deduce from (27) the simpler statement that with asymptotic probability one,

$$R_n(j, k_n) \leq (\gamma_n^*(j, k_n) + \gamma_n(j, k_n))(\sqrt{2 \log n} + \kappa_{n,\alpha} + O_p(1)) + O_p(\log n). \quad (30)$$

Now we need reasonable bounds for  $\gamma_n^*(j, k_n)^2$  in terms of  $R_n(j)$  and the minimal risk  $\rho_n = R_n(k_n)$ , where we start from the equation in (28): If  $j < k_n$ , then  $\gamma_n(j, k_n)^2 = \gamma_n^*(j, k_n)^2 + 4R_n(j, k_n)$  and  $\gamma_n^*(j, k_n)^2 = 6(k_n - j) \leq 6\rho_n$ . If  $j > k_n$ , then  $\gamma_n^*(j, k_n)^2 = \gamma_n(j, k_n)^2 + 4R_n(j, k_n)$  and

$$\gamma_n(j, k_n)^2 = \sum_{i=k_n+1}^j (4\theta_{in}^2 + 2) \leq 4\rho_n + 2R_n(j) = 6\rho_n + 2R_n(j, k_n).$$

Thus

$$\gamma_n^*(j, k_n) + \gamma_n(j, k_n) \leq 2\sqrt{6}\sqrt{\rho_n} + (\sqrt{2} + \sqrt{6})\sqrt{R_n(j, k_n)},$$



and inequality (30) leads to

$$R_n(j, k_n) \leq \left(4\sqrt{3}\sqrt{\log n} + 2\sqrt{6}\kappa_{n,\alpha} + O_p(1)\right)\sqrt{\rho_n} + O_p(\sqrt{\log n} + \kappa_{n,\alpha})\sqrt{R_n(j, k_n)} + O_p(\log n)$$

for all  $j \in \hat{\mathcal{K}}_{n,\alpha}$ . Again we may employ Lemma 8 with  $x = 0$  and  $y = R_n(j, k_n)$  to conclude that

$$\max_{j \in \hat{\mathcal{K}}_{n,\alpha}} R_n(j, k_n) \leq \left(4\sqrt{3}\sqrt{\log n} + 2\sqrt{6}\kappa_{n,\alpha} + O_p(1)\right)\sqrt{\rho_n} + O_p\left((\log(n))^{3/4} + \kappa_{n,\alpha(n)}^{3/2}\rho_n^{1/4} + \log n + \kappa_{n,\alpha(n)}^2\right)$$

uniformly in  $\alpha \geq \alpha(n)$ .

If  $\log(n)^3 + \kappa_{n,\alpha(n)}^6 = O(\rho_n)$ , then the previous bound for  $R_n(j, k_n) = R_n(j) - \rho_n$  reads

$$\max_{j \in \hat{\mathcal{K}}_{n,\alpha}} R_n(j) \leq \rho_n + \left(4\sqrt{3}\sqrt{\log n} + 2\sqrt{6}\kappa_{n,\alpha} + O_p(1)\right)\sqrt{\rho_n}$$

uniformly in  $\alpha \geq \alpha(n)$ . On the other hand, if we consider just a fixed  $\alpha > 0$ , then  $\kappa_{n,\alpha} = O(1)$ , and the previous considerations yield

$$\begin{aligned} \max_{j \in \hat{\mathcal{K}}_{n,\alpha}} R_n(j) &\leq \rho_n + (4\sqrt{3} + o_p(1))\sqrt{\log(n)\rho_n} \\ &\quad + O_p(\log(n)^{3/4}\rho_n^{1/4} + \log n) \\ &\leq \rho_n + (4\sqrt{3} + o_p(1))\sqrt{\log(n)\rho_n} + O_p(\log n). \end{aligned}$$

To verify the latter step, note that for any fixed  $\epsilon > 0$ ,

$$\log(n)^{3/4}\rho_n^{1/4} \leq \begin{cases} \epsilon^{-1}\log n & \text{if } \rho_n \leq \epsilon^{-4}\log n, \\ \epsilon\sqrt{\log(n)\rho_n} & \text{if } \rho_n \geq \epsilon^{-4}\log n. \end{cases}$$

It remains to prove claim (9) about the losses. From now on,  $j$  denotes a generic index in  $\mathcal{C}_n$ . Note first that

$$L_n(j, k_n) - R_n(j, k_n) = \sum_{i=j+1}^{k_n} (1 - \epsilon_{in}^2) = R_n(k_n, j) - L_n(k_n, j) \quad \text{if } j < k.$$

Thus Theorem 2, applied to  $\theta_n = 0$ , shows that

$$|L_n(j, k_n) - R_n(j, k_n)| \leq \gamma_n^+(j, k_n)(\sqrt{2\log n} + O_p(1)) + O_p(\log n),$$

where

$$\gamma_n^+(j, k_n) := \sqrt{2|k_n - j|} \leq \sqrt{2\rho_n} + \sqrt{2|R_n(j, k)|}.$$

It follows from  $L_n(0) = R_n(0) = \|\theta_n\|^2$  that  $L_n(j) - \rho_n$  equals

$$\begin{aligned} & L_n(j, k_n) + (L_n - R_n)(k_n, 0) \\ &= R_n(j, k_n) + O_p\left(\sqrt{\log(n)\rho_n}\right) + O_p(\sqrt{\log n})\sqrt{R_n(j, k_n)} + O_p(\log n) \\ &\geq O_p\left(\sqrt{\log(n)\rho_n} + \log n\right), \end{aligned}$$

because  $R_n(j, k_n) \geq 0$  and  $R_n(j, k_n) + O_p(r_n)\sqrt{R_n(j, k_n)} \geq O_p(r_n^2)$ . Consequently,  $\hat{\rho}_n := \min_{j \in \mathcal{C}_n} L_n(j)$  satisfies the inequality

$$\hat{\rho}_n \geq \rho_n + O_p\left(\sqrt{\log(n)\rho_n} + \log n\right) = \left(\sqrt{\rho_n} + O_p(\sqrt{\log n})\right)^2,$$

and this entails that

$$\rho_n \leq \left(\sqrt{\hat{\rho}_n} + O_p(\sqrt{\log n})\right)^2.$$

Now we restrict our attention to indices  $j \in \hat{\mathcal{K}}_{n,\alpha}$  again. Here it follows from our result about the maximal risk over  $\hat{\mathcal{K}}_{n,\alpha}$  that  $L_n(j) - \rho_n$  equals

$$\begin{aligned} & R_n(j, k_n) + O_p(\sqrt{\log(n)\rho_n}) + O_p(\sqrt{\log n})\sqrt{R_n(j, k_n)} + O_p(\log n) \\ &\leq 2R_n(j, k_n) + O_p(\sqrt{\log(n)\rho_n} + \log n) \leq O_p\left(\sqrt{\log(n)\rho_n} + \log n\right). \end{aligned}$$

Hence  $\max_{j \in \hat{\mathcal{K}}_{n,\alpha}} L_n(j)$  is not greater than

$$\begin{aligned} \rho_n + O_p\left(\sqrt{\log(n)\rho_n} + \log n\right) &= \left(\sqrt{\rho_n} + O_p(\sqrt{\log n})\right)^2 \\ &\leq \left(\sqrt{\hat{\rho}_n} + O_p(\sqrt{\log n})\right)^2. \quad \square \end{aligned}$$

PROOF OF THEOREM 5. The application of inequality (17) in Corollary 7 to the triple  $(|J|, T_n(J) - |J|, \alpha/(2M_n))$  in place of  $(n, \hat{\delta}^2, \alpha)$  yields bounds for  $\hat{\delta}_{n,\alpha,l}^2(J)$  and  $\hat{\delta}_{n,\alpha,u}^2(J)$  in terms of  $\hat{\delta}_n^2(J) := (T_n(J) - |J|)_+$ . Then we apply (15-16) to  $T_n(J)$ , replacing  $(n, \delta^2, u)$  with  $(|J|, \delta_n^2(J), \alpha'/(2M_n))$  for any fixed  $\alpha' \in (0, 1)$ . By means of Lemma 8 (ii) we obtain finally

$$\begin{aligned} \left. \begin{aligned} \hat{\delta}_{n,\alpha,u}^2(J) - \delta_n^2(J) \\ \delta_n^2(J) - \hat{\delta}_{n,\alpha,l}^2(J) \end{aligned} \right\} &\leq (1 + o_p(1))\sqrt{(16|J| + 32\hat{\delta}_n^2(J))\log M_n} \quad (31) \\ &+ (K + o_p(1))\log M_n \end{aligned}$$

for all  $J \in \mathcal{M}_n$ . Here and throughout this proof,  $K$  denotes a generic constant not depending on  $n$ . Its value may be different in different expressions. It follows from the definition of the confidence region  $\hat{\mathcal{K}}_{n,\alpha}$  that for arbitrary  $C \in \hat{\mathcal{K}}_{n,\alpha}$  and  $D \in \mathcal{C}_n$ ,

$$\begin{aligned} R_n(C) - R_n(D) &= \delta_n^2(D \setminus C) - \delta_n^2(C \setminus D) + |C| - |D| \\ &= (\delta_n^2 - \hat{\delta}_{n,\alpha,l}^2)(D \setminus C) + (\hat{\delta}_{n,\alpha,u}^2 - \delta_n^2)(C \setminus D) \\ &\quad - (\hat{\delta}_{n,\alpha,u}^2(C \setminus D) - \hat{\delta}_{n,\alpha,l}^2(D \setminus C) + |D| - |C|) \\ &\leq (\delta_n^2 - \hat{\delta}_{n,\alpha,l}^2)(D \setminus C) + (\hat{\delta}_{n,\alpha,u}^2 - \delta_n^2)(C \setminus D). \end{aligned}$$

Moreover, according to (31) the latter bound is not larger than

$$\begin{aligned} &(1 + o_p(1)) \left\{ \sqrt{(16|D \setminus C| + 32\delta_n^2(D \setminus C)) \log M_n} \right. \\ &\quad \left. + \sqrt{(16|C \setminus D| + 32\delta_n^2(C \setminus D)) \log M_n} \right\} + (K + o_p(1)) \log M_n \\ &\leq (1 + o_p(1)) \sqrt{2(16|D| + 32\delta_n^2(C^c) + 16|C| + 32\delta_n^2(D^c)) \log M_n} \\ &\quad + (K + o_p(1)) \log M_n \\ &\leq 8\sqrt{(R_n(C) + R_n(D)) \log M_n} (1 + o_p(1)) + (K + o_p(1)) \log M_n. \end{aligned}$$

Thus we obtain the quadratic inequality

$$\begin{aligned} R_n(C) - R_n(D) &\leq 8\sqrt{(R_n(C) + R_n(D)) \log M_n} (1 + o_p(1)) \\ &\quad + (K + o_p(1)) \log M_n, \end{aligned}$$

and with Lemma 8 this leads to

$$R_n(C) \leq R_n(D) + 8\sqrt{2}\sqrt{R_n(D) \log M_n} (1 + o_p(1)) + (K + o_p(1)) \log M_n.$$

This yields the assertion about the risks.

As for the losses, note that  $L_n(\cdot)$  and  $R_n(\cdot)$  are closely related in that

$$(L_n - R_n)(D) = \sum_{i \in D} \epsilon_{in}^2 - |J|$$

for arbitrary  $D \in \mathcal{C}_n$ . Hence we may utilize (15-16), replacing  $(n, \delta^2, u)$  with  $(|D|, 0, \alpha'/(2\mu_n))$ , to complement (31) with the following observation:

$$-A\sqrt{|D| \log M_n} \leq L_n(D) - R_n(D) \leq A\sqrt{|D| \log M_n} + A \log M_n \quad (32)$$

simultaneously for all  $D \in \mathcal{C}_n$  with probability tending to one as  $n \rightarrow \infty$  and  $A \rightarrow \infty$ . Note also that (32) implies that  $R_n(D) \leq A\sqrt{R_n(D)\log M_n} + L_n(D)$ . Hence

$$R_n(D) \leq (3/2)(L_n(D) + A^2 \log M_n) \quad \text{for all } D \in \mathcal{C}_n,$$

by Lemma 8 (i). Assuming that both (31) and (32) hold for some large but fixed  $A$ , we may conclude that for arbitrary  $C \in \hat{\mathcal{K}}_{n,\alpha}$  and  $D \in \mathcal{C}_n$ ,

$$\begin{aligned} L_n(C) - L_n(D) &= (L_n - R_n)(C) - (L_n - R_n)(D) + R_n(C) - R_n(D) \\ &\leq A\sqrt{2(|C| + |D|)\log M_n} + A\sqrt{2(R_n(C) + R_n(D))\log M_n} + 4A \log M_n \\ &\leq 2A\sqrt{2(R_n(C) + R_n(D))\log M_n} + 4A \log M_n \\ &\leq A'\sqrt{(L_n(C) + L_n(D))\log M_n} + A'' \log M_n \end{aligned}$$

for constants  $A'$  and  $A''$  depending on  $A$ . Again this inequality entails that

$$L_n(C) \leq L_n(D) + A'\sqrt{2L_n(D)\log M_n} + A''' \log M_n$$

for another constant  $A''' = A'''(A)$ . □

**Acknowledgements.** This work was supported by the Swiss National Science Foundation. Constructive comments by two referees and an associate editor are gratefully acknowledged.

## References

- [1] BARAUD, Y. (2004). Confidence balls in Gaussian regression. *Ann. Statist.* **32**, 528-551.
- [2] BERAN, R. (1996). Confidence sets centered at  $C_p$  estimators. *Ann. Inst. Statist. Math.* **48**, 1-15.
- [3] BERAN, R. (2000). REACT scatterplot smoothers: superefficiency through basis economy. *J. Amer. Statist. Assoc.* **95**, 155-169.
- [4] BERAN, R. and DÜMBGEN, L. (1998). Modulation of estimators and confidence sets. *Ann. Statist.* **26**, 1826-1856.
- [5] BIRGÉ, L. AND MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3**, 203-268.
- [6] CAI, T.T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.* **26**, 1783-1799.

- [7] CAI, T.T. (2002). On block thresholding in wavelet regression: adaptivity, block size, and threshold level. *Statist. Sin.* **12**, 1241-1273.
- [8] CAI, T.T. and LOW, M.G. (2006). Adaptive confidence balls. *Ann. Statist.* **34**, 202-228.
- [9] CAI, T.T. and LOW, M.G. (2007). Adaptive estimation and confidence intervals for convex functions and monotone functions. *Manuscript in preparation*.
- [10] DAHLHAUS, R. and POLONIK, W. (2006). Nonparametric quasi-maximum likelihood estimation for Gaussian locally stationary processes. *Ann. Statist.* **34**, 2790-2824.
- [11] DONOHO, D.L. and JOHNSTONE, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.
- [12] DONOHO, D.L. and JOHNSTONE, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *JASA* **90**, 1200-1224.
- [13] DONOHO, D.L. and JOHNSTONE, I.M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879-921.
- [14] DÜMBGEN, L. (2003). Optimal confidence bands for shape-restricted curves. *Bernoulli* **9**, 423-449.
- [15] DÜMBGEN, L. and SPOKOINY, V.G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.* **29**, 124-152.
- [16] DÜMBGEN, L. and WALTHER, G. (2007). Multiscale inference about a density. Technical report 56, IMSV, University of Bern.
- [17] EFROMOVICH, S. (1998). Simultaneous sharp estimation of functions and their derivatives. *Ann. Statist.* **26**, 273-278.
- [18] GENOVESE, C.R. and WASSERMANN, L. (2005). Confidence sets for nonparametric wavelet regression. *Ann. Statist.* **33**, 698-729.
- [19] GINÉ, E. and NICKL, R. (2010). Confidence Bands in Density Estimation. *Ann. Statist.* **38**, 1122-1170.
- [20] HENGARTNER, N.W. and STARK, P.B. (1995). Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.* **23**, 525-550.
- [21] HOFFMANN, M. and NICKL, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39**, 2383-2409.
- [22] JUDITSKY, A. and LAMBERT-LACROIX, S. (2003). Nonparametric confidence set estimation. *Math. Meth. of Statist.* **19**, 410-428.
- [23] 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.
- [24] LI, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17**, 1001-1008.
- [25] POLYAK, B.T. and TSYBAKOV, A.B. (1991). Asymptotic optimality of the  $C_p$ -test for the orthogonal series estimation of regression. *Theory Probab. Appl.* **35**, 293-306.
- [26] ROBINS, J. and VAN DER VAART, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* **34**, 229-253.
- [27] ROHDE, A. and DÜMBGEN, L. (2009). Adaptive confidence sets for the optimal approximating model. Technical report 73, IMSV, Univ. of Bern.

- [28] STONE, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12**, 1285-1297.