The Annals of Statistics
1994, Vol. 22, No. 2, 835–866

ORDERED LINEAR SMOOTHERS[1]

BY ALOIS KNEIP

Université de Louvain

This paper deals with the following approach for estimating the mean μ of an n-dimensional random vector Y: first, a family $\mathbf{S}$ of $n \times n$ matrices is specified. Then, an element $\widehat{S} \in \mathbf{S}$ is selected by Mallows C_L, and $\widehat{\mu} = \widehat{S} \cdot Y$. The case is considered that $\mathbf{S}$ is an "ordered linear smoother" according to some easily interpretable, qualitative conditions. Examples include linear smoothing procedures in nonparametric regression (as, e.g., smoothing splines, minimax spline smoothers and kernel estimators). Stochastic probability bounds are given for the difference $(1/n)\|\mu - \widehat{S} \cdot Y\|_2^2 - (1/n)\|\mu - \widehat{S}_\mu \cdot Y\|_2^2$, where $\widehat{S}_\mu$ denotes the minimizer of $(1/n)\|\mu - S \cdot Y\|_2^2$ for $S \in \mathbf{S}$. These probability bounds are generalized to the situation that $\mathbf{S}$ is the union of a moderate number of ordered linear smoothers. The results complement work by Li on the asymptotic optimality of C_L. Implications for nonparametric regression are studied in detail. It is shown that there exists a direct connection between James–Stein estimation and the use of smoothing procedures, leading to a decision-theoretic justification of the latter. Further conclusions concern the choice of the order of a smoothing spline or a minimax spline smoother and the rates of convergence of smoothing parameters.

1. Introduction We shall consider methods to estimate the mean of a multivariate distribution. It is assumed that there is a random vector $Y = (Y_1, \ldots, Y_n)^T$, $n \in \mathbb{N}$, satisfying the following model assumption:

$$Y_i = \mu_i + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

for some $\mu = (\mu_1, \ldots, \mu_n)^T \in \mathbb{R}^n$ and i.i.d. random variables $\epsilon_1, \ldots, \epsilon_n$ with a common probability distribution W. It holds that $E(\epsilon_1) = 0$, $\operatorname{var}(\epsilon_1) = \sigma^2 < \infty$.

Throughout the major parts of this paper the variance σ^2 is assumed to be known. The problem is to estimate μ.

The most important special case of model (1.1) is nonparametric regression. It is then assumed that μ is generated by an underlying function f, that is, (1.1) is complemented by the following assumption:

(1.2) For some known "design points" $x_1, \ldots, x_n \in J \subset \mathbb{R}^d$, $d \in \mathbb{N}$, it holds that $\mu_i = f(x_i)$, $i = 1, \ldots, n$, where $f: J \to \mathbb{R}$ is an unknown function.

Usually no quantitative information about f is available, and it is only assumed that f is "smooth." Within theoretical considerations this is represented by

Received May 1990; revised July 1993.

[1]This work has been performed as part of the research program of the Sonderforschungsbereich 303 at the University of Bonn, with financial support by Deutsche Forschungsgemeinschaft.

AMS 1991 *subject classifications.* 62G07, 62J07.

Key words and phrases. Nonparametric regression, James–Stein estimation.

requiring that f possess a number of continuous derivatives. In this context, procedures for estimating μ are usually called *smoothing methods*.

Most commonly used smoothing methods are linear smoothers. Examples are smoothing splines, kernel estimators, polynomial regression and so on. Estimates are obtained by multiplying a matrix with Y. Definition of the *smoother matrices* S_h depends on the specifications of the particular method and on the design points $x_1, \ldots, x_n$. Typically these procedures involve a *smoothing parameter* h, and $\{S_h \cdot Y\}_{h \in H}$ (for some set H) defines a whole class of possible estimators of μ. An element $\widehat{S} \equiv S_{\hat{h}} \in \mathbf{S} = \{S_h\}_{h \in H}$ is then selected by a data-adaptive method, and $\widehat{\mu} = \widehat{S} \cdot Y$ is used as final estimator of μ. The literature on smoothing methods is very large. The interested reader may consult the books of Eubank (1988) or Härdle (1990). A discussion of smoothing procedures and the smoother matrices associated with them can be found in Buja, Hastie and Tibshirani (1989).

Following Li (1985, 1986, 1987) we will consider the following general approach of constructing an estimate $\widehat{\mu}$ of μ:

1. A family $\mathbf{S}$ of real $n \times n$ matrices is specified.
2. A matrix $\widehat{S} \in \mathbf{S}$ satisfying

$$\frac{1}{n}\|Y - \widehat{S} \cdot Y\|_2^2 + \frac{2\sigma^2}{n}\operatorname{tr}(\widehat{S}) = \min_{S \in \mathbf{S}} \left(\frac{1}{n}\|Y - S \cdot Y\|_2^2 + \frac{2\sigma^2}{n}\operatorname{tr}(S) \right) \tag{1.3}$$

is determined, and $\widehat{\mu} = \widehat{S} \cdot Y$ is used as estimator of μ.

The above procedure for selecting $\widehat{S}$ is well known an *Mallows' C_L* [Mallows (1973)]. In nonparametric regression most of the popular methods for choosing the smoother matrix (or smoothing parameter) can essentially be considered as versions of (1.3) which replace the true variance by an estimated one. In particular, this holds for generalized cross-validation [Craven and Wahba (1979)] and related procedures [cf., e.g., Rice (1984) and Härdle, Hall and Marron (1988)]. Li (1985) establishes a still closer relation between C_L and generalized cross-validation. The use of Mallows' C_L can be motivated as follows:

For fixed $S \in \mathbf{S}$ the average squared error $(1/n)\|\mu - S \cdot Y\|_2^2$ quantifies quadratic loss (normalized by $1/n$), and

$$\mathrm{MASE}_\mu(S) := \frac{1}{n}\|\mu - S \cdot \mu\|_2^2 + \frac{\sigma^2}{n}\operatorname{tr}(S^T S) = E\frac{1}{n}\|\mu - S \cdot Y\|_2^2$$

yields the corresponding risk ("E" denotes expectation). It is now easily seen that for any $S \in \mathbf{S}$ we have

$$E\left(\frac{1}{n}\|Y - S \cdot Y\|_2^2 + \frac{2\sigma^2}{n}\operatorname{tr}(S) - \frac{1}{n}\|\mu - S \cdot Y\|_2^2 \right) - \sigma^2 = 0.$$

We can conclude that (1.3) relies on unbiased estimation of the true risk or loss. The resulting matrices $\widehat{S}$ allow two interpretations: they can either be

considered as estimators of S_μ or of $\widehat{S}_\mu$, where S_μ and $\widehat{S}_\mu$ denote the minimizers of $\mathrm{MASE}_\mu(S)$ and $(1/n)\|\mu - S \cdot Y\|_2^2$ with respect to $S \in \mathbf{S}$.

The main issue of this paper is to give some precise stochastic bounds for the differences

$$\frac{1}{n}\|\mu - \widehat{S} \cdot Y\|_2^2 - \frac{1}{n}\|\mu - \widehat{S}_\mu \cdot Y\|_2^2, \qquad \frac{1}{n}\|\mu - \widehat{S} \cdot Y\|_2^2 - \frac{1}{n}\|\mu - S_\mu \cdot Y\|_2^2$$

under certain assumptions on the class $\mathbf{S}$ of smoother matrices used. This then leads to interesting conclusions.

Our approach is closely related to previous work by Li (1986, 1987). Li (1986) deals with smoother matrices resulting from ridge regression, while in Li (1987) arbitrary classes $\mathbf{S}$ containing a finite number of elements are considered. Under some weak conditions Li then shows that as $n \to \infty$ the resulting estimators $\widehat{S} \cdot Y$ satisfy

$$\frac{(1/n)\|\mu - \widehat{S} \cdot Y\|_2^2}{(1/n)\|\mu - \widehat{S}_\mu \cdot Y\|_2^2} \to 1 \quad \text{in probability,} \tag{1.4}$$

provided that $\mathrm{MASE}_\mu(S_\mu)$ converges to zero not too fast.

The present paper deals with somewhat different families $\mathbf{S}$ defining an *ordered linear smoother*. A definition and several examples are given in Section 2. Many of the basic smoothing methods like smoothing splines, minimax spline smoothers, some versions of kernel estimators and so on lead to ordered linear smoothers.

Based on an additional assumption on ϵ, an exponential probability inequality is derived is Section 3 which bounds the difference $(1/n)\|\mu - \widehat{S} \cdot Y\|_2^2 - (1/n)\|\mu - \widehat{S}_\mu \cdot Y\|_2^2$ for all $n, \mu \in \mathbb{R}^n$ and each ordered linear smoother. Section 4 provides a generalization of the approach. The case is considered that, for some $m \in \mathbb{N}$, $\mathbf{S} = \mathbf{S}_1 \cup \mathbf{S}_2 \cup \cdots \cup \mathbf{S}_m$ holds, where for each $i \in \{1, \ldots, m\}$ $\mathbf{S}_i$ is an ordered linear smoother. A probability inequality similar to that of Section 3 is established, which, however, now additionally depends on $\log(m) + 1$. The inequalities allow one to infer that there exists a $d < \infty$ such that

$$\begin{aligned} \sup_{\mu \in \mathbb{R}^n} \left(\left(E\frac{1}{n}\|\mu - \widehat{S} \cdot Y\|_2^2 \right)^{1/2} - \left(E\frac{1}{n}\|\mu - \widehat{S}_\mu \cdot Y\|_2^2 \right)^{1/2} \right) \\ \leq d n^{-1/2} \big(\log(m) + 1 \big)^2 \end{aligned} \tag{1.5}$$

holds for all $n \in \mathbb{N}$ and all estimators constructed in the above way. For methods like ridge regression and smoothing splines ($m = 1$) this complements Li's results (1.4) by providing a quantitative bound which holds independently of the particular value of $E\big[(1/n)\|\mu - \widehat{S}_\mu \cdot Y\|_2^2\big] \leq \mathrm{MASE}_\mu(S_\mu)$. Relation (1.5) has several interesting consequences when considering, for example, smoothing spline or minimax spline smoothers. In particular, support is given to the idea of using the data to decide about the *order* of a smoothing spline or a minimax spline

smoother. This generalizes results by Hall and Marron (1988) on the choice of kernel order in density estimation.

Section 5 provides some further conclusions from the results of Sections 2–4. Section 5.1 contains a discussion of decision theoretic aspects. Estimating μ means to estimate the mean of a multivariate distribution. Assuming (1.2) for an arbitrarily smooth f does not imply any further restrictions. It is easily seen that for all $n, k \in \mathbb{N}$ and *any* $\mu \in \mathbb{R}^n$ there exists a k-times continuously differentiable function $f\colon J \to \mathbb{R}$ satisfying $\mu = (f(x_1), \ldots, f(x_n))^T$, unless $x_i = x_j$ for some $i \neq j$. This raises the question whether from a decision-theoretic point of view there is any justification for the use of methods which are based on such vague "assumptions." Li and Hwang (1984) and Li (1985, 1989) introduced a way to overcome this problem. Instead of using smoothing procedures directly, they consider the associated Stein estimates which possess bounded risk. It is now shown that there exists a *direct* connection between the use of smoothing methods and James–Stein estimation. In particular, any estimator $\widehat{S} \cdot Y$ obtained from an ordered linear smoother can be considered as a straightforward generalization of the James–Stein estimator [James and Stein (1961)]. Based on (1.5) the decision-theoretic motivation of the latter carries over.

Section 5.2 deals with the situation that under (1.2) we let $n \to \infty$ by sampling more and more observations from a fixed function f. This corresponds to the usual asymptotic theory for smoothing procedures. We are then interested in the rate of convergence of

$$\left| \frac{1}{n} \|\mu - \widehat{S} \cdot Y\|_2^2 - \frac{1}{n} \|\mu - \widehat{S}_\mu \cdot Y\|_2^2 \right|,$$

where, for any n, $\widehat{S}$ is obtained from an ordered linear smoother ($m = 1$). It follows from the results of Section 3 that this rate depends on the asymptotic behavior of a function $R_{\mu, \mathbf{S}}$, which quantifies the steepness of the minimum of $\mathrm{MASE}_\mu(\cdot)$ at S_μ. Relation (1.5) only provides an upper bound. In many situations the function $R_{\mu, \mathbf{S}}$ will behave reasonably well, and we will obtain

$$\left| \frac{1}{n} \|\mu - \widehat{S} \cdot Y\|_2^2 - \frac{1}{n} \|\mu - \widehat{S}_\mu \cdot Y\|_2^2 \right| = O_P(1/n).$$

The special case of kernel estimation serves as illustration. For kernel estimators $\widehat{S} \equiv S_{\hat{h}}$ and $\widehat{S}_\mu \equiv S_{\hat{h}(\mu)}$ hold for some appropriate smoothing parameters (bandwidths) $\hat{h}$ and $\hat{h}(\mu)$. The results of Section 5.2 include rates of convergence of $(\hat{h} - \hat{h}(\mu))/\hat{h}(\mu)$ under different conditions on the amount of smoothness of f. This generalizes work by Rice (1984) and Härdle, Hall and Marron (1988).

Finally, in Section 6 we analyze the case that σ^2 is unknown. It is then assumed that in (1.3) σ^2 is replaced by a consistent estimator $\hat{\sigma}^2$. It turns out that the basic results of the previous sections carry over. However, some more specific smoothness assumptions are required.

2. Ordered linear smoothers. We will consider families of smoother matrices defining an ordered linear smoother according to the following definition.

DEFINITION. A closed subspace $\mathbf{S} \subset M(n)$ is called an *ordered linear smoother* if the following conditions are satisfied:

(i) $0 \leq p^T S p \leq p^T p$ for all $S \in \mathbf{S}$, $p \in \mathbb{R}^n$;
(ii) $S \cdot \widetilde{S} = \widetilde{S} \cdot S$ for all $S, \widetilde{S} \in \mathbf{S}$;
(iii) for all $S, \widetilde{S} \in \mathbf{S}$ either $S \geq \widetilde{S}$ or $\widetilde{S} \geq S$.

Hereby, $M(n)$ denotes the space of all real symmetric $n \times n$ matrices endowed with the metric $d(A,B) := \max_{i,j} |a_{ij} - b_{ij}|$. Furthermore, for matrices $A, B \in M(n)$ we write $A \geq B$ if $A - B$ is positive semidefinite.

The conditions introduced in (i)–(iii) can equivalently be described as follows:

(2.1) There is an orthonormal basis $u_1, \ldots, u_n$ of $\mathbb{R}^n$ such that, for all $S \in \mathbf{S}, S = \sum_{i=1}^n \lambda_i(S) \cdot u_i u_i^T$ with $1 \geq \lambda_1(S) \geq \lambda_2(S) \geq \cdots \geq \lambda_n(S) \geq 0$.

(2.2) For all $S, \widehat{S} \in \mathbf{S}$ it either holds that $\lambda_i(S) \geq \lambda_i(\widetilde{S})$, $i = 1, \ldots, n$, or that $\lambda_i(\widetilde{S}) \geq \lambda_i(S)$, $i = 1, \ldots, n$.

In the following we will consider examples of ordered linear smoothers. Let us start with the simplest case.

EXAMPLE 1 (James–Stein method). Let $\mathbf{S} := \{(1-h) \cdot I\}_{h \in [0,1]}$. For any h the estimator $S_h \cdot Y = (1-h) \cdot Y$ simply shrinks each element of Y by the fixed amount $1 - h$. Clearly, $\mathbf{S}$ is an ordered linear smoother.

The above example might be called the James–Stein method since, when choosing $\widehat{S}$ by (1.3), we obtain $\widehat{S} = (1 - n\sigma^2/Y^T Y) \cdot I$. For large n the resulting estimator $\widehat{\mu} = (1 - n\sigma^2/Y^T Y) \cdot Y$ almost coincides with the James–Stein estimator $\widehat{\mu} = (1 - (n-2)\sigma^2/Y^T Y) \cdot Y$ [James and Stein (1961)].

Straightforward generalizations of Example 1 are methods which shrink by a different amount in different directions [directions are given by the eigenvectors; compare (2.1) and (2.12)]. The most important special cases are some of the basic linear smoothing methods in nonparametric regression. The idea of "smoothing" is to eliminate what are to be considered "wiggly" components of the vector Y. Based on (1.2), smoother matrices S are constructed in such a way that $S \cdot v \approx 0$, if $v \in \mathbb{R}^n$ corresponds to the functional values of a "wiggly" function, and $S \cdot v \approx v$, if $v \in \mathbb{R}^n$ corresponds to the functional values of a smooth function. This means that the matrices S have to shrink by a different amount in different directions.

The above argument shows that (i) is a quite natural condition when dealing with linear smoothers. The same holds for condition (iii) when considering *basic* smoothing procedures, such as, for example, smoothing splines. Such methods can be parametrized by a one-dimensional parameter which controls the amount of smoothing (shrinking). There then usually exists a natural ordering of these smoothing parameters in the sense that when choosing a large one the

method performs a "stronger smoothing" than when selecting a small parameter. When translating this in terms of the resulting smoother matrices, this corresponds to (iii).

Nevertheless, it has to be emphasized that conditions (i)–(iii) provide a rather narrow framework. In particular, the symmetry of $S \in \mathbf{S}$ and condition (ii) impose major restrictions. The latter, for example, excludes B-splines with knots at the sample quantiles. Furthermore, the conditions exclude complex smoothers which cannot be parameterized by a one-dimensional smoothing parameter (note that, by (iii), $S \to (1/n)\mathrm{tr}(S)$ is necessarily a bijective mapping from $\mathbf{S}$ into $[0, 1]$).

The following examples refer to some smoothing procedures which lead to ordered linear smoothers. We will assume that (1.1) and (1.2) hold.

EXAMPLE 2 (Least squares regression on nested subspaces). Consider a system $\phi_1, \phi_2, \ldots$ of functions which provide a suitable functional basis for approximating any smooth function (e.g., polynomials, harmonic polynomials, etc.). For $i \in \mathbb{N}$ let $\phi_{i,n} := (\phi_i(x_1), \ldots, \phi_i(x_n))^T$. Suppose that Φ_n is regular, where Φ_h, $h = 1, \ldots, n$, denote the $n \times h$ matrices $(\phi_{1,n}, \ldots, \phi_{h,n})$.

A straightforward idea of estimating $\mu = (f(x_1), \ldots, f(x_n))^T$ is to select an appropriate $h \in \{1, \ldots, n\}$ and to minimize $\|Y - \sum_{i=1}^{h} a_i \cdot \phi_{i,n}\|_2^2$ with respect to $a_1, \ldots, a_h \in \mathbb{R}$. The resulting estimate $\hat{\mu}_h$ is given by

$$\hat{\mu}_h = \Phi_h(\Phi_h^T\Phi_h)^{-1}\Phi_h^T \cdot Y =: S_h \cdot Y. \tag{2.3}$$

Evidently $\mathbf{S} := \{S_h\}_{h \in \{1, \ldots, n\}}$ defines an ordered linear smoother. In practice such procedures are quite frequently applied, using, for example Legendre polynomials (if $d = 1$). For some theoretical properties of least squares regression see Cox (1988).

EXAMPLE 3 (Ridge regression). Let Φ be a given regular $n \times n$ design matrix. The idea of ridge regression then consists in estimating μ by

$$\hat{\mu}_h = \Phi(\Phi^T\Phi + hI)^{-1}\Phi^T \cdot Y =: S_h \cdot Y,$$

for $h \in H := \mathbb{R}_+$. Clearly $\mathbf{S} := \{S_h\}_{h \in \{1, \ldots, n\}}$ is an ordered linear smoother.

EXAMPLE 4 (Smoothing splines and other penalized least squares methods). Let $d = 1$ and assume that J is a compact subinterval of $\mathbb{R}$. Smoothing spline estimators of order $2k$ are defined in the following way: for $h \in H := [0, \infty]$, determine the function $f_{\hat{\mu}(h)}$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - w(x_i)\big)^2 + h \cdot \int_J \big[w^{(k)}(x)\big]^2\,dx \tag{2.4}$$

with respect to the Soblev space

$$\mathcal{F}_k(J) := \Big\{ w\colon J \to \mathbb{R} \mid w \text{ has } k-1 \text{ absolutely continuous derivatives, and } \int_J \left[w^{(k)}(x)\right]^2 dx < \infty \Big\}$$

(for $h = \infty$, let $h \cdot 0 := 0$). Then set $\hat{\mu}(h) = (f_{\hat{\mu}(h)}(x_1), \ldots, f_{\hat{\mu}(h)}(x_n))^T$.

Let us assume that all design points are distinct. Then (2.4) is equivalent to determining $\hat{\mu}(h)$ by minimizing $\sum_{i=1}^n (Y_i - u_i)^2 + h \cdot \int_J [G_u^{(k)}(x)]^2\, dx$ with respect to $u \in \mathbb{R}^n$, where G_u denotes a spline interpolant of order $2k$ of $u_1, \ldots, u_n$ at $x_1, \ldots, x_n$ [G_u has knots at each x_i; cf. Reinsch (1967)]. With A denoting the symmetric $n \times n$ matrix corresponding to the bilinear map $p, q \to \int G_p^{(k)}(t) G_q^{(k)}(t)\, dt$ [see, e.g. Utreras (1983)], $\hat{\mu}(h)$ minimizes

$$\sum_{i=1}^n (Y_i - u_i)^2 + h \cdot u^T A u. \tag{2.5}$$

This yields

$$\hat{\mu}(h) = (I + h \cdot A)^{-1} \cdot Y =: S_h \cdot Y,$$

and $\mathbf{S} = \{S_h\}_{h \in H}$ defines an ordered linear smoother. More generally, most penalized least squares approaches of the type (2.5) lead to ordered linear smoothers. It is only required that A be symmetric and positive semidefinite. In particular, this holds for multivariate generalizations of smoothing splines [cf. Wahba and Wendelberger (1980)].

EXAMPLE 5 (Minimax spline smoothing). Let $d = 1, J = [0, 1]$ and assume that $f \in \mathcal{F}_k(J)$ for some k. Here, $\mathcal{F}_k(J)$ is defined as in Example 4. Following the construction of smoothing splines of order $2k$, let $u_1, \ldots, u_n$ denote an orthonormal eigensystem of the resulting matrix A such that $A = \sum_{i=1}^n \alpha_i u_i u_i^T$ with $\alpha_1 \le \alpha_2 \le \cdots \le \alpha_n$.

Following Speckman (1985), for any $h \in H := [0, \infty[$ a smoother matrix S_h now might be defined by

$$S_h := \sum_{i=1}^n \left(1 - \sqrt{h \cdot \alpha_i}\right)_+ \cdot u_i u_i^T.$$

Here, the notation $(t)_+ = \max\{t, 0\}$ is used. This establishes minimax spline smoothers of order $2k$. It is immediately seen that $\mathbf{S} = \{S_h\}_{h \in H}$ is an ordered linear smoother.

There exists an interesting minimax result characterizing this choice of $\mathbf{S}$. For $\gamma \in]0, \infty[$ let

$$\mathcal{F}(k, \gamma, n) := \Big\{ p \in \mathbb{R}^n \mid p = \big(w(x_1), \ldots, w(x_n)\big)^T \text{ for some } w \in \mathcal{F}_k(J) \text{ with } \int_J \left[w^{(k)}(x)\right]^2 dx \le \gamma \Big\}.$$

Based on some weak additional assumptions on the design, Speckman (1985) then shows that, for any fixed $\gamma > 0$,

$$\sup_{\mu \in \mathcal{F}(k,\gamma,n)} \mathrm{MASE}_\mu(S_\mu) = \inf_{\tilde{\mu} \in \mathcal{L}_n} \sup_{\mu \in \mathcal{F}(k,\gamma,n)} \frac{1}{n} E\|\mu - \tilde{\mu}(Y)\|_2^2 = O\left(n^{-2k/(2k+1)}\right), \tag{2.6}$$

where $\mathcal{L}_n$ denotes the class of all possible linear estimators of μ.

Under some further assumptions on the design, an even stronger asymptotic minimax result can be obtained from Nussbaum (1985). Nussbaum's results imply that for $k = 2$ up to a term of order $o(n^{-2k/(2k+1)})$ (2.6) remains true if $\mathcal{L}_n$ is replaced by the class of *all* estimators of μ, that is, of all measurable mappings from $\mathbb{R}^n$ into $\mathbb{R}^n$.

EXAMPLE 6 (Kernel estimators). The basic idea of kernel estimation consists in smoothing via local averaging. Assume that $J = [0, 1]$ and $X_i = (i - 1)/n$, $i = 1, \ldots, n$. Furthermore, suppose that we believe in $f(0) = f(1)$ and $f'(0) = f'(1)$. Let K denote the Epanechnikov kernel. In this setup, a kernel estimator with kernel function K and bandwidth h estimates μ by

$$\hat{\mu}(h)_i = \frac{\sum_{j=1}^n K\left(q(X_i, X_j)/h\right) \cdot Y_j}{\sum_{j=1}^n K\left(q(X_i, X_j)/h\right)}, \qquad i = 1, \ldots, n. \tag{2.7}$$

Here, $q(X_i, X_j) = \min\{|X_i - X_j|, |1 + X_i - X_j|\}$. Smoother matrices S_h are given by $S_h = (K(q(X_i, X_j)/h)/\sum_{j=1}^n K(q(X_i, X_j)/h))_{ij}$, and $\mathbf{S} = \{S_h\}_{h \in]0, 1/2]}$ is an ordered linear smoother. This generalizes to other kernels. However, for nonequidistant, noncircular design the usual definitions of kernel estimators by Nadaraya (1964), Watson (1964) or Gasser and Müller (1984) lead to asymmetric matrices which are not covered by the present approach. The same is true for k-nearest-neighbor estimators.

3. Properties of estimators based on ordered linear smoothers.

Given an ordered linear smoother $\mathbf{S}$, an element $\widehat{S} \in \mathbf{S}$ is selected by (1.3) and $\hat{\mu} = \widehat{S} \cdot Y$ is used as an estimator of μ. It is easily verified that the definition of an ordered linear smoother ensures measurability of $\hat{\mu}$.

3.1. *The function $R_{\mu,\mathbf{S}}$.* Formulation of our main theorem makes use of a function $R_{\mu,\mathbf{S}}$ which depends on the behavior of $\mathrm{MASE}_\mu(S)$, when moving away from S_μ. Thus, before stating the theorem, we will introduce this function and discuss its main properties.

Consider some ordered linear smoother $\mathbf{S}$. For all $S, \widetilde{S} \in \mathbf{S}$, set

$$q_\mu(S, \widetilde{S}) := \left| \frac{1}{n} \mu^T (S - \widetilde{S})^2 \mu + \frac{\sigma^2}{n} \mathrm{tr}\left((S - \widetilde{S})^2\right) \right|^{1/2}.$$

It is easy to check that q_μ defines a metric on $\mathbf{S}$.

For any $\eta > 0$, set

$$R_{\mu,\mathbf{S}}(\eta) := \inf\left\{\varepsilon \geq 0 \;\middle|\; \varepsilon \geq \eta \frac{q_\mu(S,S_\mu)^2}{\mathrm{MASE}_\mu(S) - \mathrm{MASE}_\mu(S_\mu)} \right.$$
$$\left. \text{for all } S \in \mathbf{S} \text{ with } q_\mu(S,S_\mu) > n^{-1/2}\varepsilon\right\}.$$

It is easy to see that, for fixed $\eta > 0$, $R_{\mu,\mathbf{S}}(\eta)$ provides some measure of the steepness of the minimum of MASE_μ at S_μ. For small values of η, $R_{\mu,\mathbf{S}}(\eta)$ will be small if $\mathrm{MASE}_\mu(S)$ is rapidly increasing when moving away from S_μ. It will tend to be large if this is not true. Recall that $\mathrm{MASE}\mu(S) = (1/n)\mu^T(I-S)^2\mu + (\sigma^2/n)\mathrm{tr}(S^2)$.

Let $I \in \mathbf{S}$, and suppose that $\mathbf{S}$ contains more than one element. Large values of $R_{\mu,\mathbf{S}}(\eta)$ also for small η arise if $\mathrm{MASE}_\mu(S) = \mathrm{MASE}_\mu(S_\mu) = \sigma^2$ for all $S \in \mathbf{S}$ (i.e., S_μ may equal any element of $\mathbf{S}$). This implies that $(1/n)\mu^T(S-S_\mu)^2\mu \leq \sigma^2$ and $(\sigma^2/n)\mathrm{tr}((S-S_\mu)^2) \leq \sigma^2$ for all $S \in \mathbf{S}$. By definition of $R_{\mu,\mathbf{S}}$ we thus obtain $R_{\mu,\mathbf{S}}(\eta) = \sup_{S\in\mathbf{S}} n^{1/2} q_\mu(S,S_\mu) \leq (2n\sigma^2)^{1/2}$ for any $\eta > 0$.

In contrast consider the situation that there is a constant $\gamma > 0$ such that $\mathrm{MASE}_\mu(S) - \mathrm{MASE}_\mu(S_\mu) \geq \gamma \cdot q_\mu(S,S_\mu)^2$ for all $S \in \mathbf{S}$. Then $R_{\mu,\mathbf{S}}(\eta) = \eta/\gamma$.

More generally, approximations for $R_{\mu,\mathbf{S}}(\eta)$ are given by the following proposition.

PROPOSITION 1.

(i) *For any ordered linear smoother* $\mathbf{S}$ *it holds that* $\mathrm{MASE}_\mu(S) - \mathrm{MASE}_\mu(S_\mu) \geq q_\mu(S,S_\mu)^2/3$ *for all* $S \in \mathbf{S}$ *with* $q_\mu(S,S_\mu)^2 > 3\,\mathrm{MASE}_\mu(S_\mu)$, *and*

$$R_{\mu,\mathbf{S}}(\eta) \leq \max\left\{\left(3n\,\mathrm{MASE}_\mu(S_\mu)\right)^{1/2}, 3\eta\right\}.$$

(ii) *If for some* $0 \leq \delta, \gamma < \infty$ *it holds that* $\mathrm{MASE}_\mu(S) - \mathrm{MASE}_\mu(S_\mu) \geq \gamma \cdot q_\mu(S,S_\mu)^2$ *for all* $S \in \mathbf{S}_\delta := \{\widetilde{S} \in \mathbf{S} \mid q_\mu(\widetilde{S},S_\mu) > \delta\}$, *then*

$$R_{\mu,\mathbf{S}}(\eta) \leq \max\left\{n^{1/2}\delta, \eta/\gamma\right\}.$$

A proof is contained in the Appendix. Relation (i) reflects the fact that if $\mathrm{MASE}_\mu(S_\mu) \ll \sigma^2$, comparably small values $\mathrm{MASE}_\mu(S)$ can only be achieved by a small number of elements $S \neq S_\mu$. For example, for all S close to the identity matrix, $\mathrm{MASE}_\mu(S) \approx \sigma^2 \gg \mathrm{MASE}_\mu(S_\mu)$. For the special case of kernel estimation, some more precise bounds for $R_{\mu,\mathbf{S}}$ are given in Section 5.

3.2. *Basic results.* Consider the model and definitions given in the Introduction, and recall that $\widehat{\mu}(Y) = \widehat{S}\cdot Y$ constitutes our estimator of μ. The following theorem now establishes the basic result of this paper.

THEOREM 1. *In addition to model* (1.1), *assume that* $\int \exp(\beta x^2) W(dx) < \infty$ *for some* $\beta > 0$. *Then there exist constants* C_W, $\widetilde{C}_W$, $0 < C_W < \infty$, $\widetilde{C}_W < \infty$,

depending only on the distribution W, *such that the following holds for all* $n \in \mathbb{N}$ *and* $\mu \in \mathbb{R}^n$, *each ordered linear smoother* $\mathbf{S} \subset M(n)$ *and any* $\eta > 0$:

(i) $P\Big((1/n)\|\mu - \widehat{S}\cdot Y\|_2^2 - (1/n)\|\mu - \widehat{S}_\mu \cdot Y\|_2^2 > \eta n^{-1} R_{\mu,\mathbf{S}}(\eta)\Big) \le \widetilde{C}_W \cdot \exp(-C_W \cdot \eta)$

(ii) *Relation* (i) *remains true when replacing* $(1/n)\|\mu - \widehat{S}\cdot Y\|_2^2 - (1/n)\|\mu - \widehat{S}_\mu \cdot Y\|_2^2$ *by either* $|(1/n)\|\mu - \widehat{S}\cdot Y\|_2^2 - (1/n)\|\mu - S_\mu \cdot Y\|_2^2|$, $(1/n)\|\mu - S_\mu \cdot Y\|_2^2 - (1/n)\|\mu - S_\mu \cdot Y\|_2^2$, $\mathrm{MASE}_\mu(\widehat{S}) - \mathrm{MASE}_\mu(S_\mu)$ *or* $\mathrm{MASE}_\mu(\widehat{S}_\mu) - \mathrm{MASE}_\mu(S_\mu)$.

The proof of the theorem is given in the Appendix. The key to gaining some intuition into the result consists in noting that all classes $\mathbf{S}$ considered are essentially "one-dimensional" in the sense that $S \to (1/n)\mathrm{tr}(S)$ establishes a homeomorphism from $\mathbf{S}$ into some subset of $[0, 1]$. This then gives rise to a chaining argument which constitutes the major part of the proof. Other requirements like the symmetry of $S \in \mathbf{S}$ are technical conditions.

The proof is based on approximations which in principle allow a direct computation of constants C_W and $\widetilde{C}_W$ for various error distributions W. However, as far as constants are concerned, some of these approximations are very rough. In order to obtain reasonable values for C_W and $\widetilde{C}_W$, a much more detailed analysis will be necessary. The theorem states that stochastic bounds for $(1/n)\|\mu - \widehat{S}\cdot Y\|_2^2 - (1/n)\|\mu - \widehat{S}_\mu \cdot Y\|_2^2$ depend on the respective steepness of the minimum of MASE_μ at S_μ, to be quantified by $R_{\mu,\mathbf{S}}$. This is certainly intuitively plausible.

We can infer from Theorem 1 and Proposition 1(i) that there exist constants $d_1, d_2 < \infty$ such that

$$\begin{aligned}
&\left|\left(E\frac{1}{n}\|\mu - \widetilde{S}\cdot Y\|_2^2\right)^{1/2} - \mathrm{MASE}_\mu(S_\mu)^{1/2}\right| \\
&\qquad \times\left(\left(E\frac{1}{n}\|\mu - \widetilde{S}\cdot Y\|_2^2\right)^{1/2} + \mathrm{MASE}_\mu(S_\mu)^{1/2}\right) \\
&\quad = \left|E\frac{1}{n}\|\mu - \widetilde{S}\cdot Y\|_2^2 - \mathrm{MASE}_\mu(S_\mu)\right| \\
&\quad \le d_1 n^{-1/2}\left(\mathrm{MASE}_\mu(S_\mu)^{1/2} + d_2 n^{-1/2}\right)
\end{aligned}$$

holds for all n, all $\mu \in \mathbb{R}^n$, each ordered linear smoother and both $\widetilde{S} = \widehat{S}$ or $\widetilde{S} = \widehat{S}_\mu$. Relation (1.5) with $m = 1$ is an immediate consequence.

This leads to an interesting result when considering Speckman's minimax spline smoothers of order $2k$, as defined by Example 5 in Section 2. We can then conclude from (2.6) and (1.5) that, as $n \to \infty$,

$$\sup_{\mu \in \mathcal{F}(k,\gamma,n)} E\frac{1}{n}\|\mu - \widehat{S}\cdot Y\|_2^2 = \inf_{\tilde{\mu} \in \mathcal{L}_n} \sup_{\mu \in \mathcal{F}(k,\gamma,n)} \frac{1}{n}E\|\mu - \tilde{\mu}(Y)\|_2^2 \cdot \big(1 + o(1)\big) \tag{3.1}$$

holds for any $0 < \gamma < \infty$. We see that asymptotically minimax risk is attained even if γ is unknown and selection of $\widehat{S}$ relies solely on the data.

4. A generalization. Results similar to those derived in Theorem 1 can be established in a more general framework. This follows when combining the present approach with some of the results of Li (1987). Li considers families $\mathbf{S}$ of smoother matrices which contain a finite number of elements. A straightforward generalization of the concept of ordered linear smoothers consists in considering classes $\mathbf{S}$ of smoother matrices which for some $m \in \mathbb{N}$ satisfy the following:

(4.1) $\mathbf{S} = \mathbf{S}_1 \cup \mathbf{S}_2 \cup \cdots \cup \mathbf{S}_m$, where, for any $i \in \{1, \ldots, m\}$, $\mathbf{S}_i$ is an ordered linear smoother.

The arguments given in the proof of Theorem 1 together with some of the basic ideas of Li (1987) now lead to the following theorem.

THEOREM 2. *In addition to model* (1.1) *assume that* $\int \exp(\beta x^2) W(dx) < \infty$ *for some* $\beta > 0$. *Then there exist constants* $D_W, \widetilde{D}_W, 0 < D_W < \infty, \widetilde{D}_W < \infty$, *depending only on the distribution* W, *such that the following holds for all* $n \in \mathbb{N}$ *and* $\mu \in \mathbb{R}^n$, *each* $\mathbf{S} \subset M(n)$ *satisfying* (4.1) *for some* $m \in \mathbb{N}$, *and any* $\eta > 0$:

$$\text{(i)} \quad P\Big(\frac{1}{n}\|\mu - \widehat{S} \cdot Y\|_2^2 - \frac{1}{n}\|\mu - \widehat{S}_\mu \cdot Y\|_2^2 > \eta^2 n^{-1/2} \mathrm{MASE}_\mu(S_\mu)^{1/2} (\log(m) + 1)^2 + \eta^4 n^{-1} (\log(m) + 1)^4\Big) \leq \widetilde{D}_W \cdot \exp(-D_W \cdot \eta).$$

(ii) *Relation* (i) *remains true when replacing* $(1/n)\|\mu - \widehat{S} \cdot Y\|_2^2 - (1/n)\|\mu - \widehat{S}_\mu \cdot Y\|_2^2$ *by either* $|(1/n)\|\mu - \widehat{S} \cdot Y\|_2^2 - (1/n)\|\mu - S_\mu \cdot Y\|_2^2|$, $(1/n)\|\mu - S_\mu \cdot Y\|_2^2 - (1/n)\|\mu - \widehat{S}_\mu \cdot Y\|_2^2$, $\mathrm{MASE}_\mu(\widehat{S}) - \mathrm{MASE}_\mu(S_\mu)$ *or* $\mathrm{MASE}_\mu(\widehat{S}_\mu) - \mathrm{MASE}_\mu(S_\mu)$.

A proof is given in the Appendix. It may be noted that the approximations given in Theorem 2 are not as precise as those of Theorem 1.

By an argument similar to that used in Section 3 it can easily be derived from Theorem 2 that (1.5) holds for all n and each $\mathbf{S}$ satisfying (4.1) for some $m \in \mathbb{N}$. Theorem 2 has some important consequences, which are best illustrated by some examples.

EXAMPLE 4 [Smoothing splines (continued)]. Assume the conditions of Example 4 in Section 2. When estimating $\mu = (f(x_1), \ldots, f(x_n))^T$ by smoothing splines, a decision has to be made about the order $2k$ of the spline. The choice of k will influence the quality of the resulting estimator. The above results now indicate that one might use the data themselves to select k. One might define $\mathbf{S} = \mathbf{S}_1 \cup \mathbf{S}_2 \cup \cdots \cup \mathbf{S}_n$, where, for $k \in \{1, \ldots, n\}$, $\mathbf{S}_k$ denotes the ordered linear smoother resulting from smoothing splines of order $2k$. We can infer that the resulting estimator $\widehat{S} \cdot Y$ will behave reasonably well. For large n,

$E[(1/n)\|\mu - \widehat{S} \cdot Y\|_2^2]$ will be close to

$$E\frac{1}{n}\|\mu - \widehat{S}_\mu \cdot Y\|_2^2 = E \min_{k \in \{1,\dots,n\}} \min_{S \in \mathbf{S}_k} \frac{1}{n}\|\mu - S \cdot Y\|_2^2.$$

EXAMPLE 5 [Minimax spline smoothers (continued)]. Assume the conditions of Example 5 in Section 2. As for ordinary smoothing splines, the order of minimax spline smoothers might be determined from the data. Thus, let $\mathbf{S} = \mathbf{S}_1 \cup \mathbf{S}_2 \cup \cdots \cup \mathbf{S}_n$, where, for $k \in \{1,\dots,n\}$, $\mathbf{S}_k$ denotes the ordered linear smoother resulting from minimax spline smoothers of order $2k$. Relations (2.6) and (1.5) then imply that, as $n \to \infty$,

$$\begin{aligned}(4.2)\quad \sup_{\mu \in \mathcal{F}(k,\gamma,n)} E\frac{1}{n}\|\mu - \widehat{S} \cdot Y\|_2^2 &= \inf_{\tilde{\mu} \in \mathcal{L}_n} \sup_{\mu \in \mathcal{F}(k,\gamma,n)} \frac{1}{n} E\|\mu - \tilde{\mu}(Y)\|_2^2 \cdot \big(1 + o(1)\big) \\ &= O\big(n^{-2k/(2k+1)}\big)\end{aligned}$$

holds for *any* $k \in \mathbb{N}$ and *each* $\gamma > 0$.

5. Conclusions.

5.1. *Decision theory.* To simplify discussion, let us assume that the error terms ϵ_i in model (1.1) are normally distributed. Then, estimating μ means nothing else but to estimate the mean of a multivariate normal distribution. As outlined in Section 1, no further restrictions are imposed when assuming (1.2) for an arbitrary smooth f. The maximum likelihood estimator of μ is given by $\widehat{\mu} = Y$. We obtain $(1/n)E\|\mu - Y\|_2^2 = \sigma^2$ for any μ. James and Stein (1961) have shown that Y is inadmissible if $n \geq 3$. It is dominated by the estimator $\widehat{\mu} = [1 - \sigma^2(n-2)/Y^TY] \cdot Y$. The James–Stein estimator is minimax, that is, $\sup_{\mu \in \mathbb{R}^n}[1/nE\|\mu - \widehat{\mu}\|_2^2] = \sigma^2$, and at the same time there exists $\mu \in \mathbb{R}^n$ with $(1/n)E\|\mu - \widehat{\mu}\|_2^2 < \sigma^2$.

In Sections 2 and 3 we have considered more general estimators $\widehat{\mu} = \widehat{S} \cdot Y$ which are obtained by using (1.3) to select a smoother matrix $\widehat{S}$ from an ordered linear smoother $\mathbf{S}$. As outlined in Section 2, the simple choice $\mathbf{S} = \{(1-h) \cdot I\}_{h \in [0,1]}$ leads to the estimator $\widehat{\mu} = \widehat{S} \cdot Y = [1 - \sigma^2 n/Y^TY] \cdot Y$, which for large n practically coincides with the James–Stein estimator. The ordered linear smoothers of Examples 2–6 differ from this simple method only insofar as the smoother matrices $S \in \mathbf{S}$ perform different amounts of shrinking in different directions (cf. Section 2). From this point of view any one of the estimators $\widehat{\mu} = \widehat{S} \cdot Y$, obtained, for example, from smoothing splines, kernel estimation and so on, can be considered as a straightforward generalization of the James–Stein estimator. The basic qualitative properties of the latter carry over:

All ordered linear smoothers given in the examples of Section 2 contain the

identity matrix I. For all $\mu \in \mathbb{R}^n$ we thus obtain

$$E\frac{1}{n}\|\mu - \widehat{S}_\mu \cdot Y\|_2^2 \le \mathrm{MASE}_\mu(S_\mu) \le \sigma^2 = E\frac{1}{n}\|\mu - Y\|_2^2.$$

Relation (1.5) now implies that

$$\sup_{\mu \in \mathbb{R}^n} E\frac{1}{n}\|\mu - \widehat{S} \cdot Y\|_2^2 \le \sigma^2 + O(n^{-1/2}). \tag{5.1}$$

This shows that for large n the estimators $\widehat{S} \cdot Y$ of μ possess "almost" minimax risk with respect to quadratic loss. On the other hand, there exist $\mu \in \mathbb{R}^n$ such that $\mathrm{MASE}_\mu(S_\mu) \ll \sigma^2$. For large n the remainder term in (5.1) is negligible, and $\mathrm{MASE}_\mu(S_\mu) \ll \sigma^2$ carries over to $E[(1/n)\|\mu - \widehat{S} \cdot Y\|_2^2] \ll \sigma^2$ by (1.5). We can then conclude that an estimator $\widehat{\mu} = \widehat{S} \cdot Y$, derived from one of these ordered linear smoothers, dominates the estimator $\widehat{\mu} = Y$.

The results of Section 4 imply that these arguments basically still apply if we combine a moderate number of ordered linear smoothers to yield an $\mathbf{S}$ of the form (4.1). Theorem 2 shows, however, that stochastic bounds for the difference $(1/n)\|\mu - \widehat{S} \cdot Y\|_2^2 - \mathrm{MASE}_\mu(S_\mu)$ tend to increase with m.

Which estimation procedures are preferable when assuming (1.2) for some smooth function f? Clearly, this lets us *expect* that, different from the James–Stein method, $\mathrm{MASE}_\mu(S_\mu) \ll \sigma^2$ will hold for any of the smoothing methods considered. We now see that in this situation the basic concept of using smoothing procedures for estimating μ is justified from a decision-theoretic point of view (at least for large n). A reasonable choice among the different smoothing methods is much more difficult. Perhaps, in the context of decision theory, minimax splines as proposed by Speckman (1985) or Nussbaum (1985) are of particular interest. Asymptotically they possess an interesting additional minimax property [cf. (2.6), (3.1) and (4.2)].

5.2. *Asymptotics.* Consider ordered linear smoothers resulting from smoothing procedures like least squares regression, smoothing splines or kernel estimation. In this section we will consider implications of our results in the context of the usual asymptotic theory for nonparametric curve estimates.

Thus, let $n \to \infty$ by assuming models (1.1) and (1.2) for all fixed function f. Suppose that, for all $n \in \mathbb{N}$, we estimate $\mu[\equiv \mu_n = (f(x_1), \ldots, f(x_n))^T]$ by a smoothing procedure defining an ordered linear smoother $\mathbf{S}[\equiv \mathbf{S}(n)]$. We will additionally require that the elements of $\mathbf{S}$ are indexed by a smoothing parameter h such that $\mathbf{S} = \{S_h\}_{h \in H}$. In the following $\widehat{h}$, $h(\mu)$ and $\widehat{h}(\mu)$ will denote the parameters with $S_{\widehat{h}} = \widehat{S}$, $S_{h(\mu)} = S_\mu$ and $S_{\widehat{h}(\mu)} = \widehat{S}_\mu$.

There has been considerable effort to derive rates of convergence for $(1/n\|\mu - S_{\widehat{h}} \cdot Y\|_2^2 - (1/n)\|\mu - S_{\widehat{h}(\mu)} \cdot Y\|_2^2$ and for $|\widehat{h} - \widehat{h}(\mu)|$ in the context of kernel estimators. Most papers concentrate on density estimation, but, for example, Rice (1984) and Härdle, Hall and Marron (1988) derive such rates for kernel regression

estimators. When considering now the results of Section 3, we can immediately derive that, for all ordered linear smoothers and any f,

$$\frac{1}{n}\|\mu - S_{\hat{h}} \cdot Y\|_2^2 - \frac{1}{n}\|\mu - S_{\hat{h}(\mu)} \cdot Y\|_2^2 = O_P\big(n^{-1/2}\,\mathrm{MASE}_{\mu}(S_{\mu})^{1/2}\big). \tag{5.2}$$

Relation (5.2) only provides an upper bound. We obtain faster rates if $R_{\mu,\mathbf{S}}(\eta) = O(\eta)$ for all $\eta > 0$. The following considerations indicate that this can be expected in many situations:

Assume that the S_h are twice continuously differentiable with respect to h, let S_h' and S_h'' denote the corresponding derivatives. Furthermore, suppose that, for sufficiently large n, $h(\mu)$ is in the interior of H. Under suitable conditions on f this assumption is, for example, satisfied for smoothing splines or kernel estimators. It is in no way necessary, but it simplifies life. Based on the respective conditions on f and on resulting asymptotic formulas, it will then often be possible to derive that for n sufficiently large there exist some d, d^*, $0 < d < 1$, $1 < d^* < \infty$, such that

$$\begin{aligned}&\mathrm{MASE}_{\mu}(S_h) \le 3 \cdot \mathrm{MASE}_{\mu}(S_{h(\mu)})\\ &\qquad\qquad\text{if and only if } h \in [d \cdot h(\mu), d^* \cdot h(\mu)] =: H_n^*.\end{aligned} \tag{5.3}$$

Then it follows from Proposition 1(i) that, in order to prove that $R_{\mu,\mathbf{S}}(\eta) = O(\eta)$ for any η, we only have to consider the behavior of $q(S_h, S_{h(\mu)})^2$ and $\mathrm{MASE}_{\mu}(S_h) - \mathrm{MASE}_{\mu}(S_{h(\mu)})$ for $h \in H_n^*$.

Clearly,

$$q(S_h, S_{h(\mu)})^2 = \big(h - h(\mu)\big)^2 \cdot \frac{1}{n}\Big(\mu^T S_{\tilde{h}}' S_{\tilde{h}}' \mu + \sigma^2 \operatorname{tr}\big(S_{\tilde{h}}' S_{\tilde{h}}'\big)\Big) =: \big(h - h(\mu)\big)^2 \cdot q^*(\tilde{h}),$$

for some suitable mean values $\tilde{h}$, By definition of q in Section 3.1 we have $q(S_h, S_{h(\mu)})^2 = O(\mathrm{MASE}_{\mu}(S_{h(\mu)}))$ for $h \in H_n^*$. Consider values of $h \in H_n^*$ such that, for some fixed $c > 0$, $(h - h(\mu))^2/h(\mu)^2 = c$. Then necessarily $h(\mu)^2 \cdot q^*(\tilde{h}) = O(\mathrm{MASE}_{\mu}(S_{h(\mu)}))$. This motivates us to expect that

$$\sup_{h \in H_n^*}\left(\frac{h(\mu)^2 \cdot q^*(h)}{\mathrm{MASE}_{\mu}(S_{h(\mu)})}\right) = O(1). \tag{5.4}$$

Since $\partial \mathrm{MASE}_{\mu}(S_h)/\partial h|_{h = h(\mu)} = 0$, we have

$$\mathrm{MASE}_{\mu}(S_h) - \mathrm{MASE}_{\mu}(S_{h(\mu)}) = \big(h - h(\mu)\big)^2 \frac{\partial^2 \mathrm{MASE}_{\mu}(S_h)}{\partial h^2}\bigg|_{h = \check{h}},$$

for some suitable mean value $\check{h}$. The same arguments as above let us presume that

$$h(\mu)^2 \frac{\partial^2 \mathrm{MASE}_{\mu}(S_h)}{\partial^2 h}\bigg|_{h = \check{h}} = O\big(\mathrm{MASE}_{\mu}(S_{h(\mu)})\big).$$

Now suppose that we can even establish the stronger relation

$$\sup_{h \in H_n^*} \frac{h(\mu)^2 \partial^2 \mathrm{MASE}_\mu(S_h)/\partial^2 h}{\mathrm{MASE}_\mu(S_{h(\mu)})} = O(1),$$
$$1 = O\left(\inf_{h \in H_n^*} \frac{h(\mu)^2\, \partial^2 \mathrm{MASE}_\mu(S_h)/\partial^2 h}{\mathrm{MASE}_\mu(S_{h(\mu)})} \right). \tag{5.5}$$

If (5.3)–(5.5) can be shown to hold, then Proposition 1(i) and the definition of $R_{\mu,\mathbf{s}}$ imply that there exist some $\gamma < \infty$ such that, for all n large enough, $R_{\mu,\mathbf{s}}(\eta) \le \gamma \cdot \eta$ holds for all $\eta > 0$. In this case, Theorem 1 allows us to infer that

$$\left| \frac{1}{n} \|\mu - S_{\hat{h}} \cdot Y\|_2^2 - \frac{1}{n} \|\mu - S_{h(\mu)} \cdot Y\|_2^2 \right| = O_P\left(\frac{1}{n}\right). \tag{5.6}$$

Relation (5.6) remains true if $h(\mu)$ is replaced by $\widehat{h}(\mu)$. Furthermore, Theorem 1 and (5.5) then allow us to derive that

$$\frac{|\widehat{h} - h(\mu)|}{h(\mu)} = O_P\left(\left(n \cdot \mathrm{MASE}_\mu(S_{h(\mu)}) \right)^{-1/2} \right). \tag{5.7}$$

Also, in (5.7) we might replace $h(\mu)$ by $\widehat{h}(\mu)$. If (5.7) holds, one recognizes the interesting effect that the rate of convergence of $|\widehat{h} - h(\mu)|/h(\mu)$ is the slower, the faster the rate of convergence of $\mathrm{MASE}_\mu(S_\mu)$. Let us consider an example.

EXAMPLE 6 [Kernel estimation (continued)]. Assume the conditions of Example 6 in Section 2. Additionally, suppose that the kernel function K is twice continuously differentiable. The following proposition now quantifies rates of convergence under different assumptions on f.

PROPOSITION 2.

(i) *Assume that f possesses finitely many discontinuities at some points $\tau_1 < \tau_2 \cdots < \tau_s \in J = [0,1]$. Furthermore, assume that f is twice continuously differentiable in each of the intervals $(0,\tau_1), (\tau_i, \tau_{i+1})$, $i = 1, \ldots, s-1$, and $(\tau_s, 1)$, and that f has finite left and right first and second derivatives at each τ_i. Then $\mathrm{MASE}_\mu(S_{h(\mu)}) = O(n^{-1/2})$ and $h(\mu) = O(n^{-1/2})$, $n^{-1/2} = O(h(\mu))$. Moreover,* (5.6) *holds, and $|\widehat{h} - h(\mu)|/h(\mu) = O(n^{-1/4})$.*

(ii) *Assume a situation as in* (i) *except that f is not differentiable, but continuous at $\tau_1 < \cdots < \tau_s \in J$. Then $\mathrm{MASE}_\mu(S_{h(\mu)}) = O(n^{-3/4})$ and $h(\mu) = O(n^{-1/4})$, $n^{-1/4} = O(h(\mu))$. Moreover,* (5.6) *holds, and $|\widehat{h} - h(\mu)|/h(\mu) = O(n^{-1/8})$.*

(iii) Assume that f is twice continuously differentiable. Then $\mathrm{MASE}_\mu(S_{h(\mu)}) = O(n^{-4/5})$ and $h(\mu) = O(n^{-1/5})$, $n^{-1/5} = O(h(\mu))$. Moreover, (5.6) *holds, and $|\widehat{h} - h(\mu)|/h(\mu) = O(n^{-1/10})$.*

PROOF. For very large h, $(1/n)\|\mu - S_h \cdot \mu\|_2^2$ is large, while for very small h $\mathrm{var}(S_h(Y))$ is close to σ^2. In between we obtain the approximations $\mathrm{MASE}_\mu(S_h)$

$= \delta_{11}h + \delta_{21}/(nh) + O(h^3 + 1/(n^2h))$, $\mathrm{MASE}_\mu(S_h) = \delta_{12}h^3 + \delta_{22}/(nh) + O(h^4 + 1/(n^2h))$ and $\mathrm{MASE}_\mu(S_h) = \delta_{13}h^4 + \delta_{22}/(nh) + O(1/(n^2h)) + o(h^4)$ for the situations described in (i)–(iii). Here, the δ's denote suitable constants which can be determined by straightforward calculations. This leads to the above conclusions about the rates of convergence of $h(\mu)$ and $\mathrm{MASE}_\mu(S_{h(\mu)})$. Furthermore, (5.3) is an immediate consequence. Some easy computations based on standard approximations show that first-order approximations to $(h^2/n)\mu^T S_h' S_h' \mu$ and $(h^2/n)\mu^T(S_h' S_h' - S_h''(I - S_h))\mu$ are of the form (constant) $\cdot\, h^p + o(h^p)$, where $p = 1, 3, 4$ under (i)–(iii). In addition, approximations to $(h^2/n)\,\mathrm{tr}(S_h' S_h')$ and $(h^2/n)\,\mathrm{tr}(S_h' S_h' + S_h'' S_h)$ are of the form (constant) $\cdot\, 1/(nh) + o(1/nh)$. This establishes (5.4) and (5.5). □

The above rates of convergence coincide with results established by van Es (1992) in the context of kernel density estimation. The results of Proposition 2(iii) have previously been derived by Härdle, Hall and Marron (1988). It should again be noted that in the above proposition we might replace $h(\mu)$ by $\hat{h}(\mu)$ without invalidating results.

6. Unknown variance. In practice usually the problem will arise that the variance σ^2 is unknown. In this section we will thus consider the situation that $\hat{S}$ is determined by (1.3) when σ^2 is replaced there by an estimator $\hat{\sigma}^2$ which is obtained from the data $[\hat{\sigma}^2 \equiv \hat{\sigma}^2(Y)]$. Li (1986, 1987) shows that his results generalize if $\hat{\sigma}^2$ is consistent. Unfortunately, this is not true for the more detailed results of Theorem 1, and we have to analyze this case more closely.

In the context of nonparametric regression (1.2), $d = 1$, one proposal of Rice (1984) is to estimate σ^2 by

$$\hat{\sigma}^2 := \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{(Y_{i+1} - Y_i)^2}{2}.$$

Other possible estimators have been established by Gasser, Sroka and Jennen-Steinmetz (1986) and Hall, Kay and Titterington (1990). Under reasonable conditions on f and on the design $x_1, \ldots, x_n$ they all satisfy $E|\sigma^2 - \hat{\sigma}^2| = O(n^{-1/2})$ as $n \to \infty$. Furthermore, they share a common structural form which leads to the following assumption on the estimator $\hat{\sigma}^2$:

$$\begin{aligned}&\hat{\sigma}^2 := (1/n)Y^T\Sigma_n Y \text{ for some real } n \times n \text{ matrix } \Sigma_n \text{ with}\\ &(1/n)\,\mathrm{tr}(\Sigma_n^T\Sigma_n) \le q_1, 0 < q_1 < \infty,\ E(1/n)\epsilon^T\Sigma_n\epsilon = \sigma^2.\end{aligned} \tag{6.1}$$

For Rice's estimator the following holds:

$$E\hat{\sigma}^2 = \sigma^2 + \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{(\mu_{i+1} - \mu_i)^2}{2} = \sigma^2 + \frac{1}{n-1}\sum_{i=1}^{n-1}\frac{\big(f(x_{i+1}) - f(x_i)\big)^2}{2}.$$

The quality of this estimator depends on μ, $\sup_{\mu \in \mathbb{R}^n}|\sigma^2 - E\hat{\sigma}^2| = \infty$. This holds for all estimators of the form (6.1). Thus, in order to generalize the results

of Theorem 1 to the case of unknown σ^2, we have to restrict the class of all possible μ's to be considered. In other words, we have to rely on an additional model assumption: for some $0 < q_2 < \infty$,

$$\mu \in V_n(q_2) := \left\{ p \in \mathbb{R}^n \,\middle|\, \left| \frac{1}{n} p^T \Sigma_n p \right| \leq q_2 \cdot n^{-1/2} \right\}. \tag{6.2}$$

If the estimator of Rice (1984) is used, (6.2) means that

$$\frac{1}{n-1} \sum_{i=1}^{n-1} \frac{\left(f(x_{i+1}) - f(x_i)\right)^2}{2} \leq q_2 \cdot n^{-1/2}.$$

This is nothing else but a more specific smoothness assumption.

As in the previous sections, we additionally require that $\int \exp(\beta x^2) W(dx) < \infty$ for some β. We then obtain the following generalizations of Theorems 1 and 2.

THEOREM 3. *Under the above assumptions, there exist constants C_{W,q_1,q_2} and $\widetilde{C}_{W,q_1,q_2}$, depending only on W, q_1, q_2, such that the assertions of Theorem 1 hold with $C_W, \widetilde{C}_W$ and $\mu \in \mathbb{R}^n$ being replaced by $C_{W,q_1,q_2}, \widetilde{C}_{W,q_1,q_2}$ and $\mu \in V_n(q_2)$.*

THEOREM 4. *Under the above assumptions, there exist constants D_{W,q_1,q_2} and $\widetilde{D}_{W,q_1,q_2}$, depending only on W, q_1, q_2, such that the assertions of Theorem 2 hold with $D_W, \widetilde{D}_W$ and $\mu \in \mathbb{R}^n$ being replaced by $D_{W,q_1,q_2}, \widetilde{D}_{W,q_1,q_2}$ and $\mu \in V_n(q_2)$.*

A proof of Theorem 3 is contained in the Appendix. Theorem 4 is easily established when combining the arguments used to derive Theorem 3 with those given in the proof of Theorem 2.

It is immediately seen that relation (1.5) generalizes, if $\mu \in \mathbb{R}^n$ is replaced by $\mu \in V_n(q_2)$. The basic conclusions about minimax spline smoothers and smoothing splines still hold. The results of Section 5.2 remain unchanged, since the requirement $\mu \in V_n(q_2)$ does not impose a real restriction in the context of ordinary asymptotic theory in nonparametric regression. Assume that for some compact $J \subset \mathbb{R}$ the design satisfies the weak condition $\sup_i |x_{i+1} - x_i| = o(n^{-1/2})$ as $n \to \infty$, and suppose that σ^2 is determined by one of the methods mentioned above. Then for any $q_2 > 0$ and any fixed, continuously differentiable function f there exists an $n(f) \in \mathbb{N}$ such that $\mu = (f(x_1), \ldots, f(x_n))^T \in V_n(q_2)$ holds for all $n \geq n(f)$. Differentiability is not even necessary. In the situation of Proposition 2(i) we also have $\mu \in V_n(q_2)$ for all n sufficiently large.

It seems to be likely that results similar to Theorems 3 and 4 can be derived for generalized cross-validation [Craven and Wahba (1979)] and related methods. It is well known that such procedures are very closely related to C_L methods which replace σ^2 by an estimate $\widehat{\sigma}^2$. An exact proof is, however, difficult. The nil-trace technique developed by Li (1985) unfortunately does not apply in the present context.

APPENDIX

PROOF OF PROPOSITION 1. Assertion (ii) is an immediate consequence of the definition of $R_{\mu,\mathbf{S}}$. We thus only have to prove assertion (i).

Consider an arbitrary ordered linear smoother $\mathbf{S}$. There then exists an orthogonal matrix U such that for any $S \in \mathbf{S}$ we have $S = U^T \Lambda_s U$ for some diagonal matrix Λ_s with diagonal entries $0 \leq \lambda_{i,s} \leq 1$, $i = 1, \ldots, n$. Since, for any $\alpha, \beta \geq 0$, $(\alpha - \beta)^2 \leq \alpha^2 + \beta^2$ holds, we can conclude that

$$\frac{1}{n}\mu^T(I-S)^2\mu + \frac{1}{n}\mu^T(I-S_\mu)^2\mu \geq \frac{1}{n}\mu^T(S-S_\mu)^2\mu,$$

$$\frac{\sigma^2}{n}\operatorname{tr}(S^2) + \frac{\sigma^2}{n}\operatorname{tr}(S_\mu^2) \geq \frac{\sigma^2}{n}\operatorname{tr}\big((S-S_\mu)^2\big).$$

It follows that for all $S \in \mathbf{S}$ with $q_\mu(S,S_\mu)^2 \geq 3 \cdot \mathrm{MASE}_\mu(S_\mu)$ we obtain $\mathrm{MASE}_\mu(S_\mu) \leq \frac{1}{2}\mathrm{MASE}_\mu(S)$ and

$$3\big(\mathrm{MASE}_\mu(S) - \mathrm{MASE}_\mu(S_\mu)\big) \geq \mathrm{MASE}_\mu(S) + \mathrm{MASE}_\mu(S_\mu) \geq q_\mu(S,S_\mu)^2.$$

Assertion (i) now is a consequence of assertion (ii). □

For the proof of Theorem 1 we need three auxiliary lemmata. Lemma 1 is merely technical. Lemma 2 provides basic inequalities which can be considered as exponential versions of the inequalities given by Whittle (1960). Lemma 3 yields the main tool for providing the theorem. Relying on a chaining argument, exponential probability inequalities are derived which, in the proof of the theorem, allow us to bound certain remainder terms quite uniformly for $\mathbf{S} \in M(n)$, $S \in \mathbf{S}$, $\mu \in \mathbb{R}^n$.

LEMMA 1. *Let X and Z denote independent, real-valued random variables with zero means.*

(a) *If, for some $t > 0$, $E\exp\big(t(X-Z)\big) < \infty$, then $E\exp(tX) < \infty$ and*

$$E\exp(tX) \leq E\exp\big(t(X-Z)\big).$$

(b) *If X and Z are similarly distributed, then $E(X-Z)^k = 0$, $k = 1,3,5,\ldots$, and*

$$E(X-Z)^k \leq 2^k EX^k, \qquad E(X+Z)^k \leq 2^k EX^k, \qquad k = 2,4,6,\ldots,$$

provided these moments exist.

The proof of the lemma is straightforward and thus omitted.

LEMMA 2. *Let W denote a one-dimensional probability distribution with the following properties:*

(i) $\int xW(dx) = 0$, $\int x^2 W(dx) = \sigma^2$;

(ii) $\int \exp(\beta x^2) W(dx) < \infty$ *for some* $\beta > 0$.

Then there exist constants $\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_W$ *and* $\delta_W \in]0, \infty$ [*such that the following hold*:

(a) *For all* $n \in \mathbb{N}, a \in \mathbb{R}^n$ *and i.i.d. random variables* $\epsilon_1, \ldots, \epsilon_n$ *with* $\epsilon_1 \sim W$,

$$P\left(\left|\frac{1}{n}\epsilon^T a\right| > \eta n^{-1/2}\left(\frac{1}{n}a^T a\right)^{1/2}\right) \leq \gamma_1 \exp(-\eta^2 \cdot \gamma_2), \qquad \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T,\ \eta > 0.$$

(b) *For all* $n \in \mathbb{N}$, *any real* $n \times n$ *matrix* A *and i.i.d. random variables* $\epsilon_1, \ldots, \epsilon_n$ *with* $\epsilon_1 \sim W$

$$P\left(\left|\frac{1}{n}\epsilon^T A\epsilon - \frac{\sigma^2}{n}\operatorname{tr}(A)\right| > \eta n^{-1/2}\left(\frac{\sigma^2}{n}\operatorname{tr}(A^T A)\right)^{1/2}\right) < \gamma_3 \exp(-\eta \cdot \gamma_4), \qquad \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T,\ \eta > 0$$

(c) *For all* $n \in \mathbb{N}$, *each* $a \in \mathbb{R}^n$, *any real* $n \times n$ *matrix* A *and i.i.d. random variables* $\epsilon_1, \ldots, \epsilon_n$ *with* $\epsilon_1 \sim W$,

$$P\left(\left|\frac{1}{n}\epsilon^T a + \frac{1}{n}\epsilon^T A\epsilon - \frac{\sigma^2}{n}\operatorname{tr}(A)\right| > \eta n^{-1/2}\left(\frac{1}{n}a^T A + \frac{\sigma^2}{n}\operatorname{tr}(A^T A)\right)^{1/2}\right) \leq \delta_W \exp(-\eta \cdot \gamma w), \qquad \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T,\ \eta > 0.$$

PROOF. We start by considering the moments of ϵ_1. By assumption, $B := E\exp(\beta\epsilon_1^2) < \infty$ for an appropriate $\beta > 0$. We thus obtain $P(|\epsilon_1| > \eta) = P(\epsilon_1^2 > \eta^2) \leq \exp(-\eta^2\beta) \cdot B$ for any $\eta > 0$. For all $k \in \mathbb{N}$ this implies [see, e.g., Serfling (1980), pages 46–47]

$$\begin{aligned} E\epsilon_1^{2k} &\leq 2k \int_0^\infty \eta^{2k-1} B \exp(-\eta^2\beta)\, d\eta \\ &= 2k \int_0^\infty \left(\frac{u}{\beta}\right)^{(2k-1)/2} B \cdot \frac{1}{2\beta}\left(\frac{u}{\beta}\right)^{-1/2} \exp(-u)\, du \\ &= B\left(\frac{1}{\beta}\right)^k \Gamma(k+1). \end{aligned} \tag{A.1}$$

In the following, $\epsilon_1^*, \ldots, \epsilon_n^*$ will denote i.i.d. random variables which are independent of $\epsilon_1, \ldots, \epsilon_n$ and satisfy $\epsilon_1^* \sim W$.

We first prove assertion (a). Choose an arbitrary $a = (a_1, \ldots, a_n)^T \in \mathbb{R}^n$.

Using (A.1) and Lemma 1 yields, for all $t > 0$ and all i,

$$
\begin{aligned}
E\exp\left(\frac{t}{n}a_i(\epsilon_i - \epsilon_i^*)\right) &= 1 + \sum_{j=1}^{\infty}\left(\frac{ta_i}{n}\right)^{2j}\frac{E(\epsilon_i - \epsilon_i^*)^{2j}}{(2j)!} \\
&\le 1 + \sum_{j=1}^{\infty}\left(\frac{ta_i}{n}\right)^{2j} 2^{2j} B\left(\frac{1}{\beta}\right)^j \frac{\Gamma(j+1)}{(2j)!} \\
&\le 1 + \sum_{j=1}^{\infty} 2^j \left(\frac{ta_i}{n}\right)^{2j} B\frac{(1/\beta)^j}{j!} \\
&\le \exp\left(\frac{2B^{1/2}}{\beta}a_i^2\frac{t^2}{n^2}\right)
\end{aligned}
\tag{A.2}
$$

(note that $B \ge 1$). For $\varepsilon > 0$ let $t_\varepsilon := \varepsilon\beta n^2/(4B^{1/2}\cdot a^T a)$. Lemma 1 and (A.2) now lead to

$$
\begin{aligned}
P\left(\frac{1}{n}\sum_{i=1}^{n} a_i\epsilon_i > \varepsilon\right) &\le \exp(-t_\varepsilon\varepsilon)E\exp\left(\frac{t_\varepsilon}{n}\sum_{i=1}^{n} a_i\epsilon_i\right) \\
&\le \exp(-t_\varepsilon\varepsilon)E\exp\left(\frac{t_\varepsilon}{n}\sum_{i=1}^{n} a_i(\epsilon_i - \epsilon_i^*)\right) \\
&\le \exp\left(-t_\varepsilon\varepsilon + \frac{2B^{1/2}}{\beta}\frac{t_\varepsilon^2}{n^2}\sum_{i=1}^{n} a_i^2\right) \\
&= \exp\left(-\frac{\beta}{8B^{1/2}}\frac{n^2}{\sum_{i=1}^{n} a_i^2}\varepsilon^2\right).
\end{aligned}
$$

Applying the same arguments to $(-1/n)\sum_{i=1}^{n} a_i\epsilon_i$ proves the assertion.

Choose an arbitrary real $n \times n$ matrix A. To prove assertion (b), we first determine moments of $(1/n)a^T\epsilon, a \in \mathbb{R}^n$. Assertion (a) implies

$$
E\left(\frac{1}{n}\sum_{j=1}^{n} a_j\epsilon_j\right)^{2k} \le \gamma_1\left(\frac{1}{\gamma_2}\right)^k\left(n^{-2}\sum_{j=1}^{n} a_j^2\right)^k \Gamma(k+1). \tag{A.3}
$$

For $i \in \{1,\dots,n\}$ now let $\psi_i := \sum_{j=1}^{n} a_{ij}(\epsilon_j + \epsilon_j^*)$ and $\xi_i := \epsilon_i - \epsilon_i^*$, where a_{ij} denote the entries of A. Evidently, $\xi_1,\dots,\xi_n$ are independent, symmetric random variables. Moreover, conditional on the sign not being zero, the sign of ξ_i is independent of all ξ_k and ψ_r, $k \ne i$, $r = 1,\dots,n$. Hence,

$$
E\left(\frac{\psi_1}{n}\xi_1\right)^{k_1}\cdots\left(\frac{\psi_n}{n}\xi_n\right)^{k_n} = 0 \tag{A.4}
$$

for $k_1,\dots,k_n \in \mathbb{N}\cup\{0\}$, whenever $k_i/2 \notin \mathbb{N}\cup\{0\}$ for some $i \in \{1,\dots,n\}$.

Using (A.1), (A.3) and Lemma 1, the Hölder inequality in its general form yields, for all $k_1, \ldots, k_n \in \mathbb{N} \cup \{0\}$ with $k_1 + \cdots + k_n = r$,

$$
\begin{aligned}
E\prod_{i=1}^{n}\left(\frac{\psi_i}{n}\xi_i\right)^{2k_i} &\le \left(E\prod_{i=1}^{n}\left(\frac{\psi_i}{n}\right)^{4k_i}\right)^{1/2}\left(E\prod_{i=1}^{n}\xi_i^{4k_i}\right)^{1/2} \\
&\le \left(\prod_{i=1}^{n}\left(E\left(\frac{\psi_i}{n}\right)^{4r}\right)^{4k_i/4r}\right)^{1/2}\left(\prod_{i=1}^{n}E\xi_i^{4k_i}\right)^{1/2} \\
&\le \left(\frac{2^4\gamma_1^{1/2r}B}{n^2\gamma_2\beta}\right)^{r}\left(\prod_{i=1}^{n}\left(\sum_{j=1}^{n}a_{ij}^2\right)^{ki}\right) \\
&\quad \times \Gamma(2r+1)^{1/2}\left(\prod_{i=1}^{n}\Gamma(2k_i+1)^{1/2}\right),
\end{aligned}
\tag{A.5}
$$

noting that $B^{\#\{k_i\, k_i \neq 0\}/2} < B^r$. For $n, r \in \mathbb{N}$ let $K(n,r)$ denote the set of all n-tuples $(k_1, \ldots, k_n) \in (\mathbb{N} \cup \{0\})^n$ with $k_1 + \cdots + k_n = r$. It follows from (A.4) and (A.5) that, for all $t > 0$ sufficiently small,

$$
\begin{aligned}
&E\exp\left(t\sum_{i=1}^{n}\frac{\psi_i}{n}\xi_i\right) \\
&\le 1 + \sum_{r=1}^{\infty}\frac{t^{2r}}{(2r)!}\left(\sum_{(k_1,\ldots,k_n)\in K(n,r)}\frac{(2r)!}{(2k_1)!\cdots(2k_n)!}E\prod_{i=1}^{n}\left(\frac{\psi_i}{x}\xi_i\right)^{2k_i}\right) \\
&\le 1 + \sum_{r=1}^{\infty}t^{2r}\left(\frac{2^4\gamma_1^{1/2r}B}{n^2\gamma_2\beta}\right)^{r}\left(\sum_{(k_1,\ldots,k_n)\in K(n,r)}\frac{(2r)!^{1/2}}{(2k_1)!^{1/2}\cdots(2k_n)!^{1/2}}\right. \\
&\qquad\qquad \left.\times\prod_{i=1}^{n}\left(\sum_{j=1}^{n}a_{ij}^2\right)^{k_i}\right) \\
&\le 1 + \sum_{r=1}^{\infty}t^{2r}\left(\frac{2^5\gamma_1^{1/2r}B}{n^2\gamma_2\beta}\right)^{r}\left(\sum_{(k_1,\ldots,k_n)\in K(n,r)}\frac{r!}{k_1!\cdots k_n!}\prod_{i=1}^{n}\left(\sum_{j=1}^{n}a_{ij}^2\right)^{k_i}\right) \\
&= 1 + \sum_{r=1}^{\infty}t^{2r}\left(\frac{2^5\gamma_1^{1/2r}B}{n^2\gamma_2\beta}\right)^{r}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}^2\right)^{r},
\end{aligned}
\tag{A.6}
$$

noting that, for any $i \in \mathbb{N}$, $i! \le (2i)!^{1/2} \le 2^i i!$. With $\gamma_4^* := (1/\sqrt{2})(\gamma_2\beta/(2^5\beta))^{1/2}$, choose $t_0 := \gamma_4^* n/(\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}^2)^{1/2} = \gamma_4^* n/\mathrm{tr}(A^TA))^{1/2}$. Now, combining Lemma

1 with (A.6) we obtain, for any $\varepsilon > 0$,

$$P\left(\frac{1}{n}\epsilon^T A\epsilon - \frac{\sigma^2}{n}\operatorname{tr}(A) > \varepsilon\right) \le \exp(-t_0\varepsilon)\cdot E\exp\left(t_0\left(\frac{1}{n}\epsilon^T A\epsilon - \frac{\sigma^2}{n}\operatorname{tr}(A)\right)\right)$$
$$\le \exp(-t_0\varepsilon)\cdot E\exp\left(t_0\frac{1}{n}\sum_{i=1}^{n}\frac{\psi_i}{n}\xi_i\right)$$
$$\le \exp(-t_0\varepsilon)\cdot\left(1+\gamma_1^{1/2}\sum_{r=1}^{\infty}\frac{1}{2r}\right).$$

Applying the same argument to $-((1/n)\epsilon^T A\epsilon - \frac{\sigma^2}{n}\operatorname{tr}(A))$ completes the proof of assertion (b). Noting that, for all $x, y > 0$, $x^{1/2} + y^{1/2} \le 2^{1/2}(x+y)^{1/2}$, assertion (c) is an immediate consequence of assertions (a) and (b). □

LEMMA 3. *Under the assumptions of Theorem* 1, *let* p_1 *and* p_2 *denote arbitrary polynomials with the following properties*:

$$\text{(A.7)}\qquad \text{For all } x, y \in [0,1], \quad |p_1(x) - p_1(y)| \le |4(x-y)|, \quad |p_2(x) - p_2(y)| \le |4(x-y)|.$$

Furthermore, for any $x \ge 0$, *let* $]x[$ *denote the smallest integer such that* $]x[\ge x$.

Then, for all $n \in \mathbb{N}$ *and* $\mu \in \mathbb{R}^n$, *each ordered linear smoother* $\mathbf{S}$, *each* $\widetilde{S} \in \mathbf{S}$ *and all* $\eta, \varepsilon > 0$

$$P\left(\frac{1}{n}\epsilon^T\big(p_1(S) - p_1(\widetilde{S})\big)\mu + \frac{1}{n}\epsilon^T\big(p_2(S) - p_2(\widetilde{S})\big)\epsilon - \frac{\sigma^2}{n}\operatorname{tr}\big(p_2(S) - p_2(\widetilde{S})\big)\right.$$
$$\left. > \eta n^{-1/2}\right]\frac{q_\mu(S,\widetilde{S})^2}{\varepsilon^2}\left[\varepsilon \text{ for some } S \in \mathbf{S}\right) \le d_W\exp(-\eta c_W),$$

for some constants $0 < c_W, d_W < \infty$, *depending only on the error distribution* W.

PROOF. For $n \in \mathbb{N}$ consider an arbitrary ordered linear smoother $\mathbf{S} \in M(n)$, and select some arbitrary $\varepsilon > 0$, $\widetilde{S} \in \mathbf{S}$ and $\mu \in \mathbb{R}^n$.

We start by parameterizing the problem. By assumption, for any $S \in \mathbf{S}$ we have $S = U^T\Lambda_S U$ for some orthogonal matrix U and a diagonal matrix Λ_S with diagonal entries $0 \le \lambda_{i,S} \le 1$, $i = 1, \ldots, n$. For all $S, \bar{S} \in \mathbf{S}$ we either have $\lambda_{i,\bar{S}} \le \lambda_{i,S}$, $i = 1, \ldots, n$ or $\lambda_{i,\bar{S}} \ge \lambda_{i,S}$, $i = 1, \ldots, n$. Furthermore, $\mathbf{S}$ is closed.

With $t_u := \sup_{S\in\mathbf{S}}\sum_{i=1}^n \lambda_{i,S} \le n$ and $t_l := \inf_{S\in\mathbf{S}}\sum_{i=1}^n \lambda_{i,S} \ge 0$, this implies that $H\colon \mathbf{S} \to [t_l, t_u]$, given by

$$H(S) := \sum_{i=1}^{n}\lambda_{i,S}, \qquad S \in \mathbf{S},$$

is a continuous injective mapping. For $t \in [t_l, t_u]$ set

$$\bar{S}(t) := H^{-1}(t) \quad \text{if } t \in H(\mathbf{S}),$$

and with

$$z_0(t) := \max_{z \in [t_l, t_u]} \{z \le t \mid z \in H(\mathbf{S})\} \quad \text{and} \quad z_1(t) := \min_{z \in [t_l, t_u]} \{z \ge t \mid z \in H(\mathbf{S})\},$$

set

$$\overline{S}(t) := H^{-1}(z_0(t)) + \frac{t - z_0(t)}{z_1(t) - z_0(t)} \left(H^{-1}(z_1(t)) - H^{-1}(z_0(t))\right)$$

if $t \notin H(\mathbf{S})$. It is immediately seen that $\overline{S}$ defines a homeomorphism from $[t_l, t_u]$ onto $\overline{S}([t_l, t_u]) \supseteq \mathbf{S}$. Moreover, it is easy to verify that, for all $t, t^* \in [t_l, t_u]$ with $t \ge t^*$, we have $\lambda_i(t) \ge \lambda_i(t^*)$, $i = 1, \ldots, n$, where $(\lambda_1(x), \ldots, \lambda_n(x)) = \operatorname{diag}(U\overline{S}(x)U^T)$.

Let $\tilde{t} := H(\tilde{S})$, and for $x, y \in [t_l, t_u]$ set

$$\begin{aligned} \delta(x, y) &:= q_\mu\left(\overline{S}(x), \overline{S}(y)\right) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} \rho_i^2 \left(\lambda_i(x) - \lambda_i(y)\right)^2 + \frac{\sigma^2}{n}\sum_{i=1}^{n} \left(\lambda_i(x) - \lambda_i(y)\right)^2\right)^{1/2}, \end{aligned}$$

where $(\rho_1, \ldots, \rho_n)^T = \rho := U \cdot \mu$. Relation (A.7) now implies that, for all $x, y \in [t_l, t_u]$,

$$\begin{aligned} 4\delta(x, y) &\ge \left(\frac{1}{n}\mu^T \left(p_1\left(\overline{S}(x)\right) - p_1\left(\overline{S}(y)\right)\right)^2 \mu \right. \\ &\qquad \left. + \frac{\sigma^2}{n}\operatorname{tr}\left(\left(p_2\left(\overline{S}(x)\right) - p_2\left(\overline{S}(y)\right)\right)^2\right)\right)^{1/2} \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} \rho_i^2 \left(p_1\left(\lambda_i(x)\right) - p_1\left(\lambda_i(y)\right)\right)^2 \right. \\ &\qquad \left. + \frac{\sigma^2}{n}\sum_{i=1}^{n} \left(p_2\left(\lambda_i(x)\right) - p_2\left(\lambda_i(y)\right)\right)^2\right)^{1/2}. \end{aligned} \tag{A.8}$$

Moreover, we obtain that, for all $x, y, z \in [t_l, t_u]$ with $x \ge y \ge z$,

$$\delta(x, z)^2 \ge \delta(x, y)^2 + \delta(y, z)^2. \tag{A.9}$$

Based on the above parameterization, the proof now relies on a chaining technique similar to that used by Pollard [(1984), pages 144–145].

Relation (A.9) gives rise to the construction of a sequence of finite subsets of $[t_l, t_u]$ which will provide a basis for the chaining argument.

Set $t_{0,1} = \tilde{t}$ and for $i \in \mathbb{N}$ define 3^i points $t_l \le t_{i,1} \le t_{i,2} \le \cdots \le t_{i,3^i} \le t_u$ by the following:

(a) $t_{i,3r-1} = t_{i-1,r}$, $r = 1, \ldots, 3^{i-1}$;

(b) $\delta(t_{i,1}, \tilde{t}) = \min\{\delta(t_l, \tilde{t}), i^{1/2}\varepsilon\}$; $\delta(t_{i,3^i}, \tilde{t}) = \min\{\delta(t_u, \tilde{t}), i^{1/2}\varepsilon\}$;

(c) $\delta(t_{i,3r}, t_{i-1}, r) = \delta(t_{i-1,r}, t_{i-1,r+1}) \cdot 1/\sqrt{3},\ r = 1, \ldots, 3^{i-1} - 1;$

(d) $\delta(t_{i,3r+1}, t_{i-1}, r) = \delta(t_{i-1,r}, t_{i-1,r+1}) \cdot \sqrt{2}/\sqrt{3},\ r = 1, \ldots, 3^{i-1} - 1.$

From (A.9) and (a)–(d) it can be inferred that, for all $i \geq 2$ and all $r \in \{1, \ldots, 3^{i-1} - 1\}$,

$$\delta(t_{i,3r}, t_{i-1,r}) \leq \varepsilon/\left(3^{(i-k)/2}\right), \tag{A.10}$$

$$\delta(t_{i,3r+1}, t_{i-1,r+1}) \leq \varepsilon/\left(3^{(i-k)/2}\right), \tag{A.11}$$

$$\delta(t_{i,3r}, t_{i,3r+1}) \leq \varepsilon/\left(3^{(i-k)/2}\right), \tag{A.12}$$

where $k := \max\{]\delta(t_{i-1,r}, \tilde{t})^2/\varepsilon^2[,]\delta(t_{i-1,r+1}\tilde{t})^2/\varepsilon^2[\}$.

In the following let $T_i := \{t_{i,1}, \ldots, t_{i,3^i}\}$, $T := \cup_{i=0}^{\infty} T_i$ and

$$Z(t) := \frac{1}{n}\epsilon^T p_1\left(\bar{S}(t)\right)\mu + \frac{1}{n}\epsilon^T p_2\left(\bar{S}(t)\right)\epsilon - \frac{\sigma^2}{n}\mathrm{tr}\left(p_2\left(\bar{S}(t)\right)\right), \qquad t \in [t_l, t_u].$$

Evidently, T is a countable dense subset of $[t_l, t_u]$ and for any $\eta > 0$ it holds that

$$\begin{aligned}
&P\left(\frac{1}{n}\epsilon^T\left(p_1(S) - p_1(\tilde{S})\right)\mu + \frac{1}{n}\epsilon^T\left(p_2(S) - p_2(\tilde{S})\right)\epsilon\right. \\
&\qquad \left. - \frac{\sigma^2}{n}\mathrm{tr}\left(p_2(S) - p_2(\tilde{S})\right) > \eta n^{-1/2}\right]\frac{q_\mu(S,\tilde{S})^2}{\varepsilon^2}\left[\varepsilon \text{ for some } S \in \mathbf{S}\right) \\
&\leq P\left(Z(t) - Z(\tilde{t}) \geq \eta n^{-1/2}\right]\frac{\delta(t,\tilde{t})^2}{\varepsilon^2}\left[\varepsilon \text{ for some } t \in [t_l, t_u]\right) \\
&= P\left(Z(t) - Z(\tilde{t}) \geq \eta n^{-1/2}\right]\frac{\delta(t,\tilde{t})^2}{\varepsilon^2}\left[\varepsilon \text{ for some } t \in T\right).
\end{aligned} \tag{A.13}$$

The last equality follows from the continuity of $\bar{S}$.

Thus, to prove the lemma, we only have to show that the last probability adopts the asserted exponential bounds. For $\eta \geq \log(3)$, $i \in \mathbb{N}$ and $k \in \{1, \ldots, i\}$, define events $A_{i,k}^{(1)}(\eta)$ and $A_i^{(2)}(\eta)$ by

$$\begin{aligned}
A_{i,k}^{(1)}(\eta) := \Big\{ &Z(t_{i+1,r}) - Z\left(N_i(t_{i+1,r})\right) \geq n^{-1/2}\delta\left(t_{i+1,r}, N_i(t_{i+1,r})\right)H_{i,k}^{(1)}(\eta) \\
&\text{for some } r \in \left\{2, \ldots, 3^{i+1} - 1\right\} \text{ with } \left]\frac{\delta(t_{i+1,r},\tilde{t})^2}{\varepsilon^2}\right[= k\Big\},
\end{aligned}$$

$$\begin{aligned}
A_i^{(2)}(\eta) := &\left\{Z(t_{i,1}) - Z(\tilde{t}) \geq n^{-1/2}\delta(t_{i,1},\tilde{t})H_i^{(2)}(\eta)\right\} \\
&\vee\left\{Z(t_{i,3^i}) - Z(\tilde{t}) \geq n^{-1/2}\delta(t_{i,3^i},\tilde{t})H_i^{(2)}(\eta)\right\},
\end{aligned}$$

where

$$H_{i,k}^{(1)}(\eta) := \frac{4\left(i\log(6) + k\eta\right)}{\gamma_W} \quad \text{and} \quad H_i^{(2)}(\eta) := \frac{4\left(i^{1/2}\log(3) + i^{1/2}\eta\right)}{\gamma_W},$$

and where, for any $s \in T_{i+1}$, $N_i(s)$ denotes the element of T_i with $\delta(s, N_i(s)) = \min_{t \in T_i} \delta(s,t)$. Constants γ_W and δ_W have to be chosen according to Lemma 2.

Evidently, $s = N_i(s)$ if $s \in T_i \cap T_{i+1}$. For any $i \in \mathbb{N}$ and $k \in \{1, \dots, i\}$, $A_{i,k}^{(1)}(\eta)$ thus consists of at most $4 \cdot 3^{i-k}$ nontrivial events, each of whose probabilities can be bounded using Lemma 2 and (A.8):

$$\begin{aligned} P\big(A_{i,k}^{(1)}(\eta)\big) &\le (4 \cdot 3^{i-k})\delta_W \exp\big(-\gamma_W \tfrac{1}{4} H_{i,k}^{(1)}(\eta)\big) \\ &\le 4 \cdot 2^{-i} 3^{-2k+1} \delta_W \exp(-\eta) \end{aligned}$$

[recall that $\eta \ge \log(3)$]. For $P(A_i^{(2)}(\eta))$ we obtain

$$P\big(A_i^{(2)}(\eta)\big) \le 2\delta_W \exp\big(-\gamma_W \tfrac{1}{4} H_i^{(2)}(\eta)\big) \le 2\delta_W 3^{-2i^{1/2}+1} \exp(-\eta).$$

With $A(\eta)$ denoting the union of all events $A_{i,k}^{(1)}(\eta), A_i^{(2)}(\eta)$, this yields, for all $\eta \ge \log(3)$,

$$\begin{aligned} P\big(A(\eta)\big) &\le \sum_{i=1}^{\infty} P\big(A_i^{(2)}(\eta)\big) + \sum_{i=1}^{\infty} \sum_{k=1}^{i} P\big(A_{i,k}^{(1)}(\eta)\big) \\ &\le \left(6\delta_W \left(\sum_{i=1}^{\infty} 3^{-2i^{1/2}} \right) + \frac{3\delta_W}{2} \right) \exp(-\eta) =: d_{1,W} \exp(-\eta) < \infty. \end{aligned} \tag{A.14}$$

Now, consider an arbitrary $s \in T$, and let m denote the smallest $i - \mathbb{N}$ such that $s \in T_i$. Clearly, $m \ge k :=]\delta(s,t)^2/\varepsilon^2[$. Set $s_m := s$ and $s_j := N_j(s_j + 1)$, for $j = k, \dots, m-1$. Trivially,

$$Z(s) - Z(\tilde{t}) = \left(\sum_{i=k+1}^{m} Z(s_i) - Z(s_{i-1}) \right) + Z(s_k) - Z(\tilde{t}).$$

To bound the sums on the right-hand side, note that either $s_k = t_{k-r,1}$, $r \in \{0,1\}$, or $s_k = t_{k-r,3^{k-r}}$, $r \in \{0,1\}$. On the complement $A(\eta)^C$ of $A(\eta)$ this leads to

$$\begin{aligned} Z(s_k) - Z(t) &\le n^{-1/2} \max\big\{\delta(t_{k,1}, \tilde{t}), \delta(t_{k,3^k}, \tilde{t})\big\} H_k^{(2)}(\eta) \\ &\le n^{-1/2} k^{1/2} \varepsilon \cdot 4k^{1/2} \frac{\big(\log(3) + \eta\big)}{\gamma_W} \end{aligned}$$

and, using (A.10) and (A.11),

$$\begin{aligned} Z(s_i) - Z(s_{i-1}) &\le n^{-1/2} \delta(s_i, s_{i-1}) H_{i-1,k}^{(1)}(\eta) \\ &\le n^{1/2} \frac{\varepsilon}{3^{(i-k)/2}} \cdot \frac{4\big((i-1)\log(6) + k\eta\big)}{\gamma_w}, \qquad i = k+1, \dots, m. \end{aligned}$$

This implies that on $A(\eta)^C, \eta \geq \log(3)$, bounds for $Z(s) - Z(\tilde{t})$ are as follows [note that $\sum_{i=1}^{\infty}(1/3^{i/2}) = 1/(\sqrt{3}-1)$, and $\sum_{i=1}^{\infty}[(i-1)/(3^{i/2})] = 1/(\sqrt{3}-1)^2$]:

$$\begin{aligned} Z(s) - Z(\tilde{t}) &\leq \frac{n^{-1/2}\varepsilon}{\gamma_W} 4\Bigg(k\log(3) + k\eta \\ &\qquad + \sum_{i=k+1}^{m} \left(\frac{k\log(6) + k\eta}{3^{(i-k)/2}} + \frac{(i-k-1)\log(6)}{3^{(i-k)/2}} \right) \Bigg) \\ &< n^{-1/2} k\varepsilon \left(\frac{4}{\gamma_W} \left(\log(3) + \frac{\log(6)}{\sqrt{3}-1} + \frac{\log(6)}{(\sqrt{3}-1)^2} \right) \right) \\ &\quad + n^{-1/2} k\varepsilon\eta \left(\frac{4\sqrt{3}}{\gamma_W(\sqrt{3}-1)} \right) \\ &=: n^{-1/2} k\varepsilon\, c_{1,W} + n^{-1/2} k\varepsilon\eta\, c_{2,W}. \end{aligned}$$

Together with (A.14) this implies that, for any $\eta > 0$,

$$\begin{aligned} &P\left(Z(t) - Z(\tilde{t}) \geq c_{1,W} n^{-1/2} \right] \frac{\delta(t,\tilde{t})^2}{\varepsilon^2} \left[\varepsilon + c_{2,W}\eta n^{-1/2} \right] \frac{\delta(t,\tilde{t})}{\varepsilon^2} \left[\varepsilon \text{ for some } t \in T \right) \\ &\quad \leq d_{2,W} \exp(-\eta), \end{aligned}$$

where $d_{2,W} := \max\{d_{1,W}, 3\}$. Consequently,

$$\begin{aligned} &P\left(Z(t) - Z(\tilde{t}) \geq \eta n^{-1/2} \right] \frac{\delta(t,\tilde{t})^2}{\varepsilon^2} \left[\varepsilon \text{ for some } t \in T \right) \\ &\quad \leq \left(d_{2,W} \exp\left(\frac{c_{1,W}}{c_{2,W}} \right) \right) \exp\left(-\eta \frac{1}{c_{2,W}} \right) =: d_W \exp(-\eta c_W). \end{aligned}$$

Together with (A.13) this yields the desired result. □

PROOF OF THEOREM 1. For some $n \in \mathbb{N}$ consider an arbitrary ordered linear smoother $\mathbf{S}$ and some arbitrary $\mu \in \mathbb{R}^n$.

We start by introducing some notation. For $S \in \mathbf{S}$ set

$$\begin{aligned} Z_1(S) &:= \frac{1}{n}\mu^T(4S - 2S^2)\epsilon + \frac{1}{n}\epsilon^T(2S - S^2)\epsilon - \frac{\sigma^2}{n}\operatorname{tr}(2S - S^2), \\ Z_2(S) &:= \frac{1}{n}2\epsilon^T S\mu + \frac{1}{n}2\epsilon^T S\epsilon - \frac{\sigma^2}{n}\operatorname{tr}(2S), \\ Z_3(S) &:= \frac{1}{n}\epsilon^T(2S - 2S^2)\mu - \frac{1}{n}\epsilon^T S^2\epsilon - \frac{\sigma^2}{n}\operatorname{tr}(S^2); \end{aligned}$$

and, for $\eta, \varepsilon > 0$ and $s = 1, 2, 3$ set

$$\begin{aligned} A_s(\varepsilon,\eta) &:= \left\{ S \in \mathbf{S} \,\middle|\, Z_s(S) - Z_s(S_\mu) \leq \eta n^{-1/2} \right] \frac{q_\mu(S,S_\mu)^2}{\varepsilon^2} \left[\varepsilon \right\}, \\ A_Q &:= \left\{ S \in \mathbf{S} \,\middle|\, Z_1(S) - Z_1(S_\mu) \geq \text{ MASE}_{\,\mu}(S) - \text{ MASE}_{\,\mu}(S_\mu) \right\}. \end{aligned}$$

Recall that $\mathrm{MASE}_\mu(S) = (1/n)\mu^T(I-S)^2\mu + (\sigma^2/n)\operatorname{tr}(S^2)$. Considering the definitions of $\widehat{S}$ and S_μ it is immediately seen that

$$\frac{1}{n}\|Y - S_\mu \cdot Y\|_2^2 + \frac{2\sigma^2}{n}\operatorname{tr}(S_\mu) \geq \frac{1}{n}\|Y - \widehat{S}\cdot Y\|_2^2 + \frac{2\sigma^2}{n}\operatorname{tr}(\widehat{S}), \tag{A.15}$$

$$\mathrm{MASE}_\mu(S_\mu) \leq \mathrm{MASE}_\mu(\widehat{S}). \tag{A.16}$$

Some easy computations show that, for all $S \in \mathbf{S}$,

$$\begin{aligned}\frac{1}{n}\|Y - S\cdot Y\|_2^2 + \frac{2\sigma^2}{n}\operatorname{tr}(S) &= \mathrm{MASE}_\mu(S) - Z_1(S) + \frac{1}{n}2\epsilon^T\mu + \frac{1}{n}\epsilon^T\epsilon \\ &= \frac{1}{n}\|\mu - S\cdot Y\|_2^2 - Z_2(S) + \frac{1}{n}2\epsilon^T\mu + \frac{1}{n}\epsilon^T\epsilon\end{aligned} \tag{A.17}$$

and

$$\frac{1}{n}\|\mu - S\cdot Y\|_2^2 = \mathrm{MASE}_\mu(S) - Z_3(S). \tag{A.18}$$

Together with (A.15) and (A.16) we thus obtain

$$\mathrm{MASE}_\mu(\widehat{S}) - \mathrm{MASE}_\mu(S_\mu) \leq Z_1(\widehat{S}) - Z_1(S_\mu), \tag{A.19}$$

$$\frac{1}{n}\|\mu - \widehat{S}\cdot Y\|_2^2 - \frac{1}{n}\|\mu - S_\mu\cdot Y\|_2^2 \leq Z_2(\widehat{S}) - Z_2(S_\mu), \tag{A.20}$$

$$\frac{1}{n}\|\mu - S_\mu\cdot Y\|_2^2 - \frac{1}{n}\|\mu - \widehat{S}\cdot Y\|_2^2 \leq Z_3(\widehat{S}) - Z_3(S_\mu). \tag{A.21}$$

By (A.19) we obtain $P(\widehat{S}\in A_Q) = 1$. Condition (A.7) of Lemma 3 is fulfilled with either of the following: $p_1 = 4x - 2x^2$ and $p_2 = 2x - x^2$; or $p_1 = 2x$ and $p_2 = 2x$; or $p_1 = 2x - 2x^2$ and $p_2 = -x^2$. Also, it follows that, for all $\varepsilon, \eta > 0$,

$$P\big(\widehat{S}\in A_r(\varepsilon,\eta)\cap A_Q\big) \geq 1 - d_W\exp(-\eta c_W), \qquad r = 1,2,3. \tag{A.22}$$

Definition of $R_{\mu,\mathbf{S}}$ implies that, for any $\eta > 0$,

$$n^{1/2}\big(n^{-1/2}R_{\mu,\mathbf{S}}(\eta)\big) \geq \eta\frac{q_\mu(S,S_\mu)^2}{\mathrm{MASE}_\mu(S) - \mathrm{MASE}_\mu(S_\mu)},$$

for all $S\in\mathbf{S}$ with

$$q_\mu(S,S_\mu) > n^{-1/2}R_{\mu,\mathbf{S}}(\eta).$$

This leads to

$$\begin{aligned}&\mathrm{MASE}_\mu(S) - \mathrm{MASE}_\mu(S_\mu) \\ &\quad\geq \eta n^{-1/2}\frac{q_\mu(S,S_\mu)^2}{\big(n^{-1/2}R_{\mu,\mathbf{S}}(\eta)\big)^2}\big(n^{-1/2}R_{\mu,\mathbf{S}}(\eta)\big) \\ &\quad> \frac{\eta}{2}n^{-1/2}\Bigg]\frac{q_\mu(S,S_\mu)^2}{\big(n^{-1/2}R_{\mu,\mathbf{S}}(\eta)\big)^2}\Bigg[\big(n^{-1/2}R_{\mu,\mathbf{S}}(\eta)\big),\end{aligned}$$

for all $S \in \mathbf{S}$ with $q_\mu(S, S_\mu) > n^{-1/2} R_{\mu,\mathbf{S}}(\eta)$.

We thus can infer that, for all $\eta > 0$,

$$A_1\Big(n^{-1/2}R_{\mu,\mathbf{S}}(\eta), \frac{\eta}{2}\Big) \cap A_Q \subseteq \{S \in \mathbf{S}) | q_\mu(S, S_\mu) \le n^{-1/2} R_{\mu,\mathbf{S}}(\eta)\}. \tag{A.23}$$

Combining (A.20), (A.22) and (A.23) we now obtain

$$\begin{aligned} &P\Big(\frac{1}{n}\|\mu - \widehat{S}\cdot Y\|_2^2 - \frac{1}{n}\|\mu - S_\mu \cdot Y\|_2^2 \le \eta n^{-1} R_{\mu,\mathbf{S}}(\eta)\Big) \\ &\quad \ge P\Big(Z_2(\widehat{S}) - Z_2(S_\mu) \le \eta n^{-1/2}\big(n^{-1/2}R_{\mu,\mathbf{S}}(\eta)\big)\Big) \\ &\quad \ge P\Big(\widehat{S} \in A_2\big(n^{-1/2}R_{\mu,\mathbf{S}}(\eta), \eta\big) \cap A_1\Big(n^{-1/2}R_{\mu,\mathbf{S}}(\eta), \frac{\eta}{2}\Big) \cap A_Q\Big) \\ &\quad \ge 1 - 2d_W \exp\Big(-\eta\frac{c_W}{2}\Big). \end{aligned} \tag{A.24}$$

Moreover, (A.21), (A.22) and (A.23) lead to

$$\begin{aligned} &P\Big(\frac{1}{n}\|\mu - S_\mu\cdot Y\|_2^2 - \frac{1}{n}\|\mu - \widehat{S} \cdot Y\|_2^2 \le \eta n^{-1} R_{\mu,\mathbf{S}}(\eta)\Big) \\ &\quad \ge P\Big(Z_3(\widehat{S}) - Z_3(S_\mu) \le \eta n^{-1/2}\big(n^{-1/2}R_{\mu,\mathbf{S}}(\eta)\big)\Big) \\ &\quad \ge P\Big(\widehat{S} \in A_3\big(n^{-1/2}R_{\mu,\mathbf{S}}(\eta), \eta\big) \cap A_1\Big(n^{-1/2}R_{\mu,\mathbf{S}}(\eta), \frac{\eta}{2}\Big) \cap A_Q\Big) \\ &\quad \ge 1 - 2d_W \exp\Big(-\eta\frac{c_w}{2}\Big). \end{aligned} \tag{A.25}$$

The final step of the proof consists of bounding $(1/n)\|\mu - \widehat{S}_\mu \cdot Y\|_2^2 - (1/n)\|\mu - S_\mu \cdot Y\|_2^2$. Obviously, relations (A.19) and (A.21) remain true when replacing $\widehat{S}$ there by $\widehat{S}_\mu$. Hence, it is immediately seen that in relations (A.22) and (A.25) $\widehat{S}$ also can be replaced by $\widehat{S}_\mu$. Since, by definition of $(1/n)\|\mu - \widehat{S}_\mu \cdot Y\|_2^2$,

$$\begin{aligned} \frac{1}{n}\|\mu - \widehat{S}\cdot Y\|_2^2 - \frac{1}{n}\|\mu - S_\mu \cdot Y\|_2^2 \le &\Big|\frac{1}{n}\|\mu - \widehat{S}\cdot Y\|_2^2 - \frac{1}{n}\|\mu - S_\mu \cdot Y\|_2^2\Big| \\ &+ \frac{1}{n}\|\mu - S_\mu \cdot Y\|_2^2 - \frac{1}{n}\|\mu - \widehat{S}_\mu \cdot Y\|_2^2, \end{aligned}$$

this establishes assertion (i) of Theorem 1. By (A.24) and (A.25), and (A.22) and (A.23), the same probability inequality holds for

$$\frac{1}{n}\|\mu - \widehat{S}\cdot Y\|_2^2 - \frac{1}{n}\|\mu - S_\mu \cdot Y\|_2^2 \quad \text{and} \quad \mathrm{MASE}_\mu\big(\widehat{S}\big) - \mathrm{MASE}_\mu\big(S_\mu\big).$$

Since, in (A.25) and in (A.22) and (A.23), $\widehat{S}$ can be replaced by $\widehat{S}_\mu$ these bounds also apply to $(1/n)\|\mu - S_\mu \cdot Y\|_2^2 - (1/n)\|\mu - \widehat{S}_\mu \cdot Y\|_2^2$, and $\mathrm{MASE}_\mu(\widehat{S}_\mu) - \mathrm{MASE}_\mu(S_\mu)$. This establishes assertion (ii). □

PROOF OF THEOREM 2. For $i \in \{1, \ldots, m\}$ let $\widehat{S}_i$, $S_{\mu,i}$ and $\widehat{S}_{\mu,i}$ denote the minimizers of $(1/n)\|Y - SY\|_2^2 + 2(\sigma^2/n)\operatorname{tr}(S)$, $\mathrm{MASE}_\mu(S)$ and $(1/n)\|\mu - SY\|_2^2$ with respect to $S \in \mathbf{S}_i$. Clearly, there exists some $m_1, m_2, m_3 \in \{1, \ldots, m\}$ such that $\widehat{S} = \widehat{S}_{m_1}, S_\mu = S_{\mu,m_2}$ and $\widehat{S}_\mu = \widehat{S}_{\mu,m_3}$.

Let Z_1 and Z_3 be defined as above. The bounds derived in the proof of Theorem 1 and in Proposition 1(i) allow us to establish the following inequality, which holds for all $n, \mu, \mathbf{S}, m, \eta > 0$, $s = 1, 3$ and both $\widetilde{S}_i = \widehat{S}_i$ or $\widetilde{S}_i = \widehat{S}_{\mu,i}$:

$$P\left(\sup_{i\in\{1,\ldots,m\}} \frac{|Z_s(S_{\mu,i}) - Z_s(\widetilde{S}_i)|}{n^{-1/2}\left(\mathrm{MASE}_\mu(S_{\mu,i}) + 1/n\right)^{1/2}} > \eta^2\right) \le mC_W \exp\left(-\eta\widetilde{C}_W\right). \tag{A.26}$$

Set $Z_1^*(S) = Z_1(S) - (1/n)2\epsilon^T\mu$ and $Z_3^*(S) = Z_3(S)$. Lemma 2(c) and the properties of $\mathbf{S}_i$ then imply the existence of constants $0 < \beta_1, \beta_2 < \infty$, depending only on W, such that, for all $n, \mu, m, \mathbf{S}$, η and $s = 1, 3$,

$$P\left(\sup_{i\in\{1,\ldots,m\}} \frac{|Z_s^*(S_{\mu,i})|}{n^{-1/2}\left(\mathrm{MASE}_\mu(S_{\mu,i}) + 1/n\right)^{1/2}} > \eta\right) \le m\beta_1 \exp(-\eta\beta_2). \tag{A.27}$$

By (A.17) and (A.18) it follows from (A.26) and (A.27) that there are constants $0 < \beta_3, \beta_4 < \infty$, depending only on W, such that, for all n, μ, m, $\mathbf{S}$ and η,

$$P\left(\sup_{i\in\{1,\ldots,m\}} \frac{|V_i|}{n^{-1/2}\left(\mathrm{MASE}_\mu(S_{\mu,i}) + 1/n\right)^{1/2}} > \eta^2\right) \le m\beta_3 \exp(-\eta\beta_4), \tag{A.28}$$

where V_i can be either one of

$$V_i := \left|\frac{1}{n}\|Y - \widehat{S}_i \cdot Y\|_2^2 + \frac{2\sigma^2}{n}\operatorname{tr}(\widehat{S}_i) - \frac{1}{n}\epsilon^T\epsilon\right| - \mathrm{MASE}_\mu(\widehat{S}_i)$$

or

$$V_i := \frac{1}{n}\|\mu - \widetilde{S}_i \cdot Y\|_2^2 - \mathrm{MASE}_\mu(\widetilde{S}_i), \qquad \widetilde{S}_i = \widehat{S}_i, \widehat{S}_{\mu,i}, S_{\mu,i}.$$

Recall the definitions of $\widehat{S}$ $(= \widehat{S}_{m_1})$ and $S_\mu (= S_{\mu,m_2})$. Since

$$\frac{1}{n}\|Y - S_\mu \cdot Y\|_2^2 + \frac{2\sigma^2}{n}\operatorname{tr}(S_\mu) - \frac{1}{n}\|Y - \widehat{S} \cdot Y\|_2^2 - \frac{2\sigma^2}{n}\operatorname{tr}(\widehat{S}) \ge 0,$$

relation (A.28) implies that

$$
\begin{aligned}
&P\Big(\mathrm{MASE}_\mu(\widehat{S}) - \mathrm{MASE}_\mu(S_\mu) \\
&\qquad > \eta^2 n^{-1/2}\big(\mathrm{MASE}_\mu(S_{\mu,m_1})^{1/2} + \mathrm{MASE}_\mu(S_\mu)^{1/2}\big) + 2\frac{1}{n}\eta^2\Big) \\
&\qquad \le m\beta_3 \exp(-\eta\beta_4).
\end{aligned} \tag{A.29}
$$

Since

$$
\mathrm{MASE}_\mu(\widehat{S})^{1/2} + \mathrm{MASE}_\mu(S_\mu)^{1/2} \ge \mathrm{MASE}_\mu(S_{\mu,m_1})^{1/2} + \mathrm{MASE}_\mu(S_\mu)^{1/2},
$$

this leads to

$$
\begin{aligned}
&P\big(\mathrm{MASE}_\mu(\widehat{S})^{1/2} - \mathrm{MASE}_\mu(S_\mu)^{1/2} > 2\eta^2 n^{-1/2}\big) \\
&\qquad \le m\beta_3 \exp(-\eta\beta_4) \quad \text{for } \eta \ge 1,
\end{aligned}
$$

and we can conclude that, for all $\eta \ge 1$,

$$
\begin{aligned}
&P\Big(\mathrm{MASE}_\mu(\widehat{S}) - \mathrm{MASE}_\mu(S_\mu) > 4\eta^2 n^{-1/2}\mathrm{MASE}_\mu(S_\mu)^{1/2} + 4\frac{1}{n}\eta^4\Big) \\
&\qquad \le m\beta_3 \exp(-\eta\beta_4).
\end{aligned} \tag{A.30}
$$

Since $(1/n)\|\mu - S_\mu \cdot Y\|_2^2 - (1/n)\|\mu - \widehat{S}_\mu \cdot Y\|_2^2 \ge 0$, we can infer from (A.28) that relation (A.29) still holds when replacing $\widehat{S}$ by $\widehat{S}_\mu$ and m_1 by m_3 (recall that $\widehat{S}_\mu = \widehat{S}_{\mu,m_3}$). We thus additionally obtain, for all $\eta > 1$,

$$
\begin{aligned}
&P\Big(\mathrm{MASE}_\mu(\widehat{S}_\mu) - \mathrm{MASE}_\mu(S_\mu) > 4\eta^2 n^{-1/2}\mathrm{MASE}_\mu(S_\mu)^{1/2} + 4\frac{1}{n}\eta^4\Big) \\
&\qquad \le m\beta_3 \exp(-\eta\beta_4).
\end{aligned} \tag{A.31}
$$

The assertions of Theorem 2 now follow from (A.28), (A.30) and (A.31). □

PROOF OF THEOREM 3. For $s = 1, 2$, let $\zeta_s(S) := Z_S(S) + 2(\sigma^2/n)\mathrm{tr}(S) - 2(\widehat{\sigma}^2/n)\mathrm{tr}(S)$, where Z_1 and Z_2 are defined as in the proof of Theorem 1. By assumption, $|\sigma^2 - E\widehat{\sigma}^2| \le q_2 n^{-1/2}$, for all $\mu \in V_n(q_2)$. Lemma 2 can be used to obtain an exponential probability bound for

$$
\big|\widehat{\sigma}^2 - E\widehat{\sigma}^2\big| = \Big|\frac{1}{n}\mu^T\Sigma_n\epsilon + \frac{1}{n}\epsilon^T\Sigma_n\mu + \frac{1}{n}\epsilon^T\Sigma_n\epsilon - \frac{\sigma^2}{n}\,\mathrm{tr}\Big(\Sigma_n\Big)\Big|.
$$

Note that $(1/n^2)(\mathrm{tr}(S) - \mathrm{tr}(S_\mu))^2 \le (1/n)\mathrm{tr}((S - S_\mu)^2)$ holds for all S. Together with Lemma 3 this implies the existence of constants $0 < d_{W,q_1,q_2}, c_{W,q_1,q_2} < \infty$, depending only on W and q, such that, for all $n, \mu, \mathbf{S}$, all $\varepsilon, \eta > 0$ and $s = 1, 2$,

$$
\begin{aligned}
&P\left(\zeta_s(S) - \zeta_s(S_\mu) > \eta n^{-1/2}\right]\frac{q_\mu(S,S_\mu)^2}{\varepsilon^2}\Big[\varepsilon \text{ for some } S \in \mathbf{S}\Big) \\
&\qquad \le d_{W,q_1,q_2}\exp(-\eta c_{W,q_1,q_2}).
\end{aligned}
$$

When replacing Z_1 and Z_2 by ζ_1 and ζ_2, $\mu \in \mathbb{R}^n$ by $\mu \in V_n(q_2)$ and d_W and c_W by d_{W,q_1,q_2} and c_{W,q_1,q_2} the rest of the proof of Theorem 3 is now analogous to the proof of Theorem 1. □

Acknowledgment. The author wishes to thank Professor Dennis D. Cox for fruitful discussions.

REFERENCES

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17** 453–555.

COX, D. D. (1988). Approximation of least squares regression on nested subspaces. *Ann. Statist.* **16** 713–732.

CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** 377–403.

EUBANK, R. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.

GASSER, TH. and MÜLLER, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11** 171–185.

GASSER, TH., SROKA, L. and JENNEN-STEINMETZ, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73** 625–633.

HALL, P., KAY, J. M. and TITERINGTON, D. M. (1990). Assymptotically optimal differene based estimation of variance in nonparametric regression. *Biometrika* **77** 521–529.

HALL, P. and MARRON, J. S. (1988). Choice of kernel order in density estimation. *Ann. Statist.* **16** 161–173.

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge Univ. Press.

HÄRDLE, W., HALL, P. and MARRON, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83** 89–99.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* 361–380. Univ. California Press, Berkeley.

LI, K.-C. (1985). From Stein's unbiased risk estimates to the method of generalized cross-validation. *Ann. Statist.* **13** 1352–1377.

LI, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.* **14** 1101–1112.

LI, K.-C. (1987). Asymptotic optimality for C_P, C_L, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.* **15** 958–976.

LI, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17** 1001–1008.

LI, K.-C. and HWANG, J. T. (1984). The data-smoothing aspect of Stein estimates. *Ann. Statist.* **12** 887–898.

MALLOWS, C. L. (1973). Some comments on C_P. *Technometrics* **15** 661–675.

NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **10** 186–190.

NUSSBAUM, M. (1985). Spline smoothing in regression models and asymptotic efficeincy in L_2. *Ann. Statist.* **13** 984–998.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

PRIESTLEY, M. B. and CHAO, M. T. (1972). Non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B* **34** 385–392.

REINSCH, C. (1967). Smoothing by spline functions. *Numer. Math.* **10** 177–183.

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1231.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

SPECKMAN, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.* **13** 970–983.

UTRERAS, F. (1983). Natural spline functions, their associated eigenvalue problem. *Numer. Math.* **42** 107–117.

VAN ES, B. (1992). Asymptotics for least squares cross-validation bandwidths in nonsmooth cases. *Ann. Statist.* **20** 1647–1657.

WAHBA, G. and WENDELBERGER, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly Weather Review* **108** 1122–1143.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359–372.

WHITTLE, P. (1960). Bounds on the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5** 302–305.

INSTITUT DE STATISTIQUE
UNIVERSITÉ CATHOLIQUE DE LOUVAIN
VOIE DU ROMAN PAYS, 34
B-1348 LOUVAIN-LA-NEUVE
BELGIUM