Bernoulli **12**(2), 2006, 271–298

Optimal estimation in additive regression models

JOEL HOROWITZ[1], JUSSI KLEMELÄ[2]*, and ENNO MAMMEN[2]**

[1]*Department of Economics, Northwestern University, 2001 Sheridan Road, Evanston IL 60208-2600, USA. E-mail: joel-horowitz@northwestern.edu*
[2]*Department of Economics, University of Mannheim, L7, 3-5, 68131 Mannheim, Germany. E-mail: *klemela@rumms.uni-mannheim.de; **emammen@rumms.uni-mannheim.de*

This paper is concerned with optimal estimation of the additive components of a nonparametric, additive regression model. Several different smoothing methods are considered, including kernels, local polynomials, smoothing splines and orthogonal series. It is shown that, asymptotically up to first order, each additive component can be estimated as well as it could be if the other components were known. This result is used to show that in additive models the asymptotically optimal minimax rates and constants are the same as they are in nonparametric regression models with one component.

Keywords: exact constants in nonparametric smoothing; kernel estimators; multivariate curve estimation; nonparametric regression; orthogonal series estimator; smoothing splines

1. Introduction

In this paper we discuss a general approach to applying one-dimensional nonparametric smoothers to an additive model. The procedure consists of two steps. In the first step, a fit to the additive model is constructed using the projection approach of Mammen *et al.* (1999). This preliminary estimator uses an undersmoothing bandwidth, so its bias terms are of asymptotically negligible higher order. In the second step, a one-dimensional smoother operates on the fitted values of the preliminary estimator. We will show that this two-step estimator is asymptotically equivalent to the estimator obtained by applying the one-dimensional smoother to a nonparametric regression model that only contains one component.

We consider an additive regression model with observations

$$Y^i = m_0 + m_1(X_1^i) + \ldots + m_D(X_D^i) + \varepsilon^i, \qquad i = 1, \ldots, n. \tag{1.1}$$

Here Y^i are the response variables. The covariates $X^i = (X_1^i, \ldots, X_D^i)$ are assumed to be random and to be independently and identically distributed. The density of X^i is denoted by p. Our results could be extended to a non-random design that approximately follows a smooth design density p. For simplicity, we assume that the covariates $X_1^i, \ldots, X_D^i$ are one-dimensional. We suppose that they lie in a bounded set that, without loss of generality, is equal to [0, 1]. Also for simplicity we assume that the error variables ε^i are independent, identically distributed and independent of $X_1^i, \ldots, X_D^i$ with $\mathrm{E}(\varepsilon^i) = 0$ and $\mathrm{var}(\varepsilon^i) = \sigma^2$. Our

results can be extended to the case of independent non-identically distributed error variables or to dependent errors that fulfil mixing conditions. The constant m_0 and the functions $m_1, \ldots, m_D$ are unknown. For identifiability we assume that

$$\mathrm{E}m_d(X_d^i) = 0, \qquad d = 1, \ldots, D. \tag{1.2}$$

We will discuss optimal estimation of an additive component, m_1, say. Many smoothing estimators are available if there is only one component, that is, $D = 1$. We will show that asymptotically m_1 can be estimated as well for $D > 1$ as for $D = 1$. In other words, consider an additive model with $D > 1$ and suppose that $m_2, \ldots, m_D$ are known. Then an estimator $\hat{m}$ of m_1 can be constructed using the data (X_1^i, Z^i), where

$$\begin{aligned} Z^i &= Y^i - m_2(X_2^i) - \ldots - m_D(X_D^i) \\ &= m_0 + m_1(X_1^i) + \varepsilon^i. \end{aligned} \tag{1.3}$$

The notation for estimates of m_1 does not include the subscript 1 because, without loss of generality, we treat only estimation of m_1. For each member of a broad class of estimators $\hat{m}$ of m_1 in the single-component model (1.3), we show that there is an estimator $\tilde{m}$ of m_1 in the multi-component model (1.1) that does not require knowledge of the functions $m_2, \ldots, m_D$ and is asymptotically equivalent to $\hat{m}$. Thus, when estimating an additive component such as m_1 in a nonparametric regression model, one can do as well asymptotically when the other components are unknown as when they are known. In particular, $\hat{m}$ can be chosen with asymptotically minimax L_2 risk for m_1 in a Sobolev ball. In the general model (1.1) where all additive components are unknown, we call $\hat{m}$ an 'oracle estimator'. It is clear that one cannot do better than $\hat{m}$ in model (1.1). Our result implies that in model (1.1) there is an estimator of m_1 that achieves the oracle lower bound. That is, the estimator has the same first-order asymptotic L_2 risk as $\hat{m}$. We emphasize that this result is specific to nonparametric estimation. It does not hold if the additive components are known up to finite-dimensional parameters as in a linear model.

Our result generalizes a classical result on optimal rates of estimation in nonparametric additive models. Stone (1985) showed that each additive component in an additive model can be estimated with the asymptotically optimal L_2 rate of convergence that is achievable in the one-dimensional model (1.3). This was one of the main motivations for the use of model (1.1). It was argued that additive fits give good insight into multivariate structure while avoiding the curse of dimensionality.

Backfitting is a popular method for estimating nonparametric additive models; see Hastie and Tibshirani (1991). Backfitting is based on iterative updates of the additive components. A one-dimensional smoother is applied at each step. Opsomer and Ruppert (1997), Opsomer (2000) and Mammen *et al.* (1999) have developed asymptotic theory for backfitting estimators. Opsomer (2000) considered backfitting estimators that use local polynomial estimators in the iterative updates of the additive components. He showed that the backfitting estimator has a normal limiting distribution with the same variance as an oracle local polynomial estimator (that is, a local polynomial estimator that uses knowledge of the additive components $m_2, \ldots, m_D$). However, the backfitting and oracle estimators have different biases. Mammen *et al.* (1999) showed that after a modification of the

backfitting algorithm, the estimators have the same asymptotic normal distribution. Linton (2000) and Horowitz and Mammen (2004) present other two-step procedures that have the same asymptotic bias and variance as an oracle estimator.

The asymptotically optimal L_2 risk over Sobolev balls for model (1.1) with $D = 1$ or, equivalently, model (1.3) is well known. Nussbaum (1985) studied minimax risks for equidistant designs and Gaussian noise. Golubev and Nussbaum (1990) and Efromovich (1996) treated general designs and non-Gaussian errors. All these authors use the fact that Sobolev balls can be represented as infinite-dimensional ellipsoids. Then they apply a result of Pinsker (1980), who showed that minimax linear estimators over an ellipsoid are asymptotically minimax in the class of all estimators. Belitser and Levit (1995) provide an elementary proof of Pinsker's result. Golubev (1992) considered minimax risks for projection pursuit regression models. These models generalize additive models by replacing the arguments $X_1^i, \ldots, X_D^i$ in the functions $m_1, \ldots, m_D$ with projections $a_1^{\mathrm{T}} X^i, \ldots, a_D^{\mathrm{T}} X^i$ with unknown projection vectors $a_1, \ldots, a_D$. But Golubev (1992) treated only a regular grid of equidistant design points. An equidistant design greatly simplifies the asymptotic analysis of an additive model, because the additive components can then be estimated directly by non-iterative smoothing methods. Nonetheless, the results of Golubev (1992) motivate our analysis, because one may conjecture that asymptotically the difference between regression models with equidistant and non-equidistant designs vanishes in the limiting white noise experiment.

The remainder of this paper is organized as follows. Section 2 discusses the construction of the preliminary estimator. Sections 3, 4 and 5, respectively, apply our approach to kernel, smoothing spline and orthogonal series estimators. Section 6 presents our result on asymptotic minimax estimation. Section 7 presents results of simulation experiments using our approach to kernel smoothing. The proofs of our results are given in Section 8.

2. Preconditioning by presmoothing

Many smoothing methods are not affected to first order if observations are replaced by averages over local neighbourhoods (presmoothing). This holds if the length of the local neighbourhood is of higher order (converges to zero more rapidly) in presmoothing than in final smoothing. This fact is the starting point of our comparison of experiments with observations Y^i from the additive model (1.1) and observations Z^i from the oracle model (1.3). We will replace Z^i by local averages $\hat{Z}^i$ and will show that, with many smoothing operators, smoothing of Z^i and $\hat{Z}^i$ leads to asymptotically equivalent estimators. In addition, in this section we construct values $\hat{Y}^i$ that are based on backfitting of Y^i and have only higher-order differences from $\hat{Z}^i$. We use this result in the sections that follow to show that the results of smoothing $\hat{Y}^i$ and $\hat{Z}^i$ are the same up to first order. This is a general strategy for showing that, for a smoothing estimator based on Z^i, there is a corresponding asymptotically equivalent estimator based on Y^i.

We now give a definition of the presmoothing values $\hat{Z}^i$. We divide $[0, 1]$ into L_n intervals $I_l = ((l-1)/L_n, l/L_n]$, $l = 1, \ldots, L_n$. The indicator functions of these intervals are denoted by $I_l(x)$. In our applications in the following sections, L_n will be chosen such

that L_n^{-1} is of smaller order than the smoothing of the estimator that is considered. Our presmoothing estimator $\hat{Z}^i$ is defined as an average of $Z^i - \overline{Z}$:

$$\hat{Z}^i = \frac{\sum_{r=1}^n (Z^r - \overline{Z}) I_l(X_1^r)}{\sum_{r=1}^n I_l(X_1^r)}, \qquad \text{for } X_1^i \in I_l. \tag{2.1}$$

Here, $\overline{Z}$ denotes the overall average $n^{-1}\sum_{i=1}^n Z^i$. By construction, for each value of l, the fitted value $\hat{Z}^i$ is constant for $X_1^i \in I_l$. Therefore, $\hat{Z}^i$ may also be denoted by $\hat{\tau}_Z(l)$. This is a regressogram estimator.

We now construct $\hat{Y}^i$. The construction is based on a least-squares fit of an additive model. Define $\hat{\tau}_{Y,1}(1), \ldots, \hat{\tau}_{Y,d}(L_n)$ as minimizers of

$$\sum_{i=1}^n \left[Y^i - \overline{Y} - \sum_{d=1}^D \sum_{l=1}^{L_n} \tau_{Y,d}(l) I_l(X_d^i) \right]^2. \tag{2.2}$$

Here, $\overline{Y}$ denotes the overall average $n^{-1}\sum_{i=1}^n Y^i$. Observe also that $\hat{\tau}_Z(l)$ minimizes the least-squares criterion function:

$$\sum_{i=1}^n \left[Z^i - \overline{Z} - \sum_{l=1}^{L_n} \tau_Z(l) I_l(X_1^i) \right]^2.$$

In (2.2) we use the same interval partition I_l ($l = 1, \ldots, L_n$) for all additive components. This has been done for simplification of notation. For applications the arguments can easily be extended to partitions that depend on d.

Because our focus is on the estimation of the first additive component, m_1, we write $\hat{\tau}_Y(l)$ instead of $\hat{\tau}_{Y,1}(l)$ for $l = 1, \ldots, L_n$. For $X_1^i \in I_l$, the value of $\hat{\tau}_Y(l)$ is also denoted by $\hat{Y}^i$.

Our first result makes use of the following assumptions.

(A1) $L_n n^{-1/2} \log n \to 0$ as $n \to \infty$.

(A2) All covariates take values in a bounded interval, [0, 1], say. The one- and two-dimensional marginal densities p_d and $p_{d,d'}$ of X_d^i and $(X_d^i, X_{d'}^i)$, respectively, are bounded away from zero and infinity.

(A3) For all additive components m_d the first derivative exists almost surely and satisfies $\int m_d'(x)^2 \, dx < \infty$ for $d = 1, \ldots, D$.

(A4) There is a finite constant $C > 0$ such that $|m_d(x) - m_d(y)| \leqslant C|x - y|$ for each $d = 1, \ldots, D$. Moreover, $\mathrm{E}[|\varepsilon^i|^{2+\delta}] < \infty$ for each $i = 1, \ldots, n$ and some $\delta > 0$.

The next theorem gives a bound for the difference between $\hat{Z}^i$ and $\hat{Y}^i$. In the statement of the theorem the conditional expectation given the covariates $X^1, \ldots, X^n$ is denoted by E^*.

Theorem 1. *Suppose that model* (1.1) *holds with* (1.2) *and that* Z^i *is defined by* (1.3).

(i) *Under assumptions* (A1), (A2) *and* (A3),

$$\mathrm{E}^*\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}^i - \hat{Z}^i\right)^2\right] = O_P(L_n^{-2}). \tag{2.3}$$

(ii) *Under assumptions* (A1), (A2) *and* (A4),

$$\max_{1\leqslant i\leqslant n}|\hat{Y}^i - \hat{Z}^i| = O_P(L_n^{-1}). \tag{2.4}$$

Results similar to Theorem 1 hold for other constructions of the presmoothers $\hat{Y}^i$ and $\hat{Z}^i$. In particular, $\hat{Y}^i$ and $\hat{Z}^i$ could be constructed by using kernel smoothers or other orthogonal series estimates. Kernel estimators should be undersmoothed, and overfitting should be used for orthogonal series estimators. Mammen *et al.* (1999) discuss kernel smoothing of additive models. Kernel estimators make sense if all additive components $m_1, \ldots, m_d$ satisfy smoothness conditions that are stronger than (A3) or (A4). We conjecture that sharper bounds on $\hat{Y}^i - \hat{Z}^i$ can be obtained under these stronger conditions. We do not pursue this here because our interest is in optimal estimation of m_1 under minimal smoothness assumptions about the other components. In particular, we want to allow the possibility that m_1 is smoother than the other components. In practice, the choice of the presmoother should depend on the method that is used in the second step. For example, it is natural to use a kernel presmoother if kernel smoothing is used in the second step. In this paper, we discuss only the piecewise constant presmoothing of (2.1) and (2.2). This approach yields a simple theory and works well, at least asymptotically, for a broad class of smoothing methods and smoothness assumptions. Practical questions about our approach will be discussed in another paper.

The conclusions of Theorem 1 hold uniformly over (sequences of) design densities and regression functions if assumptions (A2)–(A4) are replaced with uniform versions. Instead of (A2), one assumes that the one- and two-dimensional marginals are uniformly bounded away from zero and infinity. Instead of (A3), one requires the L_2 norm of the first derivative to be uniformly bounded. In (A4), the same constant C must apply uniformly for all additive components. Under these conditions, the conclusions of Theorem 1 apply to all sequences of such functions. In part (i) of the theorem, for example, this is equivalent to the statement that for every $\delta > 0$ there are a finite $c > 0$ and $n_0 > 0$ such that

$$\sup_p \sup_{m_1,\ldots,m_D} P\left(L_n^2 \mathrm{E}^* \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}^i - \hat{Z}^i\right)^2 > c\right) \leqslant \delta$$

for $n \geqslant n_0$. The suprema are over all design densities and additive components that satisfy the regularity conditions.

3. Kernel smoothing in additive models

Suppose $\hat{m}$ is of the form

$$\hat{m}_K(x) = n^{-1} \sum_{i=1}^{n} w_i(x)(Z^i - \overline{Z})$$

with random functions w_i that may depend on $X_1^1, \ldots, X_1^n$ but that are independent of $\varepsilon^1, \ldots, \varepsilon^n$. We compare this estimator with

$$\tilde{m}_K(x) = n^{-1} \sum_{i=1}^{n} w_i(x)\hat{Y}^i.$$

We will now state assumptions under which the differences between these estimators are asymptotically negligible relative to $\hat{m}_K - m_1$.

(A5) The following conditions hold:

$$\sup_{0 \leqslant x \leqslant 1} n^{-1} \sum_{i=1}^{n} |w_i(x)| = O_P(1),$$

$$\sup_{0 \leqslant i \leqslant n} \int_0^1 |w_i(x)| \mathrm{d}x = O_P(1).$$

(A6) There exist (random) functions u_r ($1 \leqslant r \leqslant n$), depending on $X_1^1, \ldots, X_1^n$ and n, and a sequence $\rho_n > 0$ with

$$\sup_{1 \leqslant s \leqslant n, |X_1^r - X_1^s| \leqslant L_n^{-1}} |w_r(x) - w_s(x)| \leqslant L_n^{-1} u_r(x), \tag{3.1}$$

$$n^{-1} \int_0^1 \sum_{i=1}^{n} u_i(x)^2 \, \mathrm{d}x = O_P(\rho_n^2). \tag{3.2}$$

(A7) The condition:

$$\sup_{0 \leqslant x \leqslant 1, 1 \leqslant i \leqslant n} |u_i(x)| = O_P(n^{1/2-\gamma} \rho_n^{1/2}) \tag{3.3}$$

holds for a constant $\gamma > 0$ with $\mathrm{E}|\varepsilon^i|^\eta < \infty$ for some $\eta > 1/\gamma$.

Theorem 2. *Suppose that model* (1.1) *holds with* (1.2) *and that* Z^i *is defined by* (1.3).

(i) *Under assumptions* (A1)–(A3), (A5) *and* (A6)

$$\int_0^1 [\hat{m}_K(x) - \tilde{m}_K(x)]^2 \, \mathrm{d}x = O_P(L_n^{-2} \rho_n^2 n^{-1} + L_n^{-2}). \tag{3.4}$$

(ii) *Under assumptions* (A1), (A2) *and* (A4)–(A7),

$$\sup_{0 \leqslant x \leqslant 1} |\hat{m}_K(x) - \tilde{m}_K(x)| = O_P(L_n^{-1} \rho_n n^{-1/2} \sqrt{\log n} + L_n^{-1}). \tag{3.5}$$

We now briefly discuss the application of Theorem 2 to Nadaraya–Watson and local polynomial estimators. A local polynomial estimator $\hat{m}^j_{\mathrm{LP},q,h}$ is defined by the following minimization problem:

$$(a_0, \ldots, a_q) = \arg\min \sum_{i=1}^{n} \left[Z^i - \bar{Z}^i - a_0 - (X_1^i - x)a_1 - \ldots - (X_1^i - x)^q a_q \right]^2 K_h(X_1^i - x).$$

We denote the minimum by $\hat{m}^j_{\mathrm{LP},q,h}(x) = a_j$. Here $K_h(u) = h^{-1}K(h^{-1}u)$ is a kernel with bandwidth h. The quantity $\hat{m}^j_{\mathrm{LP},q,h}(x)$ is the local polynomial estimator of the jth derivative of m_1 at x. We now define $\tilde{m}^j_{\mathrm{LP},q,h}(x)$ as $\hat{m}^j_{\mathrm{LP},q,h}(x)$ but with $Z^i - \bar{Z}^i$ replaced by $\hat{Y}^i$. We make the following assumptions on the kernel K and bandwidth h.

(A8) The kernel function K is a probability density function with a compact support, say $[-1, 1]$, and an absolutely bounded derivative. The density p_1 is continuous.

(A9) For constants $c_1 > 0$, $c_2 > 0$ and $0 < \alpha_2 \leqslant \alpha_1 < 1/3$, the bandwidth satisfies

$$c_1 n^{-\alpha_1} \leqslant h \leqslant c_2 n^{-\alpha_2}.$$

In addition, in part (ii) of Theorem 3 we assume that:

(A10) $\mathrm{E}|\varepsilon^i|^\eta < \infty$ for some $\eta > 2(1 - \alpha_1)^{-1}$.

The following theorem states our results for local polynomial smoothing.

Theorem 3. *Suppose that model* (1.1) *holds with* (1.2) *and that* Z^i *is defined by* (1.3).

(i) *Let assumptions* (A1)–(A3), (A8) *and* (A9) *hold. For* $q \geqslant 1$ *and* $0 \leqslant j \leqslant q$,

$$h^{2j} \int_0^1 \left[\hat{m}^j_{\mathrm{LP},q,h}(x) - \tilde{m}^j_{\mathrm{LP},q,h}(x) \right]^2 \mathrm{d}x = O_P(L_n^{-2}). \tag{3.6}$$

(ii) *Let assumptions* (A1), (A2), (A4), (A8) *and* (A10) *hold. The following result holds for* c_1, $c_2 > 0$, $0 < \alpha_2 \leqslant \alpha_1 < 1/3$, $q \geqslant 1$ *and* $0 \leqslant j \leqslant q$.

$$\sup_{0 \leqslant x \leqslant 1, c_1 n^{-\alpha_1} \leqslant h \leqslant c_2 n^{-\alpha_2}} h^j |\hat{m}^j_{\mathrm{LP},q,h}(x) - \tilde{m}^j_{\mathrm{LP},q,h}(x)| = O_P(L_n^{-1}). \tag{3.7}$$

Theorem 3 shows that the asymptotic theory of local polynomials in the classical nonparametric regression model carries over to our estimator in the additive model. The first-order difference between our two-step estimator and the oracle estimator can be made to be of nearly parametric order. Theorem 3 can be also applied to plug-in data-adaptive bandwidth choices. For an r times differentiable regression function the bandwidth h that minimizes the asymptotic mean integrated square error depends on known quantities, the variance of ε and $\int m_1^{(r)}(x)^2 \, \mathrm{d}x$. The variance can be estimated in the additive model by average of the squared residuals. In a conventional nonparametric regression model, $\int m_1^{(r)}(x)^2 \, \mathrm{d}x$ can be estimated consistently by $\int \hat{m}^r_{\mathrm{LP},r,h}(x)^2 \, \mathrm{d}x$ if $\int m_1^{(r)}(x)^2 \, \mathrm{d}x < \infty$ and, as $n \to \infty$, $h \to 0$ and $nh^r \to \infty$. Theorem 3(i) implies that in the additive model, the estimator $\int \tilde{m}^r_{\mathrm{LP},r,h}(x)^2 \, \mathrm{d}x$ is consistent under the same conditions. Theorem 3(ii) shows that the asymptotic performance of the local polynomial estimator $\tilde{m}^0_{\mathrm{LP},r,h}$ is the same with the plug-in bandwidth as it is with the asymptotically optimal bandwidth. This is because the asymptotic performance of the oracle estimator $\hat{m}^0_{\mathrm{LP},r,h}$ is the same with the plug-in and

asymptotically optimal bandwidths. Note that we obtain all of these results using only Lipschitz continuity (assumption A4) of the components $m_2, \ldots, m_D$.

4. Smoothing splines in additive models

Now suppose that $\hat{m}$ is a smoothing spline. That is,

$$\hat{m}_S = \arg\min_m \|Z - \overline{Z} - m\|_n^2 + \lambda_n^2 J_k^2(m),$$

where

$$\|Z - \overline{Z} - m\|_n^2 = \frac{1}{n}\sum_{i=1}^n \left[Z^i - \overline{Z} - m(X_1^i)\right]^2$$

and

$$J_k^2(m) = \int m^{(k)}(u)^2 \, \mathrm{d}u.$$

We make the following assumption about the smoothing parameter and the error distribution.

(A11) The smoothing parameter λ_n satisfies $\lambda_n^{-1} = O_P(n^{k/(2k+1)})$. It may be random and it is allowed to depend on the data. The error variables have sub-Gaussian tails. This means $\mathrm{E}\exp(t\varepsilon_i^2)$ is finite for $t > 0$ small enough. The function m_1 has a kth derivative with $J_k(m_1) < \infty$.

Under (A11) (see van de Geer 2000, Theorem 10.2) the estimator $\hat{m}_S$ achieves rate λ_n:

$$\|\hat{m}_S - m_1\|_n = O_P(\lambda_n), \tag{4.1}$$

$$J_k(\hat{m}_S) = O_P(1). \tag{4.2}$$

The estimator $\hat{m}_S$ achieves an optimal rate if λ_n is of order $n^{-k/(2k+1)}$. Then $\|\hat{m}_S - m_1\|_n = O_P(n^{-k/(2k+1)})$.

We compare the estimate $\hat{m}_S$ with

$$\tilde{m}_S = \arg\min_m \|\hat{Y} - m\|_n^2 + \lambda_n^2 J_k^2(m).$$

The next theorem states that the differences between these estimates are asymptotically negligible.

Theorem 4. *Suppose that model* (1.1) *holds with* (1.2) *and that* Z^i *is defined by* (1.3)*. Then, under assumptions* (A1)–(A3) *and* (A11)*,*

$$\|\hat{m}_S - \tilde{m}_S\|_n^2 = O_P(n^{-1/2} L_n^{-(2k-1)/(2k)} + L_n^{-2}), \tag{4.3}$$

$$\int_0^1 [\hat{m}_S(x) - \tilde{m}_S(x)]^2 \, dx = O_P(n^{-1/2} L_n^{-(2k-1)/(2k)} + L_n^{-2}), \tag{4.4}$$

$$J_k^2(\hat{m}_S - \tilde{m}_S) = O_P(\lambda_n^{-1} n^{-1/2} L_n^{-(2k-1)/(2k)} + L_n^{-2}). \tag{4.5}$$

The right-hand side of (4.3) is $o_P(n^{-2k/(2k+1)})$ if $L_n^{-1} = o(n^{-k/(2k+1)})$. Such a choice is possible under our conditions. So we obtain the result that in the additive model an estimate can be constructed that is asymptotically equivalent to a smoothing spline in the oracle model (1.3). If the smoothing parameter λ_n is of optimal order $n^{-k/(2k+1)}$, we further have that $J_k^2(\hat{m}_S - \tilde{m}_S) = o_P(1)$.

5. Orthogonal series estimates in additive models

We now consider orthogonal series estimates. For basis functions $e_{n,j}$, the estimate $\hat{m}_O$ is defined as

$$\hat{m}_O(x) = \sum_{j=1}^{J_n} \lambda_{n,j} \hat{\theta}_{n,j} e_{n,j}, \tag{5.1}$$

where $\hat{\theta}_n = (\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,J_n})$ minimizes

$$\frac{1}{n} \sum_{i=1}^{n} \left[Z^i - \bar{Z} - \sum_{j=1}^{J_n} \theta_{n,j} e_{n,j}(X_1^i) \right]^2, \tag{5.2}$$

and where the $\lambda_{n,j}$ are (random) shrinkage factors that may depend on $X^1, \ldots, X^n$ but are independent of $\varepsilon^1, \ldots, \varepsilon^n$.

We make the following assumptions.

(A12) The functions $e_{n,j}$ are an orthonormal system of differentiable functions on $L_2([0, 1])$. Moreover,

$$\frac{1}{n} \sum_{j,j'=1}^{J_n} \int e_{n,j}(x)^2 e_{n,j'}(x)^2 \, dx = o(1),$$

$$\sup_{1 \leq j \leq J_n} \int e'_{n,j}(x)^2 \, dx = O_P(D_n),$$

for a sequence D_n of positive numbers.

We assume that the $e_{n,j}$ are orthonormal with respect to Lebesgue measure on [0, 1]. This assumption is made for simplicity and could be achieved for other function systems by orthogonalization.

(A13) The shrinkage factors $\lambda_{n,j}$ satisfy

$$\max_{1 \leqslant j \leqslant J_n} |\lambda_{n,j}| = O_P(1).$$

We compare the estimate $\hat{m}_{\mathrm{O}}$ with $\tilde{m}_{\mathrm{O}} = \sum_{j=1}^{J_n} \lambda_{n,j} \tilde{\theta}_j e_{n,j}$ where $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_{J_n})$ minimizes

$$\frac{1}{n} \sum_{i=1}^{n} \left[\hat{Y}^i - \sum_{j=1}^{J_n} \theta_j e_{n,j}(X_1^i) \right]^2. \tag{5.3}$$

The following theorem states that the differences between these estimates are asymptotically negligible.

Theorem 5. *Suppose that model* (1.1) *holds with* (1.2) *and that* Z^i *is defined by* (1.3)*. Then, under assumptions* (A1)–(A3)*,* (A12) *and* (A13)*,*

$$\|\hat{m}_{\mathrm{O}} - \tilde{m}_{\mathrm{O}}\|_n^2 = O_P(L_n^{-2} + D_n L_n^{-2} J_n n^{-1}), \tag{5.4}$$

$$\int_0^1 [\hat{m}_{\mathrm{O}}(x) - \tilde{m}_{\mathrm{O}}(x)]^2 \, \mathrm{d}x = O_P(L_n^{-2} + D_n L_n^{-2} J_n n^{-1}). \tag{5.5}$$

We will check the assumptions of Theorem 5 for two examples: one with basis functions that localize in the frequency domain, and a second with orthonormal functions that localize in the space and frequency domains. Our first example is a Fourier basis on [0, 1]. Then

$$e_{n,0}(x) = 1, \tag{5.6}$$

$$e_{n,j}(x) = \sqrt{2} \cos(j\pi x), \qquad \text{for } 1 \leqslant j \leqslant J_n. \tag{5.7}$$

Here, J_n diverges to ∞ for $n \to \infty$. It is easy to check that for this basis and if $J_n^2/n = o(1)$, the assumptions of Theorem 5 hold with $D_n = J_n^2$. For r times differentiable functions ($r \geqslant 1$) optimal rates are achieved by choosing J_n of order $n^{1/(2r+1)}$. For such choices we obtain from Theorem 5 that $\|\hat{m}_{\mathrm{O}} - \tilde{m}_{\mathrm{O}}\|_n^2 = O_P(L_n^{-2})$. For $L_n \to \infty$ fast enough this difference is asymptotically negligible because $\|\hat{m}_{\mathrm{O}} - m_1\|_n^2 = O_P(n^{-2r/(2r+1)})$ for r times differentiable functions m_1.

Our second example is a local Fourier basis. We will use orthogonal series estimates with this basis in the next section to show that in additive models, the same L_2 minimax risk can be achieved as in the oracle model (1.3). Wavelets are an example of a basis that localizes in the time and frequency domains. Calculation of empirical wavelet coefficients by minimizing the empirical norm (5.3) is algorithmically feasible only for equidistant designs. For non-equidistant designs, it has been proposed to minimize (5.3) for preconditioned data. These methods are not covered by Theorem 5. Results on rate optimality of wavelet threshold estimates in additive regression models can be found in Zhang and Wong (2003).

The local Fourier basis is defined as follows. Let

$$e_{t0}(x) = \begin{cases} \sqrt{T}, & \text{for } \dfrac{t-1}{T} \leqslant x < \dfrac{t}{T}, \\ 0, & \text{otherwise,} \end{cases} \tag{5.8}$$

$$e_{ts}(x) = \begin{cases} \sqrt{2T}\cos\left(s\pi T\left[x - \dfrac{t-1}{T}\right]\right), & \text{for } \dfrac{t-1}{T} \leqslant x < \dfrac{t}{T}, \\ 0, & \text{otherwise,} \end{cases} \tag{5.9}$$

where $1 \leqslant t \leqslant T$, $1 \leqslant s \leqslant S$ with T and S possibly depending on n. Now the dimension J of the basis is of order $O(TS)$. It is easy to check that for this basis, the assumptions of Theorem 5 hold with $D_n = J^2$ if $(TS)^2/n = o(1)$ as $n \to \infty$. In the next section we will apply Theorem 5 to a local Fourier basis with T of order $n^{1/(2r+1)}$ and (approximately) constant S. Then for r times differentiable functions m_1, the error $\|\hat{m}_{\mathrm{O}} - m_1\|_n^2$ is of order $O_P(n^{-2r/(2r+1)})$. From Theorem 5 we obtain that $\|\hat{m}_{\mathrm{O}} - \tilde{m}_{\mathrm{O}}\|_n^2 = O_P(L_n^{-2})$. For $L_n \to \infty$ fast enough this difference is asymptotically negligible. So again we have that $\hat{m}_{\mathrm{O}}$ and $\tilde{m}_{\mathrm{O}}$ have the same first-order asymptotic performance.

6. Asymptotic minimax estimation in additive models

In Section 5 we showed that when the estimators $\hat{m}_{\mathrm{O}}$ for the oracle model (1.3) and $\tilde{m}_{\mathrm{O}}$ for the additive model (1.1) are based on empirical coefficients of a local Fourier basis, then $\hat{m}_{\mathrm{O}}$ and $\tilde{m}_{\mathrm{O}}$ have the same first-order L_2 risk. In this section we show that in the oracle model, asymptotic minimax risks can be achieved by local Fourier basis estimators. This implies that in model (1.1), $\tilde{m}_{\mathrm{O}}$ achieves the asymptotic minimax risk of $\hat{m}_{\mathrm{O}}$ in model (1.3). Therefore, models (1.1) and (1.3) have the same asymptotic minimax risk. Not knowing $m_2, \ldots, m_D$ in (1.1) leads to no loss of first-order asymptotic efficiency.

For the additive model (1.1) we make the assumption that the additive components lie in Sobolev balls. For $d = 1, \ldots, D$ we assume that

$$m_d \in \mathcal{S}_d = \left\{ g : \mathbb{R} \to \mathbb{R} \middle| \int \left(g^{(r_d)}(x)\right)^2 \mathrm{d}x \leqslant C_d \right\}, \tag{6.1}$$

where $r_d \geqslant 1$ are integers and $C_d > 0$ are positive constants.

Furthermore we assume that for a constant $C > 0$ and a function $\gamma : [0, 1] \to [0, 2]$ with $\lim_{u \to 0} \gamma(u) = 0$ the design density p_1 of X_1^i lies in a class $\mathcal{F}_1$ where $\mathcal{F}_1$ is the set of densities p_1 on $[0, 1]$ that are bounded from above by C and from below by C^{-1} and that fulfil the uniform continuity condition $|p_1(u) - p_1(v)| \leqslant \gamma(|u - v|)$ for $0 \leqslant u, v \leqslant 1$.

We define the asymptotic minimax risk $\rho = \rho(p_1)$ by the condition

$$\lim_{n \to \infty} n^{2r_1/(2r_1+1)} \inf_{\overline{m}_1} \sup_{p_1 \in \mathcal{F}_1} \rho(p_1)^{-1} \sup_{m_1 \in \mathcal{S}_1} \mathrm{E} \int_0^1 (\overline{m}_1(x) - m_1(x))^2 \, \mathrm{d}x = 1,$$

where the infimum runs over all estimates $\overline{m}_1$ based on observations $Z^1, \ldots, Z^n$ in model (1.3). In model (1.3) we have for Gaussian errors ε^i,

$$\rho = \rho(p_1) = \left\{ (2r_1+1)C_1 \left(\frac{\sigma^2 r_1}{\pi(r_1+1)} \int_0^1 p_1^{-1}(x)\mathrm{d}x \right)^{2r_1} \right\}^{1/(2r_1+1)}. \tag{6.2}$$

Equation (6.2) was proved by Efromovich (1996, Theorem 2.1); see also Golubev and Nussbaum (1990), where only non-random 'locally equispaced' designs were considered.

Our next result states that the asymptotic Gaussian minimax risk is achieved by a shrinkage estimator $\tilde{m}_1$ with appropriately chosen shrinkage factors $\lambda_{t,s}$. For simplicity, we only consider conditional L_2 risks, where we condition on the values of the design variables. As above, in the statement of the proposition the conditional expectation given the covariates $X^1, \ldots, X^n$ is denoted by E^*.

Proposition 1. *Suppose that for* $d = 1$ *model* (1.3) *holds with* (1.2). *Then there exist constants c and S and (random) choices of* $\lambda_{t,s}$ *(that depend only on the design variables, on the constants* C_1, r_1 *and on the error variance* σ^2*) with* $0 \leqslant \lambda_{t,s} \leqslant 1$ *such that the following bound holds uniformly over* $p_1 \in \mathcal{F}_1$ *and over* $m_1 \in \mathcal{S}_1$*:*

$$n^{2r_1/(2r_1+1)} \rho(p_1)^{-1} \mathrm{E}^* \int_0^1 (\hat{m}_{\mathrm{O}}(x) - m_1(x))^2 \,\mathrm{d}x \leqslant 1 + o_p(1). \tag{6.3}$$

Here $\rho(p_1)$ *is defined as in* (6.2). *The shrunken orthogonal polynomial estimate* $\hat{m}_{\mathrm{O}}$ *is defined as in* (5.1) *with local Fourier basis* (5.8), (5.9) *with S and with* $T = T_n = \lfloor c^{-1} n^{1/(2r_1+1)} \rfloor$ *where the integer part of a real number x is denoted by* $\lfloor x \rfloor$.

Uniformity in Proposition 1 means that for every $\kappa > 0$,

$$\lim_{n\to\infty} \sup_{p_1 \in \mathcal{F}_1} \sup_{m_1 \in \mathcal{S}_1} P\left(n^{2r_1/(2r_1+1)} \rho(p_1)^{-1} \mathrm{E}^* \int_0^1 (\hat{m}_1(x) - m_1(x))^2 \,\mathrm{d}x > 1 + \kappa \right) = 0.$$

From the previous section and Proposition 1 we immediately obtain that, for $\tilde{m}_{\mathrm{O}}$ with c, S and $\lambda_{t,s}$ as in Proposition 1,

$$n^{2r_1/(2r_1+1)} \rho(p_1)^{-1} \mathrm{E}^* \int_0^1 (\tilde{m}_{\mathrm{O}}(x) - m_1(x))^2 \,\mathrm{d}x \leqslant 1 + o_p(1).$$

This result holds uniformly over $p \in \mathcal{F}$ and over $m_1 \in \mathcal{S}_1^0, \ldots, m_D \in \mathcal{S}_D^0$. Here for a fixed constant C and a fixed function $\gamma : [0, 1] \to [0, 2]$ with $\lim_{u\to 0} \gamma(u) = 0$ the class $\mathcal{F}$ is the set of densities p on $[0, 1]^D$ that have one- and two-dimensional marginals p_d, $p_{d,d'}$ $(1 \leqslant d, d' \leqslant D)$ bounded from above by C and from below by C^{-1} and that, for the first argument, have a marginal p_1 with $|p_1(u) - p_1(v)| \leqslant \gamma(|u - v|)$ for $0 \leqslant u, v \leqslant 1$. Furthermore, $\mathcal{S}_d^0 = \mathcal{S}_d \cap \{m_d : \mathrm{E} m_d(\mathrm{X}_d^1) = 0\}$.

We reformulate this result as the following theorem.

Theorem 6. *Under the conditions of Proposition 1 there exists an estimator* $\tilde{m}_{\mathrm{O}}$ *in the additive model* (1.1) *with*

$$n^{2r_1/(2r_1+1)}\rho(p_1)^{-1}\mathrm{E}^*\int_0^1(\tilde{m}_{\mathrm{O}}(x)-m_1(x))^2\,\mathrm{d}x=1+o_p(1)$$

uniformly over $p\in\mathcal{F}$ *and over* $m_1\in\mathcal{S}_1^0,\ldots,m_D\in\mathcal{S}_D^0$.

For Gaussian errors, Theorem 6 shows that in an additive model there exists an estimator that achieves the same asymptotic minimax risk as for a classical regression model with one regression component. This means that asymptotically no information is lost by the introduction of additive components. But Theorem 6 and Proposition 1 are only of interest for Gaussian errors. For non-Gaussian errors better asymptotic minimax constants can be achieved by using local likelihood methods; see Golubev and Nussbaum (1990). We conjecture that at the cost of some rather technical considerations these results can be generalized to additive models with non-Gaussian errors. Another possible generalization concerns data-adaptive choices of the weights $\lambda_{t,s}$. The weights depend on the constants C_1, r_1 and on the error variances σ^2 that are typically unknown. Data-adaptive estimates of $\int(m_1^{(r_1)}(x))^2\,\mathrm{d}x$ and σ^2 can be easily achieved as discussed for kernel smoothing in Section 3. These estimates can be plugged into the definition of the weights $\lambda_{t,s}$. An extension of Theorem 6 to estimates $\tilde{m}_{\mathrm{O}}$ with these data-adaptive choices of $\lambda_{t,s}$ would require bounds on $\tilde{m}_{\mathrm{O}}-\hat{m}_{\mathrm{O}}$ that hold uniformly over possible choices of $\lambda_{t,s}$. Compare also with Section 3, where bounds on $\tilde{m}_{\mathrm{LP},h}-\hat{m}_{\mathrm{LP},h}$ are stated that hold uniformly for possible choices of the bandwidths h. Extensions of our results to data-adaptive choices of r_1 could be done along the same lines as the construction in Efromovich and Pinsker (1984) and Golubev (1987). But the mathematics would be technically rather involved.

7. Simulations

We generated two-dimensional regression data (Y^i, X_1^i, X_2^i), $i=1,\ldots,n$, where $Y^i=m_1(X_1^i)+m_2(X_2^i)+\varepsilon^i$, and ε^i are independent, identically distributed standard Gaussian. The design variables (X_1^i, X_2^i) are Gaussian, truncated to $[-1, 1]$, with standard deviations $\sigma_1=\sigma_2=1$ and correlation $\rho=0.8$. The first component $m_1(x_1)=x_1^2$, $x_1\in[-1, 1]$, of the regression function is shown as a thick solid line in Figures 1 and 2. The second component of the regression function is $m_2(x_2)=x_2^3$, $x_2\in[-1, 1]$.

We considered the kernel regression estimation with the quartic kernel (Bartlett kernel). Three estimators were studied:

1. Estimator $\tilde{m}_1$ is the kernel estimator applied to the data $\hat{Y}^i$ which was obtained by backfitting the presmoothed data. We draw the estimator $\tilde{m}_1$ with a dashed line.
2. Estimator $\hat{m}_1$ is the kernel estimator based on smoothing of $Z^i-\overline{Z}$, where $Z^i=m_1(X_1^i)+\varepsilon^i$ are the oracle data. We draw the estimator $\hat{m}_1$ with a thin solid line.
3. We study also the kernel estimator $\overline{m}_1$, which was constructed from the presmoothed oracle data $\hat{Z}^i$. We draw this estimator with a dotted line.

Figure 1 shows the results for 10 000 samples of size $n=250$. The smoothing parameter was

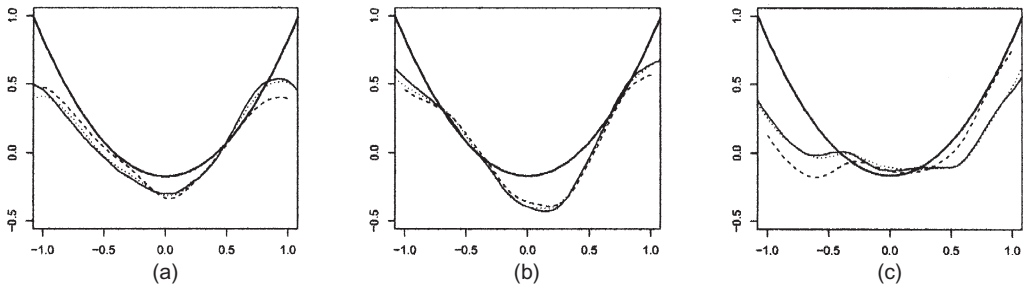

Figure 1. Results for sample size $n = 250$; three estimates of m_1 from three Monte Carlo samples: (a) 0.25 quantile; (b) 0.5 quantile; (c) 0.75 quantile. The dashed line shows $\tilde{m}_1$, the thin solid line shows $\hat{m}_1$, the dotted line shows $\overline{m}_1$, and the thick solid line shows m_1.

$h = 0.42$ and the number of bins used in presmoothing was $L_n = 12$. The bandwidth $h = 0.42$ minimizes the mean integrated squared error of the oracle estimate $\hat{m}$. Its value was determined by simulations. The number of backfitting steps in the presmoothing of the calculation of $\tilde{m}_1$ was 7. We show estimates constructed from three Monte Carlo samples. The samples were chosen so that they correspond to the sample quantiles of the normalized difference of the integrated squared error

$$ndISE = \left[\int_0^1 (m_1 - \tilde{m}_1)^2 - \int_0^1 (m_1 - \hat{m}_1)^2\right] \Big/ \int_0^1 (m_1 - \hat{m}_1)^2.$$

The 0.25, 0.5 and 0.75 quantiles are shown.

Figure 2 shows the results for 10 000 samples of size $n = 500$. The smoothing parameter was $h = 0.34$ and the number of bins in presmoothing was $L_n = 16$. Again the 0.25, 0.5 and 0.75 quantiles are shown. The bandwidth $h = 0.34$ minimizes the mean integrated squared error of the oracle estimate $\hat{m}$ and was determined by simulations.

The simulations show that the two-step procedure works quite well. We conjecture that the finite-sample performance of the two-step procedure could be improved by using some

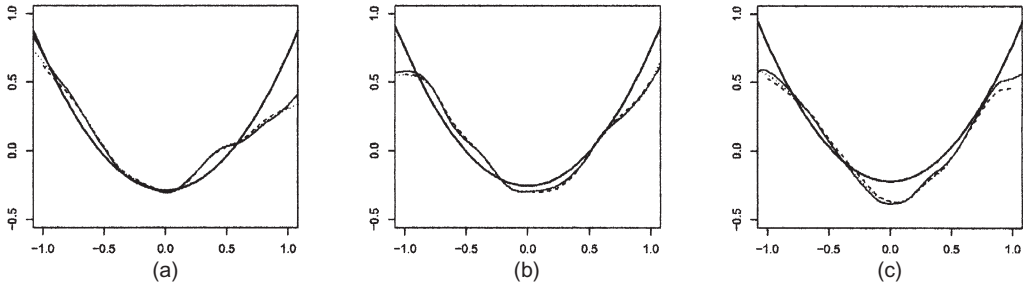

Figure 2. Results for sample size $n = 500$; three estimates of m_1 from three Monte Carlo samples: (a) 0.25 quantile; (b) 0.5 quantile; (c) 0.75 quantile. The dashed line shows $\tilde{m}_1$, the thin solid line shows $\hat{m}_1$, the dotted line shows $\overline{m}_1$, and the thick solid line shows m_1.

more refined methods. In particular, more advanced smoothing methods could be used in the presmoothing instead of our crude piecewise constant estimate. Our presmoothing method was chosen mainly to allow a rather general and simple asymptotic theory. Furthermore, a smaller bandwidth should be used in the second step to account for the smoothing of the first step. We conjecture that then the two step estimate will be closer to the oracle estimate. Note that the bandwidth was optimized for the oracle estimate.

Figure 3 shows the effect of the number of bins used in the presmoothing. We draw the simulated means of *ndISE* as function of the number of bins. The open circles show the results for sample size $n = 250$ and the filled circles show the results for sample size $n = 500$. The smoothing parameters were, as before, $h = 0.42$ for $n = 250$ and $h = 0.34$ for $n = 500$. We generated 10 000 samples. The performance of the two-step estimate depends on the choice of L_n. For both sample sizes there is a clear minimum. For $n = 250$ the best number of bins is about $L_n = 12$, and for $n = 500$ the best number is about $L_n = 16$. It is an open problem how L_n can be chosen depending on the data. Theoretically this is a very technical problem because the optimal choice depends on second-order properties of the two-step procedure. Practically it is complicated by the fact that the optimal L_n depends on the bandwidth h chosen in the second step (and the optimal h depends on L_n).

8. Proofs

8.1. Proof of Theorem 1

We start by decomposing Y^i and Z^i into signal and error components:

$$Y^i = \overline{Y} + T_1^i + T_2^i + \varepsilon_i - \bar{\varepsilon}, \tag{8.1}$$

$$Z^i = \overline{Z} + U_1^i + U_2^i + \varepsilon_i - \bar{\varepsilon}, \tag{8.2}$$

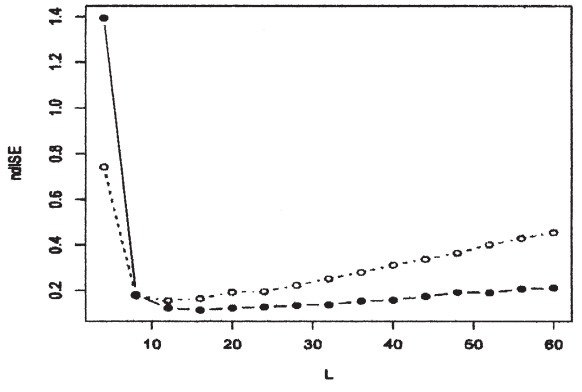

Figure 3. The *ndISE* as a function of the number of bins L_n.

where

$$T_1^i = m_1(X_1^i) - \overline{m}_1 - \hat{\mu}_{1,i} + \ldots + m_D(X_D^i) - \overline{m}_D - \hat{\mu}_{D,i},$$

$$U_1^i = m_1(X_1^i) - \overline{m}_1 - \hat{\mu}_{1,i},$$

$$\overline{m}_d = \frac{1}{n}\sum_{i=1}^n m_d(X_d^i),$$

$$\hat{\mu}_{d,i} = \frac{\sum_{r=1}^n \left[m_d(X_d^r) - \overline{m}_d\right] I_l(X_d^r)}{\sum_{r=1}^n I_l(X_d^r)}, \qquad \text{for } X_d^i \in I_l,$$

$$T_2^i = \hat{\mu}_{1,i} + \ldots + \hat{\mu}_{D,i},$$

$$U_2^i = \hat{\mu}_{1,i},$$

$$\overline{\varepsilon} = \frac{1}{n}\sum_{i=1}^n \varepsilon^i.$$

Note that for $d = 1, \ldots, D$, the function $\hat{\mu}_{d,i}$ (as a function of X_d^i) is a piecewise constant fit of $m_d(X_d^i) - \overline{m}_d$ (as $\hat{Z}^i$ is of $Z_i - \overline{Z}$).

For a vector $V = (V^i)_{i=1}^n$ we introduce linear transformations S and R. The ith element of $S(V)$ is defined by

$$S(V)^i = \frac{\sum_{r=1}^n (V^r - \overline{V}) I_l(X_1^r)}{\sum_{r=1}^n I_l(X_1^r)}, \qquad \text{for } X_1^i \in I_l.$$

Here $\overline{V}$ is defined as $n^{-1}\sum_{i=1}^n V^i$. Note that $\hat{Z} = S(Z)$.

The ith element of $R(V)$ is defined by

$$R(V)^i = \sum_{l=1}^{L_n} \tau_{1,l} I_1(X_1^i),$$

where $\tau_{1,1}, \ldots, \tau_{D,L_n}$ minimizes

$$\sum_{i=1}^n \left[V^i - \overline{V} - \sum_{d=1}^D \sum_{l=1}^{L_n} \tau_{d,l} I_d(X_d^i) \right]^2.$$

Note that $\hat{Y} = R(Y)$.

Because R and S are linear operators we have that

$$\hat{Y} = R(T_1) + R(T_2) + R(\varepsilon), \tag{8.3}$$

$$\hat{Z} = S(U_1) + S(U_2) + S(\varepsilon). \tag{8.4}$$

For part (i) of the theorem we have to show that

$$\mathrm{E}^* \|\hat{Y} - \hat{Z}\|^2 = O_P(L_n^{-2}), \tag{8.5}$$

where $\|V\|^2 = n^{-1}\sum_{i=1}^n (V^i)^2$. We will show that

$$\mathrm{E}^*\|R(T_1)\|^2 = O_P(L_n^{-2}), \tag{8.6}$$

$$\mathrm{E}^*\|S(U_1)\|^2 = O_P(L_n^{-2}), \tag{8.7}$$

$$R(T_2) = S(U_2), \tag{8.8}$$

$$\mathrm{E}^*\|R(\varepsilon) - S(\varepsilon)\|^2 = O_P(L_n^{-2}). \tag{8.9}$$

These claims immediately imply (8.5) and therefore statement (i) of Theorem 1. Claim (8.8) directly follows from the definition of the operators S and R and the structure of U_2 and T_2. Note that $R(T_2)^i = \hat{\mu}_{1,i} = U_2^i = S(U_2)^i$. For the proof of (8.7) note that $\mathrm{E}^*\|U_1\|^2 = O_P(L_n^{-2})$. Because S is a projection, $\|S(U_1)\| \leqslant \|U_1\|$. This implies (8.7). For the proof of statement (i) of Theorem 1 it remains to show (8.6) and (8.9). For the proof of (8.9) we apply Theorems 1 and 2 in Mammen *et al.* (1999). We will show that

$$\sup_{1\leqslant l\leqslant L_n}\left|L_n n^{-1}\sum_{i=1}^n I_l(X_d^i) - f_d(l/L_n)\right| = o_P(1), \qquad \text{for } 1 \leqslant d \leqslant D, \tag{8.10}$$

$$\sup_{1\leqslant l,r\leqslant L_n}\left|L_n^2 n^{-1}\sum_{i=1}^n I_l(X_d^i)I_r(X_{d'}^i) - f_{d,d'}(l/L_n,\, r/L_n)\right| = o_P(1), \qquad \text{for } 1 \leqslant d < d' \leqslant D, \tag{8.11}$$

$$\mathrm{E}^*\|W_{d,d'}\|^2 = O_P(n^{-1}), \qquad \text{for } d \neq d', \tag{8.12}$$

where $W_{d,d'}$ is defined by

$$W_{d,d'}^i = n^{-1}\sum_{j=1}^n \varepsilon_j a_{r,d,d'}(X_d^j), \qquad \text{for } X_{d'}^i \in I_r, \tag{8.13}$$

$$a_{r,d,d'}(x) = n\sum_{l=1}^{L_n}\sum_{k=1}^n \frac{I_l(x)I_l(X_d^k)I_r(X_{d'}^k)}{\sum_{t=1}^n I_l(X_d^t)\sum_{t=1}^n I_r(X_{d'}^t)}.$$

Claims (8.10) and (8.11) imply that $R(\varepsilon)$ can be calculated by backfitting. The backfitting algorithm converges with exponential rate. Compare (8.10) and (8.11) with conditions A1–A2 in Mammen *et al.* (1999) and apply their Theorem 1. Terms of type $W_{d,d'}$ appear in the first step of the backfitting algorithm for the calculation of $R(\varepsilon)$ if one starts the algorithm with $S(\varepsilon)$ as starting value. If these terms are of higher order then the difference $\|R(\varepsilon) - S(\varepsilon)\|$ is of the same higher order; see the proof of (93) in the proof of Lemma 4 in Mammen *et al.* (1999). This can be used to show that (8.10)–(8.12) imply (8.9). Claims (8.10)–(8.12) also imply that the backfitting operator has operator norm strictly less than 1; see Lemma 2 in Mammen *et al.* (1999). For this reason claim (8.6) follows from $\mathrm{E}^*\|T_1\|^2 = O_P(L_n^{-2})$. So for the first part of Theorem 1 it remains to check (8.10)–(8.12).

Claims (8.10)–(8.11) follow by application of Bernstein's inequality; see Shorack and Wellner (1986). For a proof of (8.12) one uses

$$\sup_{r,d,d',x} |a_{r,d,d'}(x)| = O_P(1). \tag{8.14}$$

We now come to the proof of part (ii) of Theorem 1. We again use decomposition (8.3)–(8.4). Claim (2.4) follows from (8.8) and

$$\|R(T_1)\|_\infty = O_P(L_n^{-1}), \tag{8.15}$$

$$\|S(U_1)\|_\infty = O_P(L_n^{-1}), \tag{8.16}$$

$$\|R(\varepsilon) - S(\varepsilon)\|_\infty = O_P\left(\sqrt{\frac{\log n}{n}}\right), \tag{8.17}$$

where $\|V\|_\infty = \max_{1\leqslant i\leqslant n}|V^i|$ denotes the supremum norm. Claim (8.16) directly follows from $\|U_1\|_\infty = O_P(L_n^{-1})$. Claim (8.15) can be shown by using $\|T_1\|_\infty = O_P(L_n^{-1})$, $\|T_1\|^2 = O_P(L_n^{-2})$ and the fact that the backfitting operator has norm strictly less than 1 (see above) and maps functions with bounded L_2 norm into functions with bounded L_∞ norm; see equation (86) in Mammen *et al.* (1999). For a proof of (8.17) one checks first that

$$\|W_{d,d'}\|_\infty = O_P\left(\sqrt{\frac{\log n}{n}}\right). \tag{8.18}$$

Then with (8.10)–(8.11) and our assumptions, claim (8.17) follows by Theorem 2 in Mammen *et al.* (1999). It remains to check (8.18). This can be done by replacing ε^i in (8.13) by variables that are absolutely bounded by $n^{1/2-\gamma}$ for a $\gamma > 0$ small enough and by using (8.14). Then claim (8.18) follows by application of the Bernstein inequality.

8.2. Proof of Theorem 2

For the proof of (3.4), note first that

$$\int [\hat{m}_K(x) - \tilde{m}_K(x)]^2\,\mathrm{d}x \leqslant 2\int\left[\frac{1}{n}\sum_{i=1}^n w_i(x)\{Z^i - \overline{Z} - \hat{Z}^i\}\right]^2 \mathrm{d}x$$

$$+ 2\int\left[\frac{1}{n}\sum_{i=1}^n w_i(x)\{\hat{Z}^i - \hat{Y}^i\}\right]^2 \mathrm{d}x.$$

The second term on the right-hand side can be bounded by

$$2\sup_{0\leqslant x\leqslant 1}\frac{1}{n}\sum_{i=1}^n |w_i(x)| \int \frac{1}{n}\sum_{i=1}^n |w_i(x)|\{\hat{Z}^i - \hat{Y}^i\}^2\,\mathrm{d}x.$$

This is of order $O_P(L_n^{-2})$ by condition (A5) and Theorem 1. It remains to bound the first term on right-hand side of (8.19). By definition of $\hat{Z}^i$, we obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} w_i(x)\{Z^i - \overline{Z} - \hat{Z}^i\} \qquad (8.19)\\
&= \frac{1}{n}\sum_{i=1}^{n} [w_i(x) - \overline{w}_i(x)]\{Z^i - \overline{Z} - \hat{Z}^i\}\\
&= \frac{1}{n}\sum_{i=1}^{n} [w_i(x) - \overline{w}_i(x)]Z^i\\
&= \frac{1}{n}\sum_{i=1}^{n} [w_i(x) - \overline{w}_i(x)]m_1(X_1^i) + \frac{1}{n}\sum_{i=1}^{n} [w_i(x) - \overline{w}_i(x)]\varepsilon_i,
\end{aligned}$$

where

$$\overline{w}_i(x) = \frac{\sum_{r=1}^{n} w_r(x) I_l(X_1^r)}{\sum_{r=1}^{n} I_l(X_1^r)}$$

for $X_1^i \in I_l$. Thus for (3.4) it suffices to show that

$$\int \left[\frac{1}{n}\sum_{i=1}^{n} [w_i(x) - \overline{w}_i(x)]m_1(X_1^i)\right]^2 \mathrm{d}x = O_P(L_n^{-2}), \qquad (8.20)$$

$$\int \left[\frac{1}{n}\sum_{i=1}^{n} [w_i(x) - \overline{w}_i(x)]\varepsilon^i\right]^2 \mathrm{d}x = O_P(L_n^{-2}\rho_n^2 n^{-1}). \qquad (8.21)$$

Claim (8.21) immediately follows from (A6). For the proof of claim (8.20) we make use of (A3).

For the proof of claim (3.5) we note first that (see (8.19))

$$\begin{aligned}
\hat{m}_K(x) - \tilde{m}_K(x) = &\frac{1}{n}\sum_{i=1}^{n} w_i(x)\{\hat{Z}^i - \hat{Y}^i\}\\
&+ \frac{1}{n}\sum_{i=1}^{n}[w_i(x) - \overline{w}_i(x)]m_1(X_1^i) + \frac{1}{n}\sum_{i=1}^{n} [w_i(x) - \overline{w}_i(x)]\varepsilon_i.
\end{aligned}$$

The supremum norm on the terms of the right-hand side can be easily bounded by using the results of Theorem 1 or condition (A4) for the first or second term, respectively. For the third term one uses condition (A7). This term can be bounded by first replacing ε^i by variables that are absolutely bounded by $n^{\gamma'}$ for a $\gamma' < \gamma$. Then by application of the Bernstein inequality one can show that the term is of order $O_P(L_n^{-1}\rho_n n^{-1/2}\sqrt{\log n})$.

8.3. Proof of Theorem 3

Note that

$$h^j\left(\hat{m}^j_{\mathrm{LP},q,h}(x)\right)_{j=0,\ldots,q} = \hat{P}_h(x)^{-1}\left(\hat{r}^j_h(x)\right)_{j=0,\ldots,q},$$

$$h^j\left(\tilde{m}^j_{\mathrm{LP},q,h}(x)\right)_{j=0,\ldots,q} = \hat{P}_h(x)^{-1}\left(\tilde{r}^j_h(x)\right)_{j=0,\ldots,q},$$

where

$$\hat{r}^j_h(x) = n^{-1}\sum_{i=1}^n (X_1^i - x)^j h^{-j} K_h(X_1^i - x)(Z^i - \overline{Z}^i),$$

$$\tilde{r}^j_h(x) = n^{-1}\sum_{i=1}^n (X_1^i - x)^j h^{-j} K_h(X_1^i - x)\hat{Y}^i,$$

and where $\hat{P}_h(x)$ is a $(q+1)\times(q+1)$ matrix with elements

$$\hat{P}_h(x)_{j,k} = n^{-1}\sum_{i=1}^n (X_1^i - x)^{j+k} h^{-j-k} K_h(X_1^i - x)$$

for $0 \leqslant j,\ k \leqslant q$. Using standard smoothing theory, one can show that

$$\sup_{0\leqslant x\leqslant 1, c_1 n^{-\alpha_1}\leqslant h\leqslant c_2 n^{-\alpha_2}} |\hat{P}_h(x) - \mathrm{E}[\hat{P}_h(x)]| = o_P(1),$$

where $|(a_{ij})| = \sup_{i,j}|a_{ij}|$ for a matrix (a_{ij}). The matrices $\mathrm{E}[\hat{P}_h(x)]$ have eigenvalues that are uniformly bounded from below and from above. Thus,

$$\sup_{0\leqslant x\leqslant 1, c_1 n^{-\alpha_1}\leqslant h\leqslant c_2 n^{-\alpha_2}} |\hat{P}_h(x)^{-1} - \mathrm{E}[\hat{P}_h(x)]^{-1}| = o_P(1).$$

Therefore it suffices for (3.6) and (3.7) that

$$\int_0^1 \left[\hat{r}^j_h(x) - \tilde{r}^j_h(x)\right]^2 \mathrm{d}x = O_P(L_n^{-2}) \tag{8.22}$$

or

$$\sup_{0\leqslant x\leqslant 1, c_1 n^{-\alpha_1}\leqslant h\leqslant c_2 n^{-\alpha_2}} |\hat{r}^j_h(x) - \tilde{r}^j_h(x)| = O_P(L_n^{-1}), \tag{8.23}$$

respectively. We now verify the conditions of Theorem 2. By standard smoothing theory one can check that under condition (A8) for a fixed sequence of bandwidths h with $c_1 n^{-\alpha_1} \leqslant h \leqslant c_2 n^{-\alpha_2}$ the weights $w_i(x) = (X_1^i - x)^j h^{-j} K_h(X_1^i - x)$ satisfy (A5) and (A6) with $\rho_n = n^{3\alpha_1/2}$. Because of $\rho_n^2 n^{-1} \to 0$, by (A9), application of Theorem 2(i) gives (8.22).

For the proof of (8.23) note first that, for sequences of bandwidths h with $c_1 n^{-\alpha_1} \leqslant h \leqslant c_2 n^{-\alpha_2}$, (A8) and (A10) imply (A7), again with $\rho_n = n^{3\alpha_1/2}$. Therefore we obtain from Theorem 2(ii) for a fixed sequence of bandwidths that

$$\sup_{0\leqslant x\leqslant 1} |\hat{r}_h^j(x) - \tilde{r}_h^j(x)| = O_P(L_n^{-1}).$$

By using a generalization of Theorem 2 which states uniformity over classes of weight functions one shows (8.23). This generalization can be proved by a slight change of the arguments in the proof of Theorem 2.

8.4. Proof of Theorem 4

For $M_1^i = m_1(X_1^i)$ define

$$\hat{M}_1^i = \frac{\sum_{r=1}^n (M_1^r - \overline{M}_1) I_l(X_1^r)}{\sum_{r=1}^n I_l(X_1^r)}$$

for $X_1^i \in I_l$, where $\overline{M}_1$ denotes the overall average $n^{-1}\sum_{i=1}^n M_1^i$. Similarly, define $\hat{\varepsilon}$. With this notation we can write $\hat{Z} = \hat{M}_1 + \hat{\varepsilon}$.

Put

$$\hat{m}_U = \arg\min_m V(U, m),$$

where

$$V(U, m) = \|U - m\|_n^2 + \lambda_n^2 J_k^2(m).$$

Then

$$\hat{m}_S = \hat{m}_{Z-\overline{Z}},$$

$$\tilde{m}_S = \hat{m}_{\hat{Y}}.$$

Now because of $\|\hat{Y} - \hat{Z}\|_n^2 = O_P(L_n^{-2})$ – see Theorem 1 – we obtain

$$\|\tilde{m}_S - \hat{m}_{\hat{Z}}\|_n^2 = \|\hat{m}_{\hat{Y}} - \hat{m}_{\hat{Z}}\|_n^2 = O_P(L_n^{-2}).$$

So it remains to compare $\hat{m}_{\hat{Z}} = \hat{m}_{\hat{M}_1} + \hat{m}_{\hat{\varepsilon}}$ with $\hat{m}_S = \hat{m}_{Z-\overline{Z}} = \hat{m}_{M_1-\overline{M}_1} + \hat{m}_{\varepsilon-\bar{\varepsilon}}$. It follows from (A3) that $\|M_1 - \overline{M}_1 - \hat{M}_1\|^2 = O_P(L_n^{-2})$. So we have

$$\|\hat{m}_{M_1-\overline{M}_1} - \hat{m}_{\hat{M}_1}\|_n^2 = O_P(L_n^{-2}).$$

So for (4.3) it suffices to show that

$$\|\hat{m}_{\hat{\varepsilon}} - \hat{m}_{\varepsilon-\bar{\varepsilon}}\|_n^2 = O_P(n^{-1/2} L_n^{-(2k-1)/(2k)}). \tag{8.24}$$

For the proof of (8.24) we show first that, for all $c > 0$,

$$\sup_{m\in\mathcal{M}_c} \left| V(\varepsilon - \bar{\varepsilon}, m) - V(\hat{\varepsilon}, m) - \Delta \right| = O_P(n^{-1/2} L_n^{-(2k-1)/(2k)}), \tag{8.25}$$

where $\Delta = \|\varepsilon - \bar{\varepsilon}\|_n^2 - \|\hat{\varepsilon}\|_n^2$ and $\mathcal{M}_c$ is the class of all functions m with $J_k^2(m) \leqslant c$ and $\|m\|_n^2 \leqslant c$.

For the proof we use the following notation. For a function m the vector M_m is defined by $M_m^i = m(X_1^i)$ – for example, then $M_{m_1}^i = M_1^i$. The vector $\hat{M}_m$ is defined as $\hat{M}_1$ but

with (the true first component) m_1 replaced by (a varying function) m. Furthermore $\langle \cdot, \cdot \rangle_n$ is the empirical scalar product $\langle A, B \rangle_n = n^{-1} \sum_{i=1}^n A_i B_i$. With this notation we obtain

$$\begin{aligned} V(\varepsilon - \bar{\varepsilon}, m) - V(\hat{\varepsilon}, m) - \Delta &= 2\langle \hat{\varepsilon} - \varepsilon + \bar{\varepsilon}, M_m \rangle_n \\ &= 2\langle \hat{\varepsilon} - \varepsilon + \bar{\varepsilon}, M_m - \hat{M}_m \rangle_n \\ &= 2\langle \varepsilon, M_m - \hat{M}_m \rangle_n. \end{aligned}$$

We now use the fact that the same entropy bound as for M_m applies for $M_m - \hat{M}_m$. Arguing as in Section 10.1 in van de Geer (2000), we obtain

$$\sup_{m \in \mathcal{M}_c} \frac{|\langle \varepsilon, M_m - \hat{M}_m \rangle_n|}{\|M_m - \hat{M}_m\|_n^{(2k-1)/(2k)}} = O_P(n^{-1/2}).$$

Claim (8.25) now follows from $\|M_m - \hat{M}_m\|_n = O_P(L_n^{-1})$. This holds because $\sup_{m \in \mathcal{M}_c} \int m'(u)^2 \, \mathrm{d}u < \infty$.

Equation (8.25) and the definition of $\hat{m}_S$ and $\tilde{m}_S$ imply

$$\begin{aligned} & V(\varepsilon - \bar{\varepsilon}, \hat{m}_{\hat{\varepsilon}}) - V(\varepsilon - \bar{\varepsilon}, \hat{m}_{\varepsilon - \bar{\varepsilon}}) \qquad (8.26) \\ &\leqslant V(\hat{\varepsilon}, \hat{m}_{\hat{\varepsilon}}) - V(\hat{\varepsilon}, \hat{m}_{\varepsilon - \bar{\varepsilon}}) + O_P(n^{-1/2} L_n^{-(2k-1)/(2k)}) \\ &= O_P(n^{-1/2} L_n^{-(2k-1)/(2k)}). \end{aligned}$$

Claim (4.3) now follows from

$$V(\varepsilon - \bar{\varepsilon}, \hat{m}_{\hat{\varepsilon}}) - V(\varepsilon - \bar{\varepsilon}, \hat{m}_{\varepsilon - \bar{\varepsilon}}) = \|\hat{m}_{\hat{\varepsilon}} - \hat{m}_{\varepsilon - \bar{\varepsilon}}\|_n^2 + \lambda_n^2 J_k^2(\hat{m}_{\hat{\varepsilon}} - \hat{m}_{\varepsilon - \bar{\varepsilon}}). \qquad (8.27)$$

Equation (8.27) follows from

$$\langle \varepsilon - \bar{\varepsilon} - \hat{m}_{\varepsilon - \bar{\varepsilon}}, \hat{m}_{\hat{\varepsilon}} - \hat{m}_{\varepsilon - \bar{\varepsilon}} \rangle_n - \lambda_n^2 \int \hat{m}_{\varepsilon - \bar{\varepsilon}}^{(k)}(u) \left[\hat{m}_{\hat{\varepsilon}}^{(k)}(u) - \hat{m}_{\varepsilon - \bar{\varepsilon}}^{(k)}(u) \right] \mathrm{d}u = 0. \qquad (8.28)$$

Claim (8.28) immediately follows from the fact that $m_{\varepsilon - \bar{\varepsilon}}$ minimizes $V(\varepsilon - \bar{\varepsilon}, m)$. This shows claim (4.3). Claim (4.5) follows by similar arguments and use of (8.27). For the proof of claim (4.4) note first that by application of Lemma 5.16 in van de Geer (2000) one obtains from (4.3) and (4.5) that

$$\int_0^1 [\hat{m}_S(x) - \tilde{m}_S(x)]^2 p_1(x) \mathrm{d}x = O_P(n^{-1/2} L_n^{-(2k-1)/(2k)} + L_n^{-2}).$$

This implies (4.4) because the density p_1 is bounded according to assumption (A2).

8.5. Proof of Theorem 5

We rewrite the functional (5.2) as

$$\|Z - \bar{Z} - B\theta\|_n^2,$$

where the matrix B has elements $B_{i,j} = e_{n,j}(X_1^i)$. Then

$$\hat{\theta} = \left[\frac{1}{n} B^{\mathrm{T}} B\right]^{-1} \frac{1}{n} B^{\mathrm{T}} \left[Z - \overline{Z}\right]$$

and

$$\hat{m}_{\mathrm{O}} = B\Lambda \left[\frac{1}{n} B^{\mathrm{T}} B\right]^{-1} \frac{1}{n} B^{\mathrm{T}} \left[Z - \overline{Z}\right], \tag{8.29}$$

where Λ is a diagonal matrix with diagonal elements $\lambda_{n,j}$ and where, in an abuse of notation, for a function m the vector with elements $m(X_1^i)$ is also denoted by m. For simplicity of notation here we have sometimes omitted the index n.

We start by showing that $n^{-1}B^{\mathrm{T}}B$ has eigenvalues stochastically bounded from above and away from zero. For the proof denote the matrix with elements

$$C_{j,j'} = \mathrm{E}\left[e_{n,j}(X_1^i) e_{n,j'}(X_1^i)\right]$$

by C. Then with constants $c_1, c_2 > 0$,

$$\sup_{a:\|a\|=1} \left| a^{\mathrm{T}} \frac{1}{n} B^{\mathrm{T}} B a - a^{\mathrm{T}} C a \right| = o_P(1), \tag{8.30}$$

$$c_1 \leqslant \inf_{a:\|a\|=1} a^{\mathrm{T}} C a \leqslant \sup_{a:\|a\|=1} a^{\mathrm{T}} C a \leqslant c_2, \tag{8.31}$$

where for a vector $a = (a_1, \ldots, a_J)$ we write $\|a\|^2 = \sum_{j=1}^J a_j^2$. Claim (8.31) immediately follows from $C_{j,j'} = \int e_{n,j}(x) e_{n,j'}(x) p_1(x) \mathrm{d}x$ and the facts that $e_{n,j}$ are orthonormal functions and that p_1 is bounded from above and below; see (A2). For the proof of (8.30) note that

$$\begin{aligned}
&\mathrm{E}\left[\sup_{a:\|a\|=1} \left| a^{\mathrm{T}} \frac{1}{n} B^{\mathrm{T}} B a - a^{\mathrm{T}} C a \right|\right] \\
&\quad \leqslant \mathrm{E} \sum_{j,j'} \left[\frac{1}{n} \sum_{i=1}^n e_{n,j}(X_1^i) e_{n,j'}(X_1^i) - \mathrm{E}\left[e_{n,j}(X_1^i) e_{n,j'}(X_1^i)\right]\right]^2 \\
&\quad \leqslant \sum_{j,j'} \mathrm{E}\left[\frac{1}{n^2} \sum_{i=1}^n e_{n,j}^2(X_1^i) e_{n,j'}^2(X_1^i)\right] \\
&\quad \leqslant \sum_{j,j'} \frac{1}{n} c_3 \int e_{n,j}^2(x) e_{n,j'}^2(x) \mathrm{d}x,
\end{aligned}$$

for a constant c_3. Thus (8.30) follows from (A12).

Now because $n^{-1}B^{\mathrm{T}}B$ has stochastically bounded eigenvalues it suffices to show that

$$\|\hat{\theta} - \tilde{\theta}\|^2 = O_P(L_n^{-2} + D_n L_n^{-2} J n^{-1}).$$

This claim follows from

$$\left\| \frac{1}{n} B^{\mathrm{T}} \left[Z - \overline{Z} - \hat{Y} \right] \right\|^2 = O_P(L_n^{-2} + D_n L_n^{-2} J n^{-1});$$

see the definitions of $\hat{\theta}$ and $\tilde{\theta}$. Because $\|\hat{Y} - \hat{Z}\|^2 = O_P(L_n^{-2})$ it remains to show that

$$\left\| \frac{1}{n} B^{\mathrm{T}} \left[Z - \overline{Z} - \hat{Z} \right] \right\|^2 = O_P(L_n^{-2} + D_n L_n^{-2} J n^{-1}). \tag{8.32}$$

We now put

$$\hat{B}_{i,j} = \frac{\sum_{r=1}^n (B_{r,j} - \overline{B}_j) I_l(X_1^r)}{\sum_{r=1}^n I_l(X_1^r)},$$

$$\hat{M}_i = \frac{\sum_{r=1}^n (M_r - \overline{M}) I_l(X_1^r)}{\sum_{r=1}^n I_l(X_1^r)},$$

for $X_1^i \in I_l$ with $M_i = m_1(X_1^i)$. Here $\overline{M} = n^{-1} \sum_{i=1}^n M_i$ and $\overline{B}_j = n^{-1} \sum_{i=1}^n B_{i,j}$. With this notation we obtain

$$\begin{aligned} \frac{1}{n} B^{\mathrm{T}} \left[Z - \overline{Z} - \hat{Z} \right] &= \frac{1}{n} (B^{\mathrm{T}} - \hat{B}^{\mathrm{T}}) \left[Z - \overline{Z} - \hat{Z} \right] \\ &= \frac{1}{n} (B^{\mathrm{T}} - \hat{B}^{\mathrm{T}}) \varepsilon + \frac{1}{n} B^{\mathrm{T}} \left[M - \overline{M} - \hat{M} \right]. \end{aligned}$$

Because $\int m_1'(x)^2 \, \mathrm{d}x$ is bounded the second term on the right-hand side has norm bounded by $O_P(L_n^{-1})$. It remains to bound the norm of the first term. Note that with a constant c_4,

$$\begin{aligned} \mathrm{E}^* \left\| \frac{1}{n} (B^{\mathrm{T}} - \hat{B}^{\mathrm{T}}) \varepsilon \right\|^2 &= \mathrm{E}^* \sum_{j=1}^J \left[\frac{1}{n} \sum_{i=1}^n (B_{i,j} - \hat{B}_{i,j}) \varepsilon^i \right]^2 \\ &\leqslant c_4 \frac{1}{n^2} \mathrm{E}^* \sum_{j=1}^J \sum_{i=1}^n (B_{i,j} - \hat{B}_{i,j})^2 \\ &= O_P(D_n L_n^{-2} J n^{-1}), \end{aligned}$$

because of (A12).

8.6. Proof of Proposition 1

The proof is based on Pinsker (1980). Because $m_1 \in \mathcal{S}_1$ the local Fourier coefficients θ_{ts} of m_1 lie in an ellipsoid. More precisely, there exist coefficients θ_{ts} ($1 \leqslant t \leqslant T$, $s \geqslant 0$) such that

$$m_1(x) = \sum_{t=1}^T \sum_{s=0}^\infty \theta_{ts} e_{ts}(x). \tag{8.33}$$

Because $m_1 \in \mathcal{S}_1$ the coefficients fulfil

$$\sum_{t=1}^{T} \sum_{s=1}^{\infty} s^{2r_1} \theta_{ts}^2 \leqslant C_1 (\pi T)^{-2r_1}. \tag{8.34}$$

Define

$$S_0(p_1) = c_T \left(\frac{\alpha(p_1)\, \sigma^2 \pi^{2r_1}}{C_1} \frac{r_1}{(r_1+1)(2r_1+1)} \right)^{-1/(2r_1+1)}, \tag{8.35}$$

where $\alpha(p_1) = \int_0^1 p_1^{-1}$. Define

$$\hat{S}_0 = S_0(\hat{p}_1) \tag{8.36}$$

where $\hat{p}_1$ is an estimator of p_1, based on observations $X_1^1, \ldots, X_n^1$. We may assume that $\hat{p}_1$ satisfies

$$\alpha(p_1) - \alpha(\hat{p}_1) = o_p(1).$$

Then $S_0(\hat{p}_1) - S_0(p_1) = o_p(1)$. Choose S such that $S \geqslant S_0(p_1) + 1$ for all $p_1 \in \mathcal{F}_1$. Then $S_0(\hat{p}_1) \leqslant S$ with probability tending to 1.

Choose the shrinking coefficients of the estimate as

$$\lambda_{ts} = \max\{0,\, 1 - (s/\hat{S}_0)^{r_1}\},$$

where $\hat{S}_0$ is defined in (8.36).

For $m_1 \in \mathcal{S}_1$ we have

$$\Delta = \int_0^1 (m_1(x) - \hat{m}_{\mathrm{O}}(x))^2 \, \mathrm{d}x = \sum_{t=1}^{T} \sum_{s=0}^{\infty} \left(\theta_{ts} - \lambda_{ts} \hat{\theta}_{ts} \right)^2. \tag{8.37}$$

Then

$$\begin{aligned} \mathrm{E}^*(\Delta) &= \sum_{t=1}^{T} \sum_{s=0}^{\infty} (1 - \lambda_{ts})^2 \theta_{ts}^2 + \mathrm{E}^* \sum_{t=1}^{T} \sum_{s=0}^{S} \lambda_{ts}^2 \left(\hat{\theta}_{ts} - \theta_{ts} \right)^2 \qquad (8.38) \\ &= \Delta_1 + \mathrm{E}^* \Delta_2. \end{aligned}$$

By application of (8.34),

$$\begin{aligned} \Delta_1 &= \sum_{t=1}^{T} \sum_{s \leqslant \hat{S}_0} \left(\frac{s}{\hat{S}_0} \right)^{2r_1} \theta_{ts}^2 + \sum_{t=1}^{T} \sum_{s > \hat{S}_0} \theta_{ts}^2 \qquad (8.39) \\ &\leqslant \sum_{t=1}^{T} \sum_{s=0}^{\infty} \left(\frac{s}{\hat{S}_0} \right)^{2r_1} \theta_{ts}^2 \\ &\leqslant \hat{S}_0^{-2r_1} C_1 (\pi T)^{-2r_1} \\ &= S_0^{-2r_1} C_1 (\pi T)^{-2r_1} (1 + o_p(1)). \end{aligned}$$

We now treat the second term on the right-hand side of (8.38). With slight changes of the notation at the beginning of Section 5 – see in particular (8.29) – we have that

$$\begin{aligned}\Delta_2 &= \|\Lambda(\hat{\theta} - \theta)\|^2 \\ &= \left\| \Lambda \left[\frac{1}{n} B^{\mathrm{T}} B \right]^{-1} \frac{1}{n} B^{\mathrm{T}} [\varepsilon - \bar{\varepsilon}] \right\|^2 .\end{aligned}$$

Here, for example, $\hat{\theta}$ and θ are vectors with double index (t, s) running over $1 \leqslant t \leqslant T$, $0 \leqslant s \leqslant \hat{S}_0$. It is easy to check that

$$\mathrm{E}^* \left\| \Lambda \left[\frac{1}{n} B^{\mathrm{T}} B \right]^{-1} \frac{1}{n} B^{\mathrm{T}} \bar{\varepsilon} \right\|^2 = O_P\left(\frac{1}{n}\right).$$

This shows that

$$\begin{aligned}\mathrm{E}^* \Delta_2 &= \mathrm{E}^* \left\| \Lambda \left[\frac{1}{n} B^{\mathrm{T}} B \right]^{-1} \frac{1}{n} B^{\mathrm{T}} \varepsilon \right\|^2 + O_P\left(\frac{1}{n}\right) \qquad (8.40) \\ &= \operatorname{trace}\left[\Lambda \left(\frac{1}{n} B^{\mathrm{T}} B \right)^{-1} \frac{1}{n^2} B^{\mathrm{T}} B \left(\frac{1}{n} B^{\mathrm{T}} B \right)^{-1} \Lambda \right] \sigma^2 + O_P\left(\frac{1}{n}\right) \\ &= \frac{1}{n} \operatorname{trace}\left[\Lambda \left(\frac{1}{n} B^{\mathrm{T}} B \right)^{-1} \Lambda \right] \sigma^2 + O_P\left(\frac{1}{n}\right).\end{aligned}$$

Because of (8.30) we have that

$$\sup_{a:\|a\|=1} \left\| \frac{1}{n} B^{\mathrm{T}} B a - \Gamma a \right\| = o_P(1),$$

where Γ is a $(T[\hat{S}_0 + 1]) \times (T[\hat{S}_0 + 1])$ diagonal matrix with diagonal elements $\Gamma_{ts,ts} = p_1(t/T)$. This implies that

$$\sup_{a:\|a\|=1} \left\| \left[\frac{1}{n} B^{\mathrm{T}} B \right]^{-1} a - \Gamma^{-1} a \right\| = o_P(1).$$

This shows that

$$\operatorname{trace}\left[\Lambda \left(\frac{1}{n} B^{\mathrm{T}} B \right)^{-1} \Lambda \right] = \sum_{t=1}^{T} \sum_{s=0}^{\hat{S}_0} p_1(t/T)^{-1} \lambda_{ts}^2 + o_P(T).$$

By plugging this into the right-hand side of (8.40) we obtain

$$\mathrm{E}^*\Delta_2 = \frac{1}{n}\,\sigma^2 \sum_{t=1}^{T} p_1\left(\frac{t}{T}\right)^{-1} \sum_{s=0}^{\hat{S}_0} \left(1 - \left\{\frac{s}{\hat{S}_0}\right\}^{r_1}\right)^2 + o_P\left(\frac{T}{n}\right) \tag{8.41}$$

$$= \frac{T\hat{S}_0}{n}\,\sigma^2 \int p_1(u)^{-1}\,\mathrm{d}u \int (1 - v^{r_1})^2\,\mathrm{d}v + o_P\left(\frac{T}{n}\right)$$

$$= \frac{T\hat{S}_0}{n}\,\sigma^2 \alpha(p_1)\,\frac{2r_1^2}{(r_1+1)(2r_1+1)} + o_P\left(\frac{T}{n}\right)$$

$$= \frac{TS_0}{n}\,\sigma^2 \alpha(p_1)\,\frac{2r_1^2}{(r_1+1)(2r_1+1)} + o_P\left(\frac{T}{n}\right).$$

The theorem now follows from (8.38), (8.39), (8.41) and the definition of S_0; see (8.35).

Acknowledgements

The research for this paper is supported in part by Deutsche Forschungsgemeinschaft MA 1026/8-2 and National Science Foundation grants SES-0352675 and SES-9910925. We wish to thank the referees for their helpful comments.

References

Belitser, E.N. and Levit, B.Y. (1995) On minimax filtering over ellipsoids. *Math. Methods Statist.*, **4**, 259–273.

Efromovich, S. (1996) On nonparametric regression for iid observations in a general setting. *Ann. Statist.*, **24**, 1126–1144.

Efromovich, S. and Pinsker, M.S. (1984) Learning algorithm for nonparametric filtering. *Avtomat. i Telemekh.*, **11**, 58–65.

Golubev, G.K. (1987) Adaptive asymptotically minimax estimators of smooth signals. *Problems Inform. Transmission*, **23**, 57–67.

Golubev, G.K. (1992) Asymptotic minimax estimation of regression in the additive model. *Problemy Peredachi Informatsii*, **28**, 3–15 (in Russian). English translation: *Problems Inform. Transmission*, **28**, 101–112 (1992).

Golubev, G.K. and Nussbaum, M. (1990) A risk bound in Sobolev class regression. *Ann. Statist.*, **18**, 758–778.

Hastie, T. and Tibshirani, R. (1991) *Generalized Additive Models*. London: Chapman & Hall.

Horowitz, J. and Mammen, E. (2004) Nonparametric estimation of an additive model with a link function. *Ann. Statist.*, **32**, 2412–2443.

Linton, O. (2000) Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory*, **16**, 502–523.

Mammen, E., Linton, O. and Nielsen, J. (1999) The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* , **27**, 1443–1490.

Nussbaum, M. (1985) Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.*, **13**, 984–997.

Opsomer, J.D. (2000) Asymptotic properties of backfitting estimators. *J. Multivariate Anal.*, **73**, 166–179.

Opsomer, J.D. and Ruppert, D. (1997) Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.*, **25**, 185–211.

Pinsker, M.S. (1980) Optimal filtering of square integrable signals in Gaussian white noise. *Problemy Peredachi Informatsii*, **16**, 52–68 (in Russian). English translation: *Problems Inform. Transmission*, **16**, 120–133 (1980).

Shorack, G.R. and Wellner, J.A. (1986) *Empirical Processes with Applications to Statistics*. New York: Wiley.

Stone, C.J. (1985) Additive regression and other nonparametric models. *Ann. Statist.*, **13**, 689–705.

van de Geer, S. (2000) *Empirical Processes in M-Estimation*. Cambridge: Cambridge University Press.

Zhang, S. and Wong, M.-Y. (2003) Wavelet threshold estimation for additive regression models. *Ann. Statist.*, **31**, 152–173.

Received May 2002 and revised August 2005