Econometric Theory, **20**, 2004, 265–300. Printed in the United States of America.
DOI: 10.1017/S026646660420202X

BOOTSTRAP INFERENCE IN SEMIPARAMETRIC GENERALIZED ADDITIVE MODELS

Wolfgang Härdle
Humboldt-Universität zu Berlin

Sylvie Huet
Institut de Recherche Agronomique

Enno Mammen
Universität Mannheim

Stefan Sperlich
Universidad Carlos III de Madrid

Semiparametric generalized additive models are a powerful tool in quantitative econometrics. With response Y, covariates X, T, the considered model is $E(Y|X;T) = G\{X^T\beta + \alpha + m_1(T_1) + \cdots + m_d(T_d)\}$. Here, G is a known link, α and β are unknown parameters, and $m_1, \ldots, m_d$ are unknown (smooth) functions of possibly higher dimensional covariates $T_1, \ldots, T_d$. Estimates of $m_1, \ldots, m_d$, α, and β are presented, and asymptotic distributions are given for both the nonparametric and the parametric part. The main focus of the paper is application of bootstrap methods. It is shown how bootstrap can be used for bias correction, hypothesis testing (e.g., component-wise analysis), and the construction of uniform confidence bands. Further, bootstrap tests for model specification and parametrization are given, in particular for testing additivity and link function specification. The practical performance of the methods is illustrated in a simulation study.

1. INTRODUCTION

Many problems in econometrics and other fields require estimating and analyzing the conditional mean $m(X,T)$ of a random response Y given covariates X and T. A traditional estimation approach for $m(x,t)$ assumes that m belongs to

This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 "Quantifikation und Simulation ökonomischer Prozesse," Humboldt-Universität zu Berlin, DFG project MA 1026/6-2, the Spanish "Dirección General de Enseñanza Superior," no. BEC2001-1270, and the grant "Nonparametric methods in finance and insurance" from the Danish Social Science Research Council. We thank Marlene Müller, Oliver Linton, and two anonymous referees for helpful discussion. Address correspondence to: Enno Mammen, Fakultät für Volkswirtschaftslehre, Universität Mannheim, L7, 3-5, D 68131 Mannheim, Germany; e-mail: emammen@rumms.uni-mannheim.de.

0266-4666/04 $12.00

a known finite-dimensional parametric family, often motivated by economic theory, identifiability conditions or practical reasons. Parameters can be estimated with $O_P(1/\sqrt{n})$ rate of convergence. Clearly, the estimation results are misleading if $m(x;t)$ is misspecified. Misspecifications may be avoided by non- or semiparametric approaches. However the nonparametric rate of convergence decreases rapidly as the dimension of the covariates increases (see, e.g., Stone, 1985), and high-dimensional nonparametric functions are difficult to interpret. A natural compromise between typical parametric and purely nonparametric models is a model of the form

$$E(Y|X = x;T = t) = G(x^T\beta + m(t)) = G(x^T\beta + \alpha + m_1(t_1) + \cdots + m_d(t_d)), \tag{1}$$

called a generalized additive partial linear model. In this paper we study the case when the link function G is known or has to be tested and coefficients α and β and the nonparametric functions $m_1,\dots,m_d$ of possibly higher dimensional covariates $T_1,\dots,T_d$ are unknown. It is well known that those models can be estimated at a rate typical for the lower dimensional explanatory variables T_j (Stone, 1985).

The special case of generalized partially linear models (with $d = 1$) is well studied (see, e.g., Ai, 1997; Mammen and van de Geer, 1997; Severini and Staniswalis, 1994). We will extend the latter approach, i.e., the iterative application of smoothed local and unsmoothed global likelihoods. The related model $E[Y|X,T] = G\{\beta^T X + m(T^T\alpha)\}$ is studied by Carroll, Fan, Gijbels, and Wand (1997). Their aim is dimension reduction of the variable T by projection, but the fitted nonparametric transformation m is quite difficult to interpret.

Additive and generalized additive models play an important role in economic theory (see, e.g., Leontief, 1947; Goldberger, 1964; Deaton and Muellbauer, 1980). Apart from their statistical advantages they allow for the analysis of subsets of regressors and permit decentralization in optimization and decision making. Projection smoothers using backfitting techniques are considered in Hastie and Tibshirani (1990), but asymptotic theory for this technique is rather complicated (see Mammen, Linton, and Nielsen, 1999; Opsomer and Ruppert, 1999). An alternative approach that allows a detailed asymptotic analysis uses approximations by linear spaces (e.g., of regression splines) with increasing dimension (see Hansen, Huang, Kooperberg, Stone, and Truong, 2002). Horowitz (2001) proposes estimates of additive components based on partial derivatives of the full-dimensional regression function. His approach also allows an unknown link function. Further, Tjøstheim and Auestadt (1994) and Linton and Nielsen (1995) introduce the marginal integration approach. Marginal integration is applied to generalized additive models by Linton and Härdle (1996). We will use the approach of Severini and Staniswalis (1994) and combine it with marginal integration. This is done for practical and theoretical reasons. In particular, this approach will allow for a detailed asymptotic distribution theory.

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

The main subject of this paper is the introduction of bootstrap procedures for (1). Nonparametric bootstrap tests for generalized partially linear models can be found in Härdle, Mammen, and Müller (1998). In our more complex case, the integration estimate of an additive component has bias terms that depend on the shape of the other additive components. This complicates the data analytic interpretation of nonparametric fits. We will show how bootstrap can be used to correct for these terms. Bootstrap tests will be considered also for variable selection, parametric specifications, and testing additivity. We will argue that bootstrap is a natural method for these problems. Alternative methods could be based on asymptotic expansions to get bias approximations or normal approximations, respectively, and to use plug-in estimates. In our setup these expansions are rather complex and may lead only to crude approximations. So we expect that the structure of the model will be better mimicked by the bootstrap.

The paper is organized as follows. In the next section we introduce estimates for the parameters and the nonparametric components of model (1). Section 3 presents several applications of bootstrap for analyzing the nonparametric components, starting with bias corrections for the nonparametric estimates. What follows are bootstrap tests for different null hypotheses about the components. In the last part procedures and theory for uniform confidence bands are given. In Section 4 the presented methodology is studied in simulations. Assumptions, asymptotic theory for the estimators, and proofs are postponed to the Appendix.

2. QUASI-LIKELIHOOD ESTIMATION IN GENERALIZED ADDITIVE MODELS

In this section we will discuss our approach for generalized additive models. Our estimation procedure starts with the iterative algorithm of Severini and Staniswalis (1994), and in a second step the additive components are fitted by marginal integration. For a better understanding we first discuss the special case of binary response models. For the general case of generalized additive models our approach will be introduced in Section 2.2. For a discussion of binary response models see also Horowitz (1998). A detailed introduction to quasi-likelihood can be found in McCullagh and Nelder (1989).

2.1. Additive Binary Response Models

In an additive binary response model independent and identically distributed (i.i.d.) tuples (Y_i, X_i, T_i) are observed $(i = 1,\dots,n)$, where $T_i = (T_{i,1},\dots,T_{i,d})$ is a random variable with components $T_{i,j}$ in $I\!R^{q_j}$, X_i is in $I\!R^p$, and $Y_i \in \{0,1\}$. Conditionally given (X_i, T_i) the variable Y_i is distributed as a Bernoulli variable with parameter $G\{X_i^T\beta + \alpha + m_1(T_{i,1}) + \cdots + m_d(T_{i,d})\}$ where G is a known (link) function, β an unknown parameter in $I\!R^p$, α an unknown scalar, and $m_j : I\!R^{q_j} \to I\!R$ unknown smooth functions. For identifiability we set

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

$E\ w_j(T_{i,1}) m_j(T_{i,1}) = 0\ \forall j$ for some weight functions w_j. Given (X_i, T_i), the likelihood of Y_i is

$$Q(\mu_i; Y_i) = Y_i \log \mu_i + (1 - Y_i) \log(1 - \mu_i), \tag{2}$$

where $\mu_i = G\{X_i^T \beta + \alpha + m_1(T_{i,1}) + \cdots + m_d(T_{i,d})\}$. The likelihood function is given by

$$\mathcal{L}(m^+, \beta) = \sum_{i=1}^{n} Q(\mu_i; Y_i), \tag{3}$$

where $m^+(t)$ is the additive function $\alpha + m_1(t_1) + \cdots + m_d(t_d)$.

We now discuss how the additive nonparametric components can be estimated. Without loss of generality, we will do this for the first component m_1. We write $r = q_1$, $s = q_2 + \cdots + q_d$ and define the smoothed likelihood

$$\mathcal{L}^S(m^+, \beta) = \int \sum_{i=1}^{n} K_h(t_1 - T_{i,1}) L_g(t_{-1} - T_{i,-1}) Q[G\{X_i^T \beta + m^+(t)\}; Y_i]\, dt, \tag{4}$$

where the vector $T_i = (T_{i,1}, \ldots, T_{i,d})$ is a random variable with components $T_{i,j}$ in $\mathbb{R}^{q_j}$. For a vector $u = (u_1, \ldots, u_d)$ with components u_j in $\mathbb{R}^{q_j}$ we denote $(u_2, \ldots, u_d)^T$ by u_{-1}; similarly, we write $T_{i,-1} = (T_{i,2}, \ldots, T_{i,d})^T$. For a kernel function L defined on $\mathbb{R}^s$ put $L_g(v) = (g_1 \cdot \cdots \cdot g_s)^{-1} L(g_1^{-1} v_1, \ldots, g_s^{-1} v_s)$ and for simplicity assume that L is a product kernel $L = \prod_{j=1}^{s} L_j$. Similarly, define $K_h(v) = h^{-1} K(h^{-1} v)$ for $v \in \mathbb{R}^r$ and bandwidth vector $h \in \mathbb{R}^r$ with product kernel $K = \prod_{j=1}^{r} K_j$. The bandwidth vector g is related to smoothing in direction of the "nuisance" covariates. The relative speed of the elements of g to the elements of h and the choice of these bandwidths will be discussed subsequently.

Following Severini and Staniswalis (1994), we base our estimates on an iterative application of smoothed local and unsmoothed global likelihood functions. We define for $\beta \in B$

$$\hat{m}_\beta(t) = \arg\max_\eta \sum_{i=1}^{n} K_h(t_1 - T_{i,1}) L_g(t_{-1} - T_{i,-1}) Q[G\{X_i^T \beta + \eta\}; Y_i], \tag{5}$$

$$\hat{\beta} = \arg\max_{\beta \in B} \mathcal{L}(\hat{m}_\beta, \beta), \tag{6}$$

$$\hat{m} = \hat{m}_{\hat{\beta}}. \tag{7}$$

Equation (5) may be written as $\hat{m}_\beta = \arg\max_m \mathcal{L}^S(m, \beta)$. The result $\hat{m}$ is a multivariate kernel estimate of m^+ that does not use the additive structure of m^+. This $\hat{m}$ will be used in an additional step to obtain estimates $\hat{\alpha}, \hat{m}_1, \ldots, \hat{m}_d$ of the additive components $\alpha, m_1, \ldots, m_d$. The final additive estimate of $m^+(t)$ will then be given by $\hat{\alpha} + \hat{m}_1(t_1) + \cdots + \hat{m}_d(t_d)$.

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

For the estimation of the nonparametric component m_1 the marginal integration method is applied. It is motivated by the fact that, up to a constant, $m_1(t_1)$ is equal to

$$\left\{ \int w_{-1}(v)\, dv \right\}^{-1} \int w_{-1}(v)\, m^{+}(t_1, v)\, dv$$

or

$$\left\{ (1/n) \sum_{i=1}^{n} w_{-1}(T_{i,-1}) \right\}^{-1} (1/n) \sum_{i=1}^{n} w_{-1}(T_{i,-1})\, m^{+}(t_1, T_{i,-1})$$

for a weight function w_{-1}. Note that this method does not use iterations so that the explicit definition allows a detailed asymptotic analysis. A weight function w_{-1} is used for two reasons: it may be useful to avoid problems at the boundary, and it can be chosen to minimize the variance (compare Fan, Härdle, and Mammen, 1998). So we define

$$\overline{m}_1(t_1) = \frac{\frac{1}{n} \sum_{i=1}^{n} w_{-1}(T_{i,-1}) \hat{m}(t_1, T_{i,-1})}{\frac{1}{n} \sum_{i=1}^{n} w_{-1}(T_{i,-1})}, \tag{8}$$

which estimates the function m_1 up to a constant. An estimate of the function m_1 is given by norming (with a weight function w_1)

$$\hat{m}_1(t_1) = \overline{m}_1(t_1) - \frac{\frac{1}{n} \sum_{i=1}^{n} w_1(T_{i,1}) \overline{m}_1(T_{i,1})}{\frac{1}{n} \sum_{i=1}^{n} w_1(T_{i,1})}. \tag{9}$$

The additive constant α can be estimated by

$$\hat{\alpha} = \frac{\frac{1}{n} \sum_{i=1}^{n} w_0(T_i)[\hat{m}(T_i) - \hat{m}_1(T_{i,1}) - \cdots - \hat{m}_d(T_{i,d})]}{\frac{1}{n} \sum_{i=1}^{n} w_0(T_i)}. \tag{10}$$

Again, the weight functions w_0 and w_1 may be useful to avoid problems at the boundary. After having estimated the remaining nonparametric components analogously, the final estimate of m is

$$\hat{m}^{+}(t) = \hat{\alpha} + \hat{m}_1(t_1) + \cdots + \hat{m}_d(t_d). \tag{11}$$

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

2.2. Semiparametric Generalized Additive Models

We now come to the discussion of estimation in semiparametric generalized additive models. Suppose that we observe an independent sample $(Y_1, X_1, T_1), \ldots, (Y_n, X_n, T_n)$ with $E[Y_i|X_i, T_i] = G\{X_i^T\beta + m(T_i)\}$. Additional assumptions on the conditional distribution of Y_i will be given subsequently. For a positive function V the quasi-likelihood (QL) function is defined as

$$Q(\mu; y) = \int_{\mu}^{y} \frac{(s-y)}{V(s)}\, ds, \tag{12}$$

where μ is the (conditional) expectation of Y, i.e., $\mu = G\{X^T\beta + m(T)\}$. The QL function has been introduced for the case that the conditional variance of Y is equal to $\sigma^2 V(\mu)$ where σ^2 is an unknown scale parameter. The function Q can be motivated by the following two considerations: clearly, $Q(\mu; y)$ is equal to $-\frac{1}{2}(\mu - y)^2 v^{-1}$ where v^{-1} is a weighted average of $1/V(s)$ for s between μ and y. Consequently, maximum QL estimates can be interpreted as a modification of weighted least squares. Another motivation comes from the fact that for exponential families the maximum QL estimate coincides with the maximum likelihood estimate. Note that the maximum likelihood estimate $\hat{\theta}$, based on an i.i.d. sample $Y_1, \ldots, Y_n$ from an exponential family with mean $\mu(\theta)$ and variance $V\{\mu(\theta)\}$, is given by

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} Q(\mu(\theta); Y_i) = 0.$$

We consider three models.

Model A. $(Y_1, X_1, T_1), \ldots, (Y_n, X_n, T_n)$ is an i.i.d. sample with $E[Y_i|X_i, T_i] = G\{X_i^T\beta + m(T_i)\}$.

Model B. Model A holds, and the conditional variance of Y_i is equal to $\mathrm{Var}[Y_i|X_i, T_i] = \sigma^2 V(\mu_i)$ where $\mu_i = G\{X_i^T\beta + m(T_i)\}$ and where σ^2 is an unknown scale parameter.

Model C. Model A holds, and the conditional distribution of Y_i belongs to an exponential family with mean μ_i and variance $V(\mu_i)$ with μ_i as in Model B.

The QL function is well motivated for Models B and C. The more general Model A is included for discussion of robustness issues, i.e., to discuss the case of a wrongly specified conditional variance in Models B and C. If not stated otherwise, all the following remarks and results treat the most general Model A. The QL function and the smoothed QL function are now defined as in (3) and (4) with (2) replaced by (12). The estimates $\hat{m}_\beta$, $\hat{\beta}$, $\hat{m}$, $\overline{m}_1$, $\hat{m}_1$, $\hat{m}^+$, and $\hat{\alpha}$ are defined as in (5)–(10). Asymptotics for $\hat{m}_1$ are presented in Section A.2 of the Appendix. In particular, Lemma A2.1 shows that

$$\sqrt{nh}\{\hat{m}_1(t_1) - m_1(t_1) - \delta_n^1(t_1)\}$$

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

converges to a centered Gaussian variable where the bias $\delta_n^1(t_1)$ is of the form $Ah_+^2 + Bg_+^2 + o_P(h_+^2 + g_+^2)$, where $h_+ = \max_{1 \le j \le r} h_j$ and $g_+ = \max_{1 \le j \le s} g_j$. For a definition of the terms Ah_+^2 and Bg_+^2 see Lemma A2.1. This lemma does not require that g_+ is of smaller order than h_+, an assumption that has been made in previous papers. Clearly, then the bias term Bg_+^2 would be asymptotically negligible, and therefore asymptotics suggests the choice $g_+ = o(h_+)$. However, stochastic and numerical stability of the preestimator $\hat{m}$ demand that $h_1 \times \cdots \times h_r \cdot g_1 \times \cdots \times g_s$ is large. Otherwise too few observations would lie in the local support of the multidimensional kernel. Often, in practice even larger values for g_j than for h_l are needed for a satisfactory performance of $\hat{m}$. The constant A in the bias depends on the value of m_1' and m_1'' at t_1, whereas the constant B depends on averages of powers of $m_j'(t_j)$ and $m_j''(t_j)$ over t_j and over $j \neq 1$. Typically the averaging leads to small values of B. For more discussion, especially on optimal rates and efficiency, we refer to Härdle, Huet, Mammen, and Sperlich (1998).

The remaining additive components m_j for $j = 2, \ldots, d$ are estimated in analogy to m_1. It can be checked that the estimates $\hat{m}_1(t_1), \ldots, \hat{m}_d(t_d)$ are asymptotically independent. The variance of the estimate $\hat{m}_1(t_1)$ can be consistently estimated (see Section A.2 of the Appendix). Consistency and asymptotic normality of β are shown in Lemma A2.2. It turns out that for asymptotic unbiasedness with rate $\sqrt{n}$ no undersmoothing is required in the nonparametric estimation. Further, an explicit expression for the asymptotic variance is given that, however, depends on unknown terms as, e.g., on the function $m(\cdot)$.

3. BOOTSTRAP APPLICATIONS IN GENERALIZED ADDITIVE MODELS

Three versions of bootstrap will be considered here. The first version is wild bootstrap, which is related to proposals of Wu (1986), Beran (1986), and Mammen (1992) and was first proposed by Härdle and Mammen (1993) in nonparametric settings. Note that in Model A the conditional distribution of Y is not specified besides the conditional mean. The wild bootstrap procedure works as follows.

Step 1. Calculate residuals $\hat{\varepsilon}_i = Y_i - \hat{\mu}_i$.

Step 2. Generate n i.i.d. random variables $\varepsilon_1^*, \ldots, \varepsilon_n^*$ with mean 0 and variance 1 and that fulfill for a constant C that $|\varepsilon_i^*| \le C$ (a.s.) for $i = 1, \ldots, n$.

Step 3. Put $Y_i^* = \hat{\mu}_i + \hat{\varepsilon}_i \varepsilon_i^*$ for $i = 1, \ldots, n$, where

$$\hat{\mu}_i = G\{X_i^T \hat{\beta} + \hat{\alpha} + \hat{m}_1(T_{i,1}) + \hat{m}_2(T_{i,2}) + \cdots + \hat{m}_d(T_{i,d})\}.$$

For Model B we propose a resampling scheme that takes care of the specification of the conditional variance of Y. For this reason, we modify Step 3 by putting $Y_i^* = \hat{\mu}_i + \hat{\sigma} V\{\hat{\mu}_i\}^{1/2} \varepsilon_i^*$ for $i = 1, \ldots, n$. Here $\hat{\sigma}^2$ is a consistent estimate of σ^2. In this case the condition that $|\varepsilon_i^*|$ is bounded can be weakened to the assumption that ε_i^* has subexponential tails; i.e., for a constant C it holds

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

that $E(\exp\{[|\varepsilon_i^*|/C]\}) \le C$ for $i = 1, \dots, n$ (compare Assumption (A2) in the Appendix).

In the special situation of Model C (semiparametric generalized linear model), $Q(y;\mu)$ is the log-likelihood. Then the conditional distribution of Y_i is specified by $\mu_i = G\{X_i^T\beta + m^+(T)\}$. In this model we generate n independent $Y_1^*, \dots, Y_n^*$ with distributions defined by $\hat{\mu}_i$, respectively. In the binary response example that we considered in Section 2, Y_i is a Bernoulli variable with parameter $\mu_i = G[X_i^T\beta + m^+(T)]$. Hence, here we resample from a Bernoulli distribution with parameter $\hat{\mu}_i$.

3.1. Bias Correction

Lemma A2.1 in the Appendix shows that if the elements of the bandwidth vectors h and g are of the same order, the bias of $\hat{m}_1(t_1)$ depends on the shape of the other additive components $m_2, \dots, m_d$. This may lead to wrong interpretations of the estimate $\hat{m}_1$. Bootstrap bias estimates will help to judge such effects.

In all three resampling schemes, one uses the data $(X_1, T_1, Y_1^*), \dots, (X_n, T_n, Y_n^*)$ to calculate the estimate $\hat{m}_1^*$. This is done with the same bandwidth h for the component t_1 and with the same g for the other $d-1$ components. The bootstrap estimate of the mean of $\hat{m}_1(t_1)$ is given by $E^*\hat{m}_1^*(t_1)$, where E^* denotes the conditional expectation given the sample $(X_1, T_1, Y_1), \dots, (X_n, T_n, Y_n)$. The bias corrected estimate of $m_1(t_1)$ is defined by

$$\hat{m}_1^B(t_1) = \hat{m}_1(t_1) - \hat{\delta}_n^1(t_1), \quad \text{where } \hat{\delta}_n^1(t_1) = E^*\hat{m}_1^*(t_1) - \hat{m}_1(t_1).$$

The theorem shows that the bias terms of order g_+^2 are removed by this construction.

THEOREM 3.1. *Assume that Model A, Model B, or Model C holds and that the corresponding version of bootstrap is used. Suppose further that Assumptions (A1)–(A11) in the Appendix apply and that assumptions analogous to (A3) and (A4) hold for the estimation of the other additive components m_j for $j = 2, \dots, d$ (h being always the bandwidth used for the estimated component m_j and g the bandwidth for the nuisance components). Furthermore, suppose that the elements of h and g tend to zero and that $nh_1 \cdot \dots \cdot h_r g_1^2 \cdot \dots \cdot g_s^2 (\log n)^{-2}$ tends to infinity. Then it holds that*

$$\hat{m}_1^B(t_1) - m_1(t_1) = O_p\{h_+^4 + g_+^4 + (nh_1 \cdot \dots \cdot h_r)^{-1/2}\}, \tag{13}$$

where again $h_+ = \max_{1\le j\le r} h_j$ and $g_+ = \max_{1\le j\le s} g_j$.

Bootstrap applications in nonparametric regression often use resampling from a modified estimate of the regression function. Suppose, e.g., that in the third step of the bootstrap algorithm $\hat{\mu}_i$ is replaced by $G\{X_i^T\hat{\beta} + \hat{\alpha} + \hat{m}_1^O(T_{i,1}) + \hat{m}_2(T_{i,2}) + \dots + \hat{m}_d(T_{i,d})\}$, where $\hat{m}_1^O$ is defined as $\hat{m}_1$ but with bandwidth vector h^O instead of h. Then if $h_j^O/h_+ \to \infty$ $(1 \le j \le r)$ one can show that the left-hand side of (13) is of order $O_p\{(h_+^O)^4 + g_+^4 + (nh_1^O \cdots h_r^O)^{-1/2}\}$, where h_+^O

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

is the maximal element of h^O. For appropriate choices of h^O, e.g., for h^O with $(h_+^O)^4$ and $(nh_1^O \cdots h_r^O)^{-1/2}$ of the same asymptotic order, this is of smaller order than the right-hand side of (13) with resampling from $\hat{m}_1$.

3.2. Componentwise Hypothesis Testing

Interesting shape characteristics may be visible in plots of estimates of the additive components. The complicated nature of the model, though, makes it difficult to judge the statistical significance of such findings. Hypothesis tests in addition to uniform confidence bands are useful tools to analyze and interpret fitted components. We now discuss tests of the hypothesis that one component is linear:

$$H_0 : m_1(t_1) = \gamma_1 t_1 \quad \text{for all } t_1 \text{ and a scalar } \gamma_1. \tag{14}$$

Extensions to variable selection problems ($H_0 : m_1 \equiv 0$) or tests of polynomial forms are straightforward; see also the discussion that follows.

Our test is a modification of a general approach by Hastie and Tibshirani (1990). In semiparametric setups they propose to apply likelihood ratio tests and to use χ^2 approximations for the calculation of critical values. Approximate degrees of freedom are heuristically derived by calculating the expectation of asymptotic expansions of the test statistic under the null hypothesis. Here we propose more accurate distributional approximations. Furthermore, in the definition of the test statistic we correct for the bias of the nonparametric estimate. Our test statistic is asymptotically normal, but the convergence to the normal limit is very slow as mathematical arguments and simulations indicate. Therefore we propose the bootstrap for the calculation of critical values. Bias correction will be used in the test because otherwise it will have a nonnegligible effect on the power. For this reason, $\hat{m}_1(t_1)$ is compared with a bootstrap estimate of its expectation under the hypothesis.

First, we calculate semiparametric estimates for the hypothetic model

$$E(Y_i|X_i,T_i) = G\{X_i^T\beta + \alpha + \gamma_1 T_{i,1} + m_2(T_{i,2}) + \cdots + m_d(T_{i,d})\}.$$

Note that the α occurring in the preceding equation can be different from the α defined in Section 2.1 because X_i is replaced by $(X_i, T_{i,1})$. Estimation of the parametric components β, α, and γ_1 and of nonparametric components $m_2,\ldots,m_d$ can be done as in Section 2.1. This defines estimates $\tilde{\beta}, \tilde{\alpha}, \tilde{\gamma}_1, \tilde{m}_2, \ldots, \tilde{m}_d$. Set

$$\tilde{\mu}_i = G\{X_i^T\tilde{\beta} + \tilde{\alpha} + \tilde{\gamma}_1 T_{i,1} + \tilde{m}_2(T_{2,i}) + \cdots + \tilde{m}_d(T_{i,d})\}.$$

Second, for the bootstrap we proceed as follows: generate independent samples $(Y_1^*,\ldots,Y_n^*)$ (compare Section 3) but now with μ_i replaced by $\tilde{\mu}_i$. Then, using the data $(X_1,T_1,Y_1^*),\ldots,(X_n,T_n,Y_n^*)$ calculate the estimate $\hat{m}_1^*$. The bootstrap estimate of the mean of $\hat{m}_1(t_1)$ is given by $E^*\hat{m}_1^*(t_1)$, where E^* denotes the conditional expectation given the sample $(X_1,T_1,Y_1),\ldots,(X_n,T_n,Y_n)$. Third, we define the test statistic

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

$$R = \sum_{i=1}^{n} w(T_i) \frac{[G'\{X_i^T\hat{\beta} + \hat{m}^+(T_i)\}]^2}{V(G\{X_i^T\hat{\beta} + \hat{m}^+(T_i)\})} \{\hat{m}_1(T_{i,1}) - E^*\hat{m}_1^*(T_{i,1})\}^2 \qquad \textbf{(15)}$$

with $\hat{m}^+(t) = \hat{\alpha} + \hat{m}_1(t_1) + \cdots + \hat{m}_d(t_d)$. The weights $[G'\{\dots\}]^2/V(G\{\dots\})$ in the summation of the test statistic are motivated by likelihood considerations (see Härdle et al., 1998) but could be replaced by some other weights. The test statistic R has an asymptotic normal distribution (see Lemma A3.1 in the Appendix). Mean and variance can be consistently estimated, and thus critical values for the test could be calculated using the normal approximation. But as mentioned before this approximation does not perform well. Again we recommend using bootstrap approximations. The bootstrap estimate of the distribution of R is given by the conditional distribution of the test statistic R^*, defined by

$$R^* = \sum_{i=1}^{n} w(T_i) \frac{[G'\{X_i^T\hat{\beta} + \hat{m}^+(T_i)\}]^2}{V\{X_i^T\hat{\beta} + \hat{m}^+(T_i)\}} \{\hat{m}_1^*(T_{i,1}) - E^*\hat{m}_1^*(T_{i,1})\}^2. \qquad \textbf{(16)}$$

The conditional distribution $\mathcal{L}^*(R^*)$ (given the original data $(X_1, T_1, Y_1), \dots, (X_n, T_n, Y_n)$) is our bootstrap estimate of $\mathcal{L}(R)$ (on the hypotheses (14)). Here, $\mathcal{L}(R)$ denotes the distribution of R. The following theorem states consistency of the bootstrap.

THEOREM 3.2. *Assume that Model A, Model B, or Model C holds and that the corresponding version of bootstrap is used. Furthermore suppose that assumptions (A1)–(A11) in the Appendix hold with X_i replaced by $(X_i, T_{i,1})$. Then, if additionally, $n^{1/2}h_1 \cdot \cdots \cdot h_r g_1^2 \cdot \cdots \cdot g_s^2 (\log n)^{-1} \to \infty$ and if all elements of h and g are of order $o(n^{-1/8})$, on the hypotheses (14), it holds that*

$$d_K\{\mathcal{L}^*(R^*), \mathcal{L}(R)\} \xrightarrow{P} 0,$$

where d_K denotes the Kolmogorov distance, which is defined for two probability measures μ and ν (on the real line) as $d_K(\mu,\nu) = \sup_{t\in\mathbb{R}}|\mu(X \le t) - \nu(X \le t)|$.

With similar arguments as in Härdle and Mammen (1993) one shows that the test R has nontrivial asymptotic power for deviations from the linear hypothesis of order $n^{-1/2}(h_1 \cdot \cdots \cdot h_r)^{-1/4}$. This means that the test does not reject alternatives that have a distance of order $n^{-1/2}$. However, the test also detects local deviations (of order $n^{-1/2}(h_1 \cdot \cdots \cdot h_r)^{-1/4}$) that are concentrated on shrinking intervals with length of order h. The test may be compared with overall tests that achieve nontrivial power for deviations of order $n^{-1/2}$. Typically, such tests have poorer power performance for deviations that are concentrated on shrinking intervals. For our test, the choice of the bandwidth determines how sensitively the test reacts on local deviations; i.e., for smaller h the test detects deviations that are more locally concentrated but at the cost of a poorer power performance for more global deviations. In particular, as an extreme

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

case one can consider the case of a constant bandwidth h. This case is not covered by our theory. It can be shown that in that case R is an $n^{-1/2}$-consistent overall test.

Finally we want to emphasize that the same procedure works for any other linearly parameterized hypothesis

$$H_0 : m_1(t_1) = \theta_1 f_1(t_1) + \cdots + \theta_q f_q(t_1),$$

where $\theta_1, \ldots, \theta_q$ are unknown parameters but $f_1, \ldots, f_q$ are given. Moreover, the results of this section can be extended to tests of other parametric hypotheses on m_1:

$$H_0 : m_1(t_1) = m_\theta(t_1) \quad \text{for all } t_1 \text{ and a parameter } \theta, \tag{17}$$

where $\{m_\theta : \theta \in \Theta\}$ is a parametric family. This can be done similarly as in Härdle and Mammen (1993). However, this requires an asymptotic study of parametric estimates in the model (1) with parametric specification (17) for m_1.

Using an approach similar to the approach described earlier, one can construct F-type tests on the coefficients β. For testing $H_0 : H\beta = c$ versus $H_1 : H\beta \neq c$ (with a $k \times p$ matrix H of rank $k \leq p$ and a constant $c \in \mathbb{R}^k$ for a $k \geq 1$) a natural test statistic is defined as $R_\beta = (H\hat{\beta} - c)^T \times (H\hat{I}^{-1}H^T)^{-1}(H\hat{\beta} - c)$, where $\hat{I}$ is a consistent estimate of the matrix I, defined in Lemma A2.2. A natural estimate of I would be the bootstrap estimate. According to Lemma A2.2, on the hypothesis R_β has a central χ^2 distribution. This asymptotic result could be used for the approximate calculation of critical values. As before we recommend applying bootstrap. Then R_β will be compared with its bootstrap analog $R_\beta^* = (H\hat{\beta}^* - c)^T(H\hat{I}^{-1}H^T)^{-1}(H\hat{\beta}^* - c)$. For simplicity, the same (bootstrap) covariance estimate has been used in the calculation of R_β and R_β^*.

3.3. Testing Separability and Interactions

First note that our estimate of m_1 is robust against nonadditivity of the other components. In fact, in the construction of the estimate it is only used that $m(x;t)$ is of the form

$$G\{x^T\beta + \alpha + m_1(T_1) + m_{2,\ldots,d}(T_2, \ldots, T_d)\} \tag{18}$$

for an arbitrary function $m_{2,\ldots,d}$. It is not assumed that the function $m_{2,\ldots,d}$ is additive, i.e., $m_{2,\ldots,d}(T_2, \ldots, T_d) = m_2(T_2) + \cdots + m_d(T_d)$. Also in the case that $m(x;t)$ is not of the form (18), the estimate $\hat{m}_1$ makes sense because then it estimates the average (or marginal) effect of T_1. Nevertheless the hypothesis of additivity is of interest in its own right and an important step in a model choice procedure. Following the idea of Sperlich, Tjøstheim, and Yang (2002), we con-

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

sider a split of the first covariate T_1 into two components $T_{1:1}$ and $T_{1:2}$ and consider the hypothesis

$$H_0 : m_1(t_1) = m_{1:1}(t_{1:1}) + m_{1:2}(t_{1:2}). \tag{19}$$

For other approaches to test additivity, see also Gozalo and Linton (2001). Estimates of $m_{1:1}$ and $m_{1:2}$ are constructed by marginal integration:

$$\hat{m}_{1:1}(t_{1:1}) = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_1(t_{1:1}, T_{i,1:2}) w(T_{i,1:2}),$$

$$\hat{m}_{1:2}(t_{1:2}) = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_1(T_{i,1:1}, t_{1:2}) w(T_{i,1:1})$$

so that $\hat{m}_{1:1,2}(t_1) = \hat{m}_1(t_1) - \hat{m}_{1:1}(t_{1:1}) - \hat{m}_{1:2}(t_{1:2})$ is an estimate for the first-order interaction of $T_{1:1}$ and $T_{1:2}$.

For testing hypothesis (19) we proceed similarly as in Section 3.2. We define

$$\begin{aligned} R_{inter} &= \sum_{i=1}^{n} w(T_i) \frac{[G'\{X_i^T\hat{\beta} + \hat{m}^+(T_i)\}]^2}{V(G\{X_i^T\hat{\beta} + \hat{m}^+(T_i)\})} \\ &\quad \times \{\hat{m}_{1:1,2}(T_{i,1:1}, T_{i,1:2}) - E^*\hat{m}^*_{1:1,2}(T_{i,1:1}, T_{i,1:2})\}^2, \end{aligned}$$

where $m^*_{1:1,2}$ is an estimate based on a bootstrap sample. Bootstrap samples are generated as in Section 3.2 but now with $\tilde{\mu}_i$ replaced by

$$G\{X_i^T\hat{\beta} + \hat{\alpha} + \hat{m}_{1:1}(T_{i,1:1}) + \hat{m}_{1:2}(T_{i,1:2}) + \hat{m}_2(T_{i,2}) + \cdots + \hat{m}_d(T_{i,d})\}.$$

The test statistic R_{inter} has an asymptotic normal distribution (see Lemma A3.2 in the Appendix). The bootstrap estimate of the distribution of R_{inter} is given by the conditional distribution of the test statistic R^*_{inter}, with

$$\begin{aligned} R^*_{inter} &= \sum_{i=1}^{n} w(T_i) \frac{[G'\{X_i^T\hat{\beta} + \hat{m}^+(T_i)\}]^2}{V\{X_i^T\hat{\beta} + \hat{m}^+(T_i)\}} \\ &\quad \times \{\hat{m}^*_{1:1,2}(T_{i,1:1}, T_{i,1:2}) - E^*\hat{m}^*_{1:1,2}(T_{i,1:1}, T_{i,1:2})\}^2, \end{aligned} \tag{20}$$

where $\hat{m}^*_{1:1,2}$ is defined as $\hat{m}_{1:1,2}$ but now from a bootstrap sample instead of the original sample.

THEOREM 3.3. *Under the assumptions of Theorem 3.2, on the hypotheses (19), it holds that*

$$d_K\{\mathcal{L}^*(R^*_{inter}), \mathcal{L}(R_{inter})\} \xrightarrow{P} 0.$$

3.4. Testing the Link Function

Härdle, Mammen, and Proenca (2001) introduce a bootstrap test for the null hypothesis of a parametric generalized linear versus a single index model. We extend here their approach to test

$$H_0 : E[Y|X,T] = G\{v(T,X,\beta)\} \text{ versus} \tag{21}$$

$$H_1 : E[Y|X,T] = H\{v(T,X,\beta)\} \text{ where } H(\cdot) \text{ is an unknown function} \tag{22}$$

with $v(T,X,\beta) = \beta^T X + \alpha + m_1(T_1) + \cdots + m_d(T_d)$. We recommend a test statistic of the form

$$h_L^{1/2} \sum_{i=1}^{n} w(\hat{v}_i) \frac{\sum_{j \neq i}^{n} [Y_j - G(\hat{v}_j)] K(\{\hat{v}_j - \hat{v}_i\}/h_L)}{\sum_{j \neq i}^{n} K(\{\hat{v}_j - \hat{v}_i\}/h_L)} [Y_i - G(\hat{v}_i)], \tag{23}$$

where h_L is an additional bandwidth and where $\hat{v}_i = \hat{\beta}^T X_i + \hat{\alpha} + \hat{m}_1(T_{i,1}) + \cdots + \hat{m}_d(T_{i,d})$. For further details see also Section 4.

3.5. Uniform Bootstrap Confidence Bands

To construct uniform confidence bands we first define

$$S = \sup_{t_1} w_1(t_1) |\hat{m}_1(t_1) - m_1(t_1) - \delta_n^1(t_1)| \hat{\sigma}_1^{-1}(t_1),$$

where $\hat{\sigma}_1^2(t_1)$ is the estimate of the variance of $\hat{m}_1(t_1)$, defined in equation (A.2) in the Appendix. In the simulation study in Section 4 we also use a bootstrap estimate of $\sigma_1(t_1)$. The distribution of S can be estimated by bootstrap as introduced in the beginning of Section 3. This defines the statistic $S^* = \sup_t w_1(t_1)|\hat{m}_1^*(t_1) - E^*\hat{m}_1^*(t_1)|\hat{\sigma}_1^{-1}(t_1)$. In the definition of S^* the norming $\hat{\sigma}(t_1)$ could be replaced by $\hat{\sigma}_1^*(t_1)$. We write $S^{**} = \sup_t w_1(t_1)|\hat{m}_1^*(t_1) - E^*\hat{m}_1^*(t_1)|[\hat{\sigma}_1^*]^{-1}(t_1)$. Here $\hat{\sigma}_1^*(t_1)$ is an estimate of the variance of $\hat{m}_1^*(t_1)$, that is defined similarly as $\hat{\sigma}_1(t_1)$ but that is calculated with a bootstrap resample instead of with the original sample. The first norming helps to save computation time; for the second choice bootstrap theory from other setups suggests higher order accuracy of bootstrap. Nevertheless, both bootstrap procedures can be used to construct valid uniform confidence bands:

THEOREM 3.4. *Assume that Model A, Model B, or Model C holds and that the corresponding version of bootstrap is used. Furthermore suppose that assumptions (A1)–(A11) apply, that all elements of h and g are of order $o(n^{-1/8})$, and that $nh_1 \cdot \cdots \cdot h_r g_1^2 \cdot \cdots \cdot g_s^2 (\log n)^{-2} \to \infty$. Then it holds that*

$$d_K\{\mathcal{L}^*(S^*), \mathcal{L}(S)\} \xrightarrow{P} 0, \qquad d_K\{\mathcal{L}^*(S^{**}), \mathcal{L}(S)\} \xrightarrow{P} 0.$$

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

This gives uniform confidence intervals for $m_1(t_1) - \delta_n^1(t_1)$. For confidence bands of m_1 one needs a consistent estimate of $\delta_n^1(t_1)$. This could be done by plug-in or by bootstrap. Both approaches require oversmoothing, i.e., choice of a bandwidth vector h^O with $h_j^O/h_+ \to \infty$; see also the discussion after Theorem 3.1. For related discussions in nonparametric estimation see Eubank and Speckman (1993) and Neumann and Polzehl (1998).

4. A SIMULATION STUDY

We now illustrate the performance of our methods in small samples. Simulation results are given for different tests and for confidence bands. Level accuracy is checked for testing linearity of an additive component and for testing the specification of the link function. For the first test problem power functions also are calculated. Furthermore, coverage probabilities of our bootstrap confidence bands are checked.

Binary response data are generated from

$$E(Y|X=x,T=t) = P(Y=1|X=x,T=t) = G\{\beta^T x + m^+(t)\}, \tag{24}$$

where G is the logit distribution function and $m^+(t) = \alpha + \sum_{j=1}^2 m_j(t_j)$. The explanatory variables X_1, X_2, T_1, and T_2 are independent, X_1 and X_2 are standard normal, and T_1 and T_2 have a uniform distribution on $[-2,2]$. We generate $n = 250$ data points with $\beta = (0.3,-0.7)^T$, $m_1(t_1) = 2\sin(-2t_1)$, $m_2(t_2) = t_2^2 - E[T_2^2]$, and $\alpha = 0$. For all computations the quartic kernel is used. In this section h_1 denotes the bandwidth that is used for the estimation of β. In the simulations we set all weight functions w_{-1}, w_0, and w_1 equal to 1; i.e., we applied no trimming and no optimal weighting.

First, we consider the test problem (14) $H_0: m_1(t_1)$ *is linear*. It can be seen from Figure 1 that the normal approximation of Lemma A3.1 is quite inaccu-

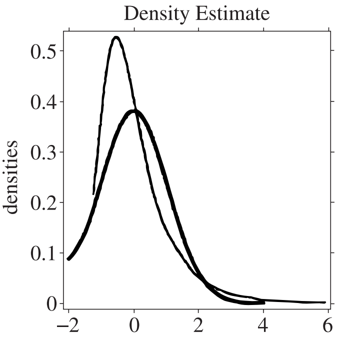

FIGURE 1. Standardized density estimate of the test statistic (thin line) and convoluted standard normal density (thick line).

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

rate. In this plot a density estimate for the test statistic R, based on 500 Monte Carlo replications, is plotted together with its limiting normal density. The parameters are chosen on the null hypothesis, with $m_1(t_1) = t_1$ and β, m_2, and α as before. The density estimate for R is a kernel estimate with bandwidth according to Silverman's rule of thumb, i.e., $1.06 \cdot 2.62 \cdot n^{-1/5}$ times the empirical standard deviation. For better comparison, the normal density is convoluted with the quartic kernel using the same bandwidth.

In a simulation (500 runs) the level of the bootstrap test is estimated for $B = 249$ bootstrap repetitions. We get a relative number of rejections of 0.03 for theoretical level 0.05 and 0.06 for theoretical level 0.1; i.e., the bootstrap test keeps its level but is conservative for such a small sample. The power is investigated for the alternatives $m_1(t_1) = (1 - v)t_1 + v\{2\sin(-2t_1)\}$, $0 \le v \le 1$. The other parameters are chosen as before. For comparison, we perform the same simulations for a parametric likelihood ratio test testing the hypothesis $\gamma_1 = \gamma_2 = 0$ in the parametric model

$$P(Y = 1|X = x, T = t) = G[\beta^T x + (1 - \gamma_1)t_1 + \gamma_2\{2\sin(-2t_1)\} + \gamma_3 m_2(t_2) + \gamma_4].$$

Clearly, this comparison is far away from being fair because for the parametric test the alternative and also m_2 are assumed to be known. Figure 2 plots the power of these tests at theoretical levels 0.05 and 0.1. Note that the better performance of the parametric test is partly due just to the fact that the test R is conservative (see the preceding discussion). (One could compare the power of R in the right plot with the power of the likelihood ratio test in the left plot.) We conclude that the bootstrap test performs quite well.

Second, for bootstrap confidence bands we investigate the following questions: What is the coverage accuracy in a small sample? How much does the width of the band vary with the chosen coverage probability? Does it really matter how we estimate $\sigma_1^2(t_1)$? In the simulations we use two estimates of $\sigma_1^2(t_1)$: $\hat{\sigma}_1^2(t_1)$ as defined in equation (A.2) (see Section 3.5) and the empirical variance

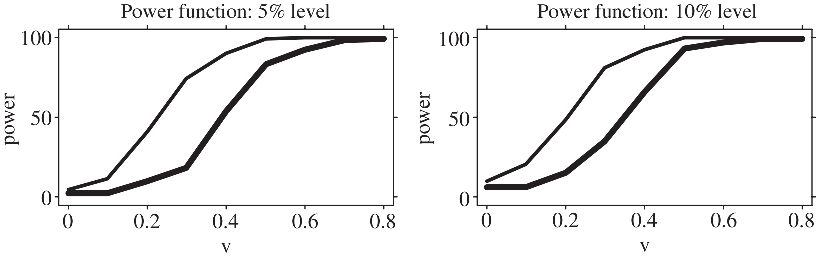

FIGURE 2. Power functions for theoretical levels 0.05 (left) and 0.1 (right), for the nonparametric bootstrap test (thick line) and the likelihood ratio test (thin line).

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

TABLE 1. Coverage probabilities for bootstrap confidence bands with $h_1 = h = g = 0.5$.

Theoretical coverage	95%	90%	85%	80%
Using $\hat{\sigma}_1^2(t_1)$	0.963	0.912	0.846	0.776
Using $\tilde{\sigma}_1^2(t_1)$	0.948	0.904	0.839	0.776

of $m_1^*(t_1)$ in the bootstrap resamples, denoted by $\tilde{\sigma}_1^2(t_1)$. The simulated model is again (24) with n, m_1, m_2, X_1, and X_2 as before. But the variables T are now uniformly distributed on $[-1,1]$. The confidence bands are only investigated for m_1. For $h_1 = h = g = 0.3$ to 0.6 we obtain reasonable coverage accuracies; results for $h_1 = h = g = 0.5$ are given in Table 1. The empirical coverage probabilities are close to the theoretical ones for all levels and for both variance estimates. It is surprising how well the bootstrap fits the different coverage probabilities in such small samples. For smaller and larger bandwidths they are less accurate. This is caused by poorer bootstrap bias correction. In contrast, the variance of the estimates is always well caught by the bootstrap. In Figures 3 and 4 we compare 95% and 85% confidence bands. Despite their different levels the bands hardly differ.

In our last simulation, we verify the performance of the test for the link function (see Section 3.4). The data are generated as in the simulations on confidence bands. Bandwidth h_L (see (23)) is set to $0.4 \cdot \hat{s}_I$, where $\hat{s}_I$ is an estimate of the standard deviation s_I of the index; otherwise we set $h_1 = h = g = 0.35$. The simulation results for level accuracy for the theoretical 1, 5, 10, and 15% levels are 0.014, 0.046, 0.090, and 0.13. Thus the accuracy is quite good. We also tried different bandwidths but found no major differences in the results.

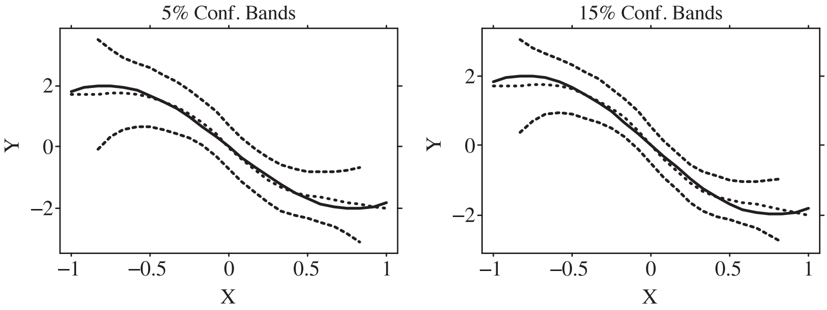

FIGURE 3. 95% and 85% confidence bands, using $\hat{\sigma}$. Dashed lines are the confidence bands and corresponding estimates; solid lines are the data-generating functions.

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

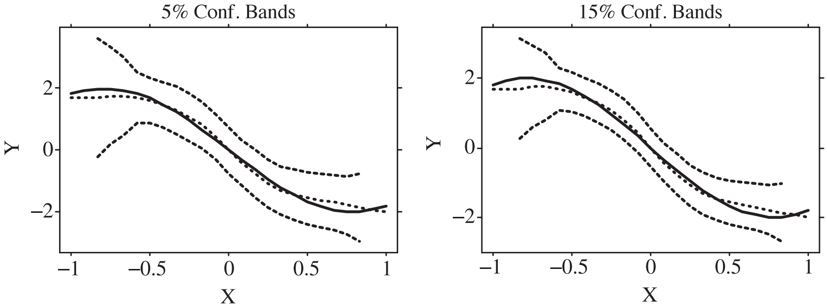

FIGURE 4. 95% and 85% confidence bands, using $\tilde{\sigma}$. Dashed lines are the confidence bands and corresponding estimates; solid lines are the data-generating functions.

REFERENCES

Ai, C. (1997) A semiparametric maximum likelihood estimator. *Econometrica* 65, 933–963.

Beran, R. (1986) Comment on "Jackknife, bootstrap, and other resampling methods in regression analysis" by C.F.J. Wu. *Annals of Statistics* 14, 1295–1298.

Carroll, R.J., J. Fan, I. Gijbels, & M.P. Wand (1997) Generalized partially linear single-index models. *Journal of the American Statistical Association* 92, 477–489.

Deaton, A. & J. Muellbauer (1980) *Economics and Consumer Behavior.* Cambridge University Press.

Eubank, R.L. & P.L. Speckman (1993) Confidence bands in nonparametric regression. *Journal of the American Statistical Association* 88, 1287–1301.

Fan, J., W. Härdle, & E. Mammen (1998) Direct estimation of low dimensional components in additive models. *Annals of Statistics* 26, 943–971.

Goldberger, A.S. (1964) *Econometric Theory.* Wiley.

Gozalo, P.L. & O.B. Linton (2001) Testing additivity in generalized nonparametric regression models. *Journal of Econometrics* 104, 1–48.

Hansen, M.H., J.Z. Huang, C. Kooperberg, C.J. Stone, & Y.K. Truong (2002) *Statistical Modeling with Spline Functions: Methodology and Theory.* Springer-Verlag. In press.

Härdle, W., S. Huet, E. Mammen, & S. Sperlich (1998) Semiparametric additive indices for binary response and generalized additive models. Discussion Paper 95, Sanderforschungsbereich 373, Berlin.

Härdle, W. & E. Mammen (1993) Testing parametric versus nonparametric regression. *Annals of Statistics* 21, 1926–1947.

Härdle, W., E. Mammen, & M. Müller (1998) Testing parametric versus semiparametric modelling in generalized linear models. *Journal of the American Statistical Association* 93, 1461–1474.

Härdle, W., E. Mammen & I. Proenca (2001) A bootstrap test for single index models. *Statistics* 35, 427–452.

Hastie, T.J. & R.J. Tibshirani (1990) *Generalized Additive Models.* Chapman and Hall.

Horowitz, J.L. (1998) *Semiparametric Methods in Econometrics.* Lecture Notes in Statistics 131, Springer-Verlag.

Horowitz, J.L. (2001) Nonparametric estimation of a generalized additive model with an unknown link function. *Econometrica* 69, 499–513.

Leontief, W. (1947) Introduction to a theory of the internal structure of functional relationships. *Econometrica*, 15 361–373.

Linton, O.B. & W. Härdle (1996) Estimating additive regression models with known links. *Biometrika* 83, 529–540.

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

Linton, O.B. & J.P. Nielsen (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82, 93–101.

Mammen, E. (1992) *When Does Bootstrap Work? Asymptotic Results and Simulations*. Lecture Notes in Statistics 77, Springer-Verlag.

Mammen, E. & S. van de Geer (1997) Penalized quasi-likelihood estimation in partial linear models. *Annals of Statistics* 25, 1014–1035.

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

McCullagh, P. & J.A. Nelder (1989) *Generalized Linear Models.* Chapman and Hall.

Neumann, M. & J. Polzehl (1998) Simultaneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics* 9, 307–333.

Opsomer, J.D. & D. Ruppert (1999) A root-*n* consistent estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics* 8, 715–732.

Severini, T.A. & J.G. Staniswalis (1994) Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association* 89, 501–511.

Sperlich, S., D. Tjøstheim, & L. Yang (2002) Nonparametric estimation and testing of interaction in additive models. *Econometric Theory* 18, 197–251.

Stone, C.J. (1985) Additive regression and other nonparametric models. *Annals of Statistics* 13, 685–705.

Tjøstheim, D.J. & B.H. Auestadt (1994) Nonparametric identification of nonlinear time series: Projections. *Journal of the American Statistical Association* 89, 1398–1409.

Wu, C.F.G. (1986) Jackknife, bootstrap, and other resampling methods in regression analysis. (with discussion) *Annals of Statistics* 14, 1291–1380.

APPENDIX

A.1. Assumptions. We now state the assumptions that are used in the results in Sections 2.1 and 2.2 and Section A.3 of this Appendix. We use the notation

$$h_{\max} = \max\{h_1, \dots, h_r, g_1, \dots, g_s\},$$

$$h_{prod} = h_1 \cdot \dots \cdot h_r \, g_1 \cdot \dots \cdot g_s,$$

$$\rho_1 = h_{\max}^2 + (nh_{prod})^{-1/2},$$

$$\rho_2 = h_{\max}^2 + (\log n)^{1/2}(nh_{prod})^{-1/2}.$$

Furthermore, we put $\lambda_i(u) = Q\{G(u);Y_i\}$, $\lambda(u) = Q\{G(u);Y\}$. With this notation we have

$$\lambda_i'(u) = \frac{Y_i - G(u)}{V[G(u)]}\, G'(u),$$

$$\lambda_i''(u) = \{Y_i - G(u)\}\left[\frac{G''(u)}{V[G(u)]} - \frac{V'(G(u))\, G'(u)^2}{V[G(u)]^2}\right] - \frac{G'(u)^2}{V[G(u)]}. \qquad \textbf{(A.1)}$$

For our asymptotic expansions we use the following assumptions.

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

(A1) $(X_1, T_1, Y_1), \dots, (X_n, T_n, Y_n)$ are i.i.d. tuples. The expression $T_i = (T_{i,1}, \dots, T_{i,d})$ is a vector with components $T_{i,j}$ in $\mathbb{R}^{q_j}$, X_i is $\mathbb{R}^p$ valued, and Y_i is $\mathbb{R}$ valued. We write $r = q_1$ and $s = q_2 + \cdots + q_d$.

(A2) $E(Y|X,T) = G\{X^T\beta + m^+(T)\}$ with $\beta \in \mathbb{R}^p$. Here m^+ denotes the function $m^+(t) = \alpha + m_1(t_1) + \cdots + m_d(t_d)$, with $E\, m_j(T_{i,j}) = 0$ for $j = 1, \dots, d$. The conditional variance $\operatorname{Var}(Y_i|T_i = t)$ has a bounded second derivative. Furthermore the Laplace transform $E \exp t|Y_i|$ is finite for $t > 0$ small enough.

(A3) X_i and T_i have compact support S_X, S_T. The support S_T is of the form $S_{T,1} \times S_{T,-1}$ with $S_{T,1} \subset \mathbb{R}^r$ and $S_{T,-1} \subset \mathbb{R}^s$. Here T has a twice continuously differentiable density f_T with $\inf_{t \in S_T} f_T(t) > 0$.

(A4) For compact sets $B \subset \mathbb{R}^p$ and $H \subset \mathbb{R}$ we define

$$\hat{\beta} = \arg\max_{\beta \in B} \mathcal{L}(\hat{m}_\beta, \beta),$$

where, as before,

$$\mathcal{L}(\eta, \beta) = \sum_{i=1}^{n} Q\{G(X_i^T\beta + \eta(T_i)); Y_i\}.$$

The term $\hat{m}_\beta(t)$ is defined as

$$\hat{m}_\beta(t) = \arg\max_{\eta \in H} \sum_{i=1}^{n} K_h(t_1 - T_{i,1}) L_g(t_{-1} - T_{i,-1}) Q[G\{X_i^T\beta + \eta\}; Y_i].$$

For $\beta \in B$ we put

$$m_\beta(t) = \arg\max_{\eta \in H} E[\lambda(X^T\beta + \eta)|T = t].$$

We assume that $m_\beta(t)$ lies in the interior of H for all $t \in S_T$ and $\beta \in B$. This implies $E\{\lambda'(\beta^T X + m_\beta(t))|T = t\} = 0$. We assume also that $E[\lambda''\{\beta^T X + m_\beta(T)\}|T = t] \neq 0$ for all $t \in S_T$ and $\beta \in B$ and that for all $\varepsilon > 0$ there exists a $\delta > 0$ such that for all $\eta \in H, t \in S_T, \beta \in B$

$$|E[\lambda'(X^T\beta + \eta)|T = t]| \le \delta$$

implies that

$$|\eta - m_\beta(t)| \le \varepsilon.$$

(A5) There exists a $\delta > 0$ such that $G^{(k)}(u)$, $k = 1, \dots, 3$, and $G'(u)^{-1}$ are bounded on $u \in S^+ = \{x^T b + \eta + \kappa : x \in S_X, b \in B \text{ and } \eta \in H, \kappa \in \mathbb{R} \text{ with } |\kappa| \le \delta\}$. Furthermore V^{-1}, V', and V'' are bounded on $G(S^\delta)$.

(A6) $m_1, \dots, m_d$ are twice continuously differentiable functions from $\mathbb{R}^{q_j}$ to $\mathbb{R}$. The weight functions w, w_{-1}, and w_1 are positive and twice continuously differentiable. To avoid problems on the boundary, we assume that for a $\delta > 0$ we have that $w_{-1}(t) = 0$, $w_1(t) = 0$, and $w(t) = 0$ for $t \in S_{T,-1}^- = \{s : \text{there exists a } u \notin S_{T,-1} \text{ with } \|s - u\| \le \delta\}$, $t \in S_{T,1}^- = \{s : \text{there exists a } u \notin S_{T,1} \text{ with } \|s - u\| \le \delta\}$, and $t \in S_T^- = \{s : \text{there exists a}$

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

$u \notin S_T$ with $\|s - u\| \leq \delta\}$, respectively. Furthermore, the weight function w_1 is such that $\int_{S_{T,1}} w_1(t_1) m_1(t_1) f_{T_1}(t_1)\, dt_1 = 0$, where f_{T_1} denotes the density of T_1.

(A7) The kernels K and L are product kernels $K(v) = K_1(v_1) \cdot \dots \cdot K_r(v_r)$ and $L(v) = L_1(v_1) \cdot \dots \cdot L_s(v_s)$. The kernels K_i and L_j are symmetric probability densities with compact support ($[-1,1]$, say).

(A8) $E[\lambda_1''\{X_1^T\beta_0 + m^+(T_1)\}|T_1 = t]$ and $E[\lambda_1''\{X_1^T\beta_0 + m^+(T_1)\}X_1|T_1 = t]$ are twice continuously differentiable functions for $t \in S_T$.

(A9) The matrix $E\, Z^2 \widetilde{X}\widetilde{X}^T$ is strictly positive definite. The random vectors Z and $\widetilde{X}$ are defined in Lemmas A2.1 and A2.2 in Section A.2 of this Appendix, respectively.

This assumption implies that X does not contain an intercept. Note that if the first element of X were constant, a.s., e.g., $X_{i1} \equiv 1$, then $\widetilde{X}_{i1} \equiv 0$.

(A10) $m_1, \dots, m_d$ are four times continuously differentiable on $\mathbb{R}$.

(A11) The kernels K_i and L_j are twice continuously differentiable.

Assumptions (A1)–(A3) and (A5) and (A6) contain boundedness conditions on covariates and standard smoothness conditions on regression functions, design densities, link function, and variance function. Condition (A4) contains a slightly modified definition of our estimates. We now assume that in the definition of the parametric and nonparametric estimates the minimization of the QL only runs over a bounded set (denoted by B or H, respectively). This assumption together with (A8) and the other assumptions of (A4) enables us to prove consistency of the parametric and nonparametric estimates and to derive a stochastic expansion of these estimates. Condition (A7) is a standard assumption on the kernels K and L. Condition (A8) guarantees that the Fisher information of the parametric estimate is positive definite. Conditions (A10) and (A11) are used for second-order bounds on expansions of bias terms.

A.2. Asymptotic Theory for Estimation. This section contains asymptotic results on the marginal integration estimates $\hat{m}_j$ and the parametric estimate $\hat{\beta}$.

LEMMA A2.1. *Suppose that Assumptions (A1)–(A9) apply. If the elements of h and g tend to zero and $nh_1 \cdot \dots \cdot h_r g_1^2 \cdot \dots \cdot g_s^2 (\log n)^{-2}$ tends to infinity, then*

$$\sqrt{nh}\{\hat{m}_1(t_1) - m_1(t_1) - \delta_n^1(t_1)\}$$

converges to a centered Gaussian variable with variance

$$\sigma_1^2(t_1) = \int K^2(u)\, du \frac{f_1(t_1)}{\{Ew_{-1}(T_{-1})\}^2} E\left[\frac{Z_1}{Z_2} \middle| T_1 = t_1 \right],$$

where $f_{T_{-1}}$ and f_T are the densities of T_{-1} or $T = (T_1, T_{-1})$, respectively. (For a vector $(v_1, \dots, v_d)$ with $v_j \in \mathbb{R}^{q_j}$ we denote the vector $(v_1, \dots, v_{j-1}, v_{j+1}, \dots, v_d)$ by v_{-j}.) The terms Z_1 and Z_2 are defined in the following way:

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

$$Z_1 = w_1^2(T_{-1}) \frac{Z^2}{V[G\{X^T\beta + m^+(T)\}]} f_{T_{-1}}^2(T_{-1}) \mathrm{Var}(Y|X,T),$$

$$Z_2 = E[Z^2|T_1 = t_1, T_{-1}]^2 f_T^2(t_1, T_{-1}),$$

$$Z^2 = \frac{G'(X^T\beta + m^+(T))^2}{V[G\{X^T\beta + m^+(T)\}]}.$$

For the asymptotic bias $\delta_n^1(t_1)$, one has

$$\delta_n^1(t_1) = d_n^1(t_1) - \int d_n^1(v_1) w_1(v_1) f_{T_1}(v_1)\, dv_1 \Big/ \int w_1(v_1) f_{T_1}(v_1)\, dv_1 + o_P(h_+^2 + g_+^2),$$

where

$$d_n^1(t_1) = g_+^2 \int_{\mathbb{R}^{d-1}} E\left[a^1(X, t_1, u) \sum_{j=2}^{d} \sigma_{L,j}^2 b_j(X, t_1, u) \,\middle|\, T = (t_1, u) \right] f_{T_{-1}}(u)\, du$$
$$+ h_+^2 \int_{\mathbb{R}^{d-1}} E[a^1(X, t_1, u)\, \sigma_K^2 b_1(X, t_1, u) | T = (t_1, u)]\, f_{T_{-1}}(u)\, du.$$

Here f_{T_1} denotes the density of T_1. We write $f'_{Tj}(v) = (\partial/\partial v_j) f_T(v)$. Furthermore, $\sigma_{L,j}^2 = \int s^2\, dL_j$, $\sigma_K^2 = \int s^2\, dK$, and

$$a^1(x,v) = \frac{w_{-1}(v_{-1}) G'(x^T\beta + m^+(v))}{E[w_{-1}(T_{-1})]\, E[Z^2|T = v]\, f_T(v) V[G(x^T\beta + m^+(v))]},$$

$$b_j(x,v) = \frac{1}{2}[G''(x^T\beta + m^+(v)) m_j'(v_j)^T H_j^2 m_j'(v_j)$$
$$+ G'(x^T\beta + m^+(v)) \mathrm{trace}[m_j''(v_j) H_j^2]]$$
$$\times f_T(v) + G'(x^T\beta + m^+(v)) m_j'(v_j)^T H_j^2 f'_{Tj}(v),$$

where H_1 is a diagonal matrix with diagonal elements

$$h_1/h_+, \dots, h_{q_1}/h_+$$

and where for $j = 2, \dots, d$ the matrix H_j is a diagonal matrix with diagonal elements

$$g_{q_2+\cdots+q_{j-1}}/g_+, \dots, g_{q_2+\cdots+q_j}/g_+.$$

Under the additional assumption of (A10) the rest term $o_P(h_+^2 + g_+^2)$ in the expansion of $\delta_n^1(t_1)$ can be replaced by $O_P(h_+^4 + g_+^4)$.

The estimation of the other additive components m_j for $j = 2, \dots, d$ can be done in the same way as the estimation of m_1 in Lemma A2.1. If assumptions analogous to (A1)–(A10) hold for the other components, then the corresponding limit theorems apply for their estimates. (In the assumptions h always denotes the bandwidth of the estimated component, and g is chosen as bandwidth of the other components.) Then under these conditions the estimates $\hat{m}_1(t_1),\ldots,\hat{m}_d(t_d)$ are asymptotically independent. This leads to a multidimensional result. The random vector

$$\sqrt{nh}\begin{pmatrix}\hat{m}_1(t_1)-m_1(t_1)-\delta_n^1(t_1)\\ \vdots \\ \hat{m}_d(t_d)-m_d(t_d)-\delta_n^d(t_d)\end{pmatrix}\xrightarrow[n\to\infty]{D} N\left(0;\begin{bmatrix}\sigma_1(t_1) & 0 & \ldots & 0\\ \vdots & & \ddots & \vdots\\ 0 & \ldots & 0 & \sigma_d(t_d)\end{bmatrix}\right).$$

The variance $\sigma_1^2(t_1)$ of $(\hat{m}_1(t_1)-E\hat{m}_1(t_1))$ can be estimated by

$$\hat{\sigma}_1^2(t_1)=\sum_{i=1}^n\hat{\tau}_i^2, \tag{A.2}$$

where

$$\hat{\tau}_i=\left[\sum_{j=1}^n w_{-1}(T_{j,-1})\right]^{-1}\frac{1}{n}\sum_{j=1}^n w_{-1}(T_{j,-1})\kappa_i(t_1,T_{j,-1})$$
$$\times\left[\frac{1}{n}\sum_{l=1}^n\frac{G'(X_l^T\hat{\beta}+\hat{m}^+(T_l))^2}{V[G\{X_l^T\hat{\beta}+\hat{m}^+(T_l)\}]}\kappa_l(t_1,T_{j,-1})\right]^{-1}\frac{G'(X_i^T\hat{\beta}+\hat{m}^+(t_1,T_{j,-1}))}{V[G\{X_i^T\hat{\beta}+\hat{m}^+(t_1,T_{j,-1})\}]}\hat{s}_i,$$

$$\kappa_i(t)=\frac{K_h(t_1-T_{i,1})L_g(t_{-1}-T_{i,-1})}{\frac{1}{n}\sum_{j=1}^n K_h(t_1-T_{j,1})L_g(t_{-1}-T_{j,-1})},$$

$$\hat{s}_i^2=\begin{cases}[Y_i-\hat{\mu}_i]^2 & \text{in the case of Model A,}\\ \hat{s}^2V(\hat{\mu}_i) & \text{in the case of Model B,}\\ V(\hat{\mu}_i) & \text{in the case of Model C}\end{cases} \tag{A.3}$$

with

$$\hat{s}^2=\frac{1}{n}\sum_{i=1}^n\frac{[Y_i-\hat{\mu}_i]^2}{V(\hat{\mu}_i)}\quad\text{and}\quad\hat{\mu}_i=G\{X_i^T\hat{\beta}+\hat{\alpha}+\hat{m}_1(T_{i,1})+\cdots+\hat{m}_d(T_{i,d})\}.$$

The estimation of the nonparametric components also yields an estimate of the parameter β. We show that under certain conditions a rate of order $O_P(n^{-1/2})$ can be achieved. This is a consequence of the iterative application of smoothed local and unsmoothed global likelihood function in the definition of $\hat{\beta}$. Our conditions imply that $s+r\le 3$. This constraint can be weakened by assumption of higher order smoothness of $m_1,\ldots,m_d$ and by the use of higher order kernels.

LEMMA A2.2. *Suppose that Assumptions (A1)–(A9) apply. Then, if* $hg^{d-1}\times n^{1/2}(\log n)^{-1}$ *tends to infinity and* h *and* $g=o(n^{-1/8})$, *it holds that*

$$n^{1/2}\{\hat{\beta}-\beta\}\xrightarrow[n\to\infty]{D} N(0;I^{-1}),$$

where Z^2 is defined as in Lemma A2.1 and

$$I = EZ^2\widetilde{X}\widetilde{X}^T \quad \textit{with } \widetilde{X} = X - \{E(Z^2|T)\}^{-1}E(Z^2X|T).$$

Our estimate of β achieves the efficiency bound in the partial linear model $m(x;t) = G\{x^T\beta + \alpha + m(T_1,\dots,T_d)\}$ (see Mammen and van de Geer, 1997). An estimate that takes care of additivity is given by

$$\hat{\beta} = \arg\max_{\beta\in B} \mathcal{L}(\hat{m}_\beta^+,\beta),$$

where $\hat{m}_\beta^+(t)$ is defined as $\hat{m}^+(t)$ with $\hat{m}$ replaced by $\hat{m}_\beta$ in equation (8). We expect that this estimate achieves higher efficiency. However this estimate has two drawbacks. Calculation of this estimate would need several nested iterative algorithms and is therefore computationally unattractive for large data sets. Moreover, such an estimator is not robust against deviations from additivity.

Compared to $\hat{\beta}$ root-n consistency of $\hat{\alpha}$ requires additional conditions. The estimate $\hat{\alpha}$ inherits by construction the biases of the nonparametric estimates $\hat{m}, \hat{m}_1,\dots,\hat{m}_d$. These biases are only of order $o(n^{-1/2})$ if the elements of h and g are of order $o(n^{-1/4})$. Note that this is not necessary for $\hat{\beta}$. On the other hand it can be checked that $\hat{\alpha}$ has, as does $\hat{\beta}$, asymptotic variance of order $O(n^{-1})$. Clearly, this is not essential as for most applications the parameter α has no direct interpretation.

A.3. Proofs. For simplicity of notation we give all proofs only for the case $q_1 = \cdots = q_d = 1$. Then $r = 1$ and $s = d - 1$. Furthermore we suppose that $g_1 = \cdots = g_{d-1}$ and denote this bandwidth by g. The bandwidth h_1 is denoted by h.

Proof of Lemma A2.1. We start by showing consistency of the estimate $\hat{\beta}$:

$$\hat{\beta} = \beta_0 + o_P(1). \tag{A.4}$$

For the proof of (A.4) we show first that

$$\sup_{t,\beta}|\hat{m}_\beta(t) - m_\beta(t)| = o_p(1). \tag{A.5}$$

Proof of (A.5). For the proof of claim (A.5) we show first that

$$\sup_{\eta,t,\beta}|\Delta(m_\beta(t),t,\beta)| = O_p(\rho_2), \tag{A.6}$$

where the following notation has been used:

$$\Delta(\eta,t,\beta) = \Delta_1(\eta,t,\beta) - \Delta_2(\eta,t,\beta),$$

$$\Delta_1(\eta,t,\beta) = \frac{1}{n}\sum_i \lambda_i'(X_i^T\beta + \eta)\kappa_i(t),$$

$$\Delta_2(\eta,t,\beta) = E[\lambda'(X^T\beta + \eta)|T = t],$$

$$\kappa_i(t) = \frac{K_h(t_1 - T_{i,1})L_g(t_{-1} - T_{i,-1})}{\frac{1}{n}\sum_{j=1}^n K_h(t_1 - T_{j,1})L_g(t_{-1} - T_{j,-1})}. \tag{A.7}$$

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

For the proof of (A.6) we remark first that

$$E\Delta(\eta,t,\beta) = O(h^2+g^2).$$

This can be seen by standard smoothing arguments. Furthermore, $\Delta_1(\eta,t,\beta)$ is a sum of i.i.d. random variables with bounded Laplace transform (see Assumption (A2)). By standard application of exponential inequalities we get for every $\nu_1 > 0$ that for C' large enough

$$P\{|\Delta(\eta,t,\beta)| > C'\rho_2\} = o(n^{-\nu_1}). \tag{A.8}$$

We consider the partial derivatives of the summands of $\Delta(\eta,t,\beta)$ with respect to η, t, and β. They are bounded by $C''n^{\nu_2}$ for C'' and ν_2 large enough. Together with (A.8), following the same argument as in Härdle and Mammen (1993), we obtain (A.6).

For the proof of (A.5), one can conclude from (A.6) that, with probability tending to one, $\hat m_\beta(t)$ lies in the interior of H (see Assumption (A4)). This gives

$$\Delta_1(\hat m_\beta(t),t,\beta) = 0. \tag{A.9}$$

With (A.6) we obtain

$$\sup_{t,\beta}|\Delta_2(\hat m_\beta(t),t,\beta)| = O_p(\rho_2).$$

With Assumption (A4) this yields (A.5). ■

We use (A.5) to prove (A.4) (consistency of $\hat\beta$).

Proof of (A.4). Let $k(\beta) = E[Q\{X^T\beta + m_\beta(T);Y\}]$. We will show that

$$\sup_{\beta\in B}\left|\frac{1}{n}\mathcal{L}(\hat m_\beta,\beta) - k(\beta)\right| \to 0 \quad \text{(in probability).} \tag{A.10}$$

This implies claim (A.4) because

$$\begin{aligned} k''(\beta_0) &= E\left[\lambda''\{X^T\beta_0 + m^+(T)\}\left\{X + \frac{\partial m_\beta}{\partial\beta}(\beta_0,T)\right\}\left\{X + \frac{\partial m_\beta}{\partial\beta}(\beta_0,T)\right\}^T\right] \\ &= -E(Z^2\widetilde{X}\widetilde{X}^T) \end{aligned}$$

is strictly negative definite and $k(\beta_0) = \sup_{\beta\in H} k(\beta)$.

It remains to prove (A.10). This follows from

$$\sup_{\beta\in B}\left|\frac{1}{n}\mathcal{L}(m_\beta,\beta) - k(\beta)\right| \to 0 \quad \text{(in probability),} \tag{A.11}$$

$$\sup_{\beta\in B}\left|\frac{1}{n}\mathcal{L}(\hat m_\beta,\beta) - \frac{1}{n}\mathcal{L}(m_\beta,\beta)\right| \to 0 \quad \text{(in probability).} \tag{A.12}$$

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

Claim (A.11) holds because $\mathcal{L}(m_\beta,\beta)/n$ converges to $k(\beta)$ by the law of large numbers and because $\{\mathcal{L}(m_\beta,\beta)/n, \beta \in B\}$ is tight. For the proof of tightness note first that

$$\left|\frac{1}{n}\,\mathcal{L}(m_{\beta_1},\beta_1) - \frac{1}{n}\,\mathcal{L}(m_{\beta_2},\beta_2)\right| \le T_{n,1}\|\beta_1-\beta_2\| + T_{n,2}\sup_t|m_{\beta_1}(t) - m_{\beta_2}(t)|$$

$$\le T_{n,1}\|\beta_1-\beta_2\| + T_{n,2}\sup_{t,\beta}\left\|\frac{\partial}{\partial\beta}\,m_\beta(t)\right\|\|\beta_1-\beta_2\|,$$

where

$$T_{n,1} = \sup_{\beta,\eta}\frac{1}{n}\sum_{i=1}^{n}\lambda'(X_i^T\beta+\eta)\|X_i\|,$$

$$T_{n,2} = \sup_{\beta,\eta}\frac{1}{n}\sum_{i=1}^{n}\lambda'(X_i^T\beta+\eta).$$

Under our conditions, $T_{n,1}$ and $T_{n,2}$ are bounded in probability. To see that $(\partial/\partial\beta)m_\beta(t)$ is uniformly bounded in β and t note that

$$\frac{\partial m_\beta}{\partial\beta}(\beta,t) = -\frac{E[\lambda''\{\beta^TX + m_\beta(T)\}X|T=t]}{E[\lambda''\{\beta^TX + m_\beta(t)\}|T=t]}. \tag{A.13}$$

Equation (A.13) follows by differentiation of $E\{\lambda'(\beta^TX + m_\beta(t))|T=t\} = 0$. This shows (A.11). Claim (A.12) follows from

$$\sup_\beta\left|\frac{1}{n}\,\mathcal{L}(\hat{m}_\beta,\beta) - \frac{1}{n}\,\mathcal{L}(m_\beta,\beta)\right| \le \sup_{\beta,\eta}|\lambda'(X^T\beta+\eta)|\sup_{t,\beta}|\hat{m}_\beta(t) - m_\beta(t)|.$$

Thus finally (A.4) is shown. ■

Next, we establish uniform stochastic expansions of $\hat\beta$ and $\hat{m}(t)$.

$$\hat\beta = \beta + \{E(Z^2\widetilde{X}\widetilde{X}^T)\}^{-1}\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i\,\lambda_i'\{X_i^T\beta + m^+(T_i)\} + O_p(\rho_2^2), \tag{A.14}$$

$$\sup_{t\in S_T^*}|\Delta(t)| = O_p(\rho_2^2), \tag{A.15}$$

with

$$\Delta(t) = \hat{m}(t) - \Bigg\{\overline{m}(t) + \{E(Z^2|T=t)\}^{-1}E(Z^2X^T|T=t)\{E(Z^2\widetilde{X}\widetilde{X}^T)\}^{-1}$$

$$\times\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i\,\lambda_i'\{X_i^T\beta + m^+(T_i)\}\Bigg\}, \tag{A.16}$$

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

$$\overline{m}(t) = m^{+}(t) + \{E(Z^2|T=t)\}^{-1}\,\frac{1}{n}\sum_{i=1}^{n}\kappa_i(t)\lambda_i'\{X_i^T\beta + m^{+}(t)\}, \tag{A.17}$$

$$S_T^{*} = \{t \in S_T : t+\eta \in S_T \text{ for all } \eta \text{ with } |\eta_1| \le g \text{ and } |\eta_j| \le h\ (j=2,\ldots,d)\},$$

$$\widetilde{X}_i = X_i - \{E[Z_i^2|T_i]\}^{-1}E[Z_i^2X_i|T_i], \tag{A.18}$$

$$Z_i^2 = \frac{G'(X_i^T\beta + m^{+}(T_i))^2}{V[G(X_i^T\beta + m^{+}(T_i))]}. \tag{A.19}$$

Equations (A.14) and (A.15) follow from a slight modification of Lemma A3.3 and Corollary A3.4 in Härdle et al. (1998). There it has been assumed that the likelihood is maximized for β in a neighborhood of β_0 with radius ρ_1 (see Härdle et al., 1998, Assumption (A7)). In our setup we have that for a sequence δ_n' with $\delta_n' \to 0$ with probability tending to one

$$\hat{\beta} = \arg\max_{\beta:\|\beta-\beta_0\|\le\delta_n'} \mathcal{L}(\hat{m}_\beta, \beta).$$

Using the same arguments as in Härdle et al. (1998), one can show that

$$\hat{\beta} = \beta + \{E(Z^2\widetilde{X}\widetilde{X}^T)\}^{-1}\,\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i\,\lambda_i'\{X_i^T\beta + m^{+}(T_i)\} + O_p(\rho_2^2) + \|\hat{\beta}-\beta\|^2 O_p(1).$$

This shows (A.14). Equation (A.15) can be shown similarly.

With the help of (A.15) we arrive at

$$\begin{aligned}\overline{m}_1(t_1) &= \frac{\sum_{i=1}^{n} w_{-1}(T_{i,-1})\overline{m}(t_1, T_{i,-1})}{\sum_{i=1}^{n} w_{-1}(T_{i,-1})} + O_P(\rho_2^2 + n^{-1/2})\\ &= m_1(t_1) + R_1 + \Delta_1(t_1) + O_P(\rho_2^2 + n^{-1/2}),\end{aligned} \tag{A.20}$$

where

$$R_1 = \frac{1}{\sum_{i=1}^{n} w_{-1}(T_{i,-1})}\sum_{i=1}^{n} w_{-1}(T_{i,-1})[m_2(T_{i,2}) + \cdots + m_d(T_{i,d})],$$

$$\Delta_1(t_1) = \frac{1}{\sum_{i=1}^{n} w_{-1}(T_{i,-1})}\,\frac{1}{n}\sum_{i,j=1}^{n}\frac{w_{-1}(T_{i,-1})\kappa_j(t_1, T_{i,-1})}{E(Z_i^2|T_{i,1}=t_1, T_{i,-1})}\,\lambda_j'\{X_j^T\beta + m^{+}(t_1, T_{i,-1})\},$$

where λ_j', κ_j, and Z_i are defined by equations (A.1), (A.3), and (A.19), respectively. Given $\mathcal{Z}_n = ((X_1, T_{1,1}, \ldots, T_{1,d}), \ldots, (X_n, T_{n,1}, \ldots, T_{n,d}))$, the term $\Delta_1(t_1)$ is a sum of independent variables. For the conditional variance the following convergence holds in probability:

$$nh\,\mathrm{Var}(\Delta_1(t_1)|\mathcal{Z}_n)$$
$$\to \int L^2(u)\,du\,E\left[\frac{w^2(T_{-1})}{\{Ew_{-1}(T_{-1})\}^2}\,\frac{E(Z^2|T_1=t_1)}{E(Z^2|T_1=t_1,T_{-1})^2}\,\frac{f^2_{T_{-1}}(T_{-1})}{f^2_T(t_1,T_{-1})}\right].$$

For this convergence, one uses, e.g.,

$$\left|\sup_{t=(t_1,t_{-1})\in S_T^-} n^{-1}\sum_{k=1}^n K_h(t_1-T_{1,k})L_g(t_{-1}-T_{-1,k}) - f_T(t_1,t_{-1})\right| = o_P(1),$$

$$n^{-1}\sum_{k=1}^n K_h(t_1-T_{1,k}) - f_{T_1}(t_1) = o_P(1).$$

Asymptotic normality of $\Delta_1(t_1) - E(\Delta_1(t_1)|\mathcal{Z}_n)$ follows from the convergence of the conditional variance and from

$$P(d_K(\mathcal{L}(\Delta_1(t_1)-E(\Delta_1(t_1)|\mathcal{Z}_n)),N(0,\mathrm{Var}(\Delta_1(t_1)|\mathcal{Z}_n)))>\delta)\to 0 \tag{A.21}$$

for all $\delta > 0$. Here d_K is the Kolmogorov distance, which is for two probability measures μ and ν (on the real line) defined as

$$d_K(\mu,\nu) = \sup_{t\in I\!R}|\mu(X\le t)-\nu(X\le t)|.$$

For the proof of (A.21) one shows that a conditional Lindeberg condition holds with probability tending to one. It remains to study the conditional expectation $E(\Delta_1(t_1)|\mathcal{Z}_n)$. This can be done by showing first that

$$\begin{aligned}E(\Delta_1(t_1)|\mathcal{Z}_n) = {}&\frac{1}{n}\sum_{i=1}^n\int K_h(t_1-v_1)L_g(T_{i,-1}-v_{-1})\\ &\times E[\{G(X^T\beta+m^+(v))-G(X^T\beta+m^+(t_1,T_{i,-1}))\}\\ &\times a^1(X,t_1,T_{i,-1})|T_{i,1}=t_1,T_{i,-1}]\,f_T(v)\,dv + r_n,\end{aligned} \tag{A.22}$$

where the function a^1 is defined in Lemma A2.1 and $r_n = O_P(\rho_2^2 + n^{-1/2}) + o_P(h^2 + g^2)$. Furthermore, $r_n = O_P(\rho_2^2 + n^{-1/2} + h^4 + g^4)$ under the additional assumption (A10). The proof of (A.22) follows by standard but tedious calculations. The asymptotic form of $E(\Delta_1(t_1)|\mathcal{Z}_n)$ can be easily calculated from (A.22). Note that the asymptotic bias of $\hat{m}_1(t_1)$ is asymptotically equal to

$$E(\Delta_1(t_1)|\mathcal{Z}_n) - \int E(\Delta_1(v_1)|\mathcal{Z}_n)w_1(v_1)f_{T_1}(v_1)\,dv_1 \Big/ \int w_1(v_1)f_{T_1}(v_1)\,dv_1$$

because we assumed that $\int w_1(v_1)m_1(v_1)f_{T_1}(v_1)\,dv_1 = 0$. Furthermore, note that up to first order, $\hat{m}_1(t_1)$ and $\tilde{m}_1(t_1)$ have the same asymptotic variance. ■

Proof of Lemma A2.2. The conditions on h and g imply $\rho_2^2 = o(n^{-1/2})$. Therefore the statement of Lemma A2.2 can be followed from (A.14). ■

Proof of Theorem 3.1. The statement of the theorem follows from

$$2\hat{m}_1(t_1) - E^*\hat{m}_1^*(t_1) - m_1(t_1) = O_P(h^4 + g^4 + (nh)^{-1/2}). \tag{A.23}$$

Claim (A.23) follows from

$$2\overline{m}_1(t_1) - E^*\,\overline{m}_1^*(t_1) - m_1(t_1) = R_1 - \hat{R}_1 + O_P(h^4 + g^4 + (nh)^{-1/2}), \tag{A.24}$$

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n w_1(T_{i,1})[2\overline{m}_1(T_{i,1}) - E^*\,\overline{m}_1^*(T_{i,1}) - m_1(T_{i,1})]\\ &\qquad = [R_1 - \hat{R}_1]\,\frac{1}{n}\sum_{i=1}^n w_1(T_{i,1}) + O_P(h^4 + g^4 + (nh)^{-1/2}),\end{aligned} \tag{A.25}$$

where

$$\hat{R}_1 = \frac{1}{\sum_{i=1}^n w_{-1}(T_{i,-1})}\sum_{i=1}^n w_{-1}(T_{i,-1})[\hat{m}_2(T_{i,2}) + \cdots + \hat{m}_d(T_{i,d})]$$

and where R_1 has been defined after (A.20).

We give only the proof of (A.24). Claim (A.25) follows similarly. By (A.20) we have that

$$\overline{m}_1(t_1) = m_1(t_1) + R_1 + D_1(t_1) + O_P(h^4 + g^4 + (nh)^{-1/2}),$$

where

$$\begin{aligned}D_1(t_1) = {}& \frac{1}{\sum_{i=1}^n w_{-1}(T_{i,-1})}\,\frac{1}{n}\sum_{i,j=1}^n \frac{w_{-1}(T_{i,-1})\kappa_j(t_1, T_{i,-1})}{E(Z_i^2|T_{i,1} = t_1, T_{i,-1})}\,\frac{G'\{X_j^T\beta + m^+(t_1, T_{i,-1})\}}{V(G\{X_j^T\beta + m^+(t_1, T_{i,-1})\})}\\ &\times [G\{X_j^T\beta + m^+(T_j)\} - G\{X_j^T\beta + m^+(t_1, T_{i,-1})\}].\end{aligned}$$

Similarly, one obtains

$$E^*\overline{m}_1^*(t_1) = \overline{m}_1(t_1) + \hat{R}_1 + \hat{D}_1(t_1) + O_P(h^4 + g^4 + (nh)^{-1/2}),$$

where

$$\begin{aligned}\hat{D}_1(t_1) = {}& \frac{1}{\sum_{i=1}^n w_{-1}(T_{i,-1})}\,\frac{1}{n}\sum_{i,j=1}^n \frac{w_{-1}(T_{i,-1})\kappa_j(t_1, T_{i,-1})}{E(Z_i^2|T_{i,1} = t_1, T_{i,-1})}\,\frac{G'\{X_j^T\hat{\beta} + \hat{m}^+(t_1, T_{i,-1})\}}{V(G\{X_j^T\hat{\beta} + \hat{m}^+(t_1, T_{i,-1})\})}\\ &\times [G\{X_j^T\hat{\beta} + \hat{m}^+(T_j)\} - G\{X_j^T\hat{\beta} + \hat{m}^+(t_1, T_{i,-1})\}].\end{aligned}$$

For claim (A.24) it suffices to show

$$D_1(t_1) - \hat{D}_1(t_1) = O_P(h^4 + g^4 + (nh)^{-1/2}). \tag{A.26}$$

This can be done by lengthy but straightforward calculations. We do not want to give all details here. In a first step one shows that

$$\begin{aligned} D_1(t_1) - \hat{D}_1(t_1) = \sum_{i,j=1}^{n} W_{i,j}[&G\{X_j^T\beta + m^+(T_j)\} - G\{X_j^T\beta + m^+(t_1, T_{i,-1})\} \\ &- G\{X_j^T\hat{\beta} + \hat{m}^+(T_j)\} + G\{X_j^T\hat{\beta} + \hat{m}^+(t_1, T_{i,-1})\}] \\ &+ O_P(h^4 + g^4 + (nh)^{-1/2}), \end{aligned} \tag{A.27}$$

where

$$W_{i,j} = \frac{1}{\sum_{i=1}^{n} w_{-1}(T_{i,-1})} \frac{1}{n} \frac{w_{-1}(T_{i,-1})\kappa_j(t_1, T_{i,-1})}{E(Z_i^2|T_{i,1} = t_1, T_{i,-1})} \frac{G'\{X_j^T\beta + m^+(t_1, T_{i,-1})\}}{V(G\{X_j^T\beta + m^+(t_1, T_{i,-1})\})}.$$

The left-hand side of (A.27) can be treated by using Taylor expansions of G and the stochastic expansions of $\hat{m}_j$ given in (A.20). Consider, e.g., for $k \neq 1$

$$\begin{aligned} C_k(t_1) = \sum_{i,j=1}^{n} W_{i,j} G'\{X_j^T\beta + m^+(T_j)\}[&m_k(T_{j,k}) - m_k(T_{i,k}) \\ &- \hat{m}_k(T_{j,k}) + \hat{m}_k(T_{i,k})]. \end{aligned}$$

Then by using the expansions of $\hat{m}_k$ given in (A.20) and the expansion of the bias of $\hat{m}_k$ (see Lemma A2.1) one sees that

$$C_k(t_1) = C_{k1}(t_1) + C_{k2}(t_1) + O_P(h^4 + g^4 + (nh)^{-1/2}),$$

where

$$C_{k1}(t_1) = \sum_{i,j=1}^{n} W_{i,j} G'\{X_j^T\beta + m^+(T_j)\}[-\delta_n^k(T_{j,k}) + \delta_n^k(T_{i,k})]$$

and where

$$C_{k2}(t_1) = \frac{1}{n}\sum_{i=1}^{n} \omega_{i,n}(\mathcal{Z}_n, t_1)\varepsilon_i$$

with some uniformly bounded constants $\omega_{i,n}(\mathcal{Z}_n, t_1)$:

$$\sup_{1 \le i \le n} \sup_{t_1 \in S_{T,1}^-} \omega_{i,n}(\mathcal{Z}_n, t_1) = O_P(1).$$

It can be easily seen that $C_{k1}(t_1) = O_P(h^4 + g^4 + n^{-1/2})$ and $C_{k2}(t_1) = O_P(n^{-1/2})$. We have discussed this term because it shows how the terms of order g^2 cancel in $\hat{m}_1^B(t_1) - m_1(t_1)$. By similar calculations for the other terms one can show the theorem. ■

Proof of Theorem 3.2. For the proof we make use of the following lemma.

LEMMA A3.1. *Under the assumptions of Theorem 3.2, it holds that*

$$v_n^{-1}(R - e_n) \xrightarrow{\mathcal{L}} N(0,1)$$

with

$$e_n = (h_1 \cdot \cdots \cdot h_r)^{-1} \prod_{j=1}^{r} \int K_j(u)^2 \, du E[Af_{T_1}(T_1)],$$

$$v_n^2 = (h_1 \cdot \cdots \cdot h_r)^{-1} \prod_{j=1}^{r} \int K_j^{(2)}(u)^2 \, du E\{E[A|T_1]^2 f_{T_1}(T_1)^3\},$$

$$A = \frac{1}{E[w_{-1}(T_{-1})]} \frac{w_{-1}(T_{-1})w(T)Z^4 f_{T_{-1}}^2(T_{-1})}{E[Z^2|T]^2 f_T^2(T)} \frac{\operatorname{Var}[Y|X,T]}{V\{X^T\beta + m^+(T)\}},$$

where $K_j^{(2)}(u) = \int K_j(u - v)K(v)\, dv$ *is the convolution of* K_j *with itself.*

We now give a proof of Lemma A3.1. Theorem 3.2 follows by replication of the arguments for the "bootstrap world."

We consider the statistic

$$U = \sum_{i=1}^{n} W_i \{\hat{m}_1(T_{i,1}) - E^* \hat{m}_1^*(T_{i,1})\}^2,$$

where

$$W_i = w(T_i) \frac{[G'\{X_i^T\beta + m^+(T_i)\}]^2}{V\{X_i^T\beta + m^+(T_i)\}}.$$

Note that

$$R = \sum_{i=1}^{n} \hat{W}_i \{\hat{m}_1(T_{i,1}) - E^* \hat{m}_1^*(T_{i,1})\}^2$$

with

$$\hat{W}_i = w(T_i) \frac{[G'\{X_i^T\hat{\beta} + \hat{m}^+(T_i)\}]^2}{V\{X_i^T\hat{\beta} + \hat{m}^+(T_i)\}}.$$

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

We will show that

$$U = V + o_p(h^{-1/2}), \tag{A.28}$$

$$R = U + o_p(h^{-1/2}), \tag{A.29}$$

where

$$V = \sum_{i=1}^{n} W_i \{\hat{m}_1^{APPR1}(T_{i,1})\}^2,$$

$$\hat{m}_1^{APPR1}(t_1) = \frac{1}{n}\sum_{i=1}^{n} a^1(X_i, t_1, T_{i,-1}) f_{T_{-1}}(T_{i,-1}) K_h(t_1 - T_{i,1})\,\varepsilon_i,$$

$$\varepsilon_i = Y_i - \mu(X_i, T_i),$$

$$\mu(x,t) = G[x^T\beta + \alpha + \gamma_1 t_1 + m_2(t_2) + \cdots + m_d(t_d)].$$

The function a^1 has been defined in the statement of Lemma A2.1. Asymptotic normality of V can be shown as in Härdle and Mammen (1993). In particular, one gets (with pairwise different indices i, j, k, and l)

$$\begin{aligned} EV &= E\{W_i a^1(X_j, T_{i,1}, T_{j,-1}) f_{T_{-1}}(T_{j,-1})^2 K_h^2(T_{i,1} - T_{j,1}) \operatorname{Var}[Y_j | X_j, T_j]\} \\ &\quad + O(n^{-1}h^{-2}) \\ &= e_n + O(h + n^{-1}h^{-2}), \end{aligned}$$

$$\begin{aligned} \operatorname{Var}[V] &= E\{W_i W_l a^1(X_j, T_{i,1}, T_{j,-1}) a^1(X_j, T_{l,1}, T_{j,-1}) a^1(X_k, T_{i,1}, T_{k,-1}) \\ &\qquad \times a^1(X_k, T_{l,1}, T_{k,-1}) f_{T_{-1}}^2(T_{j,-1}) f_{T_{-1}}^2(T_{k,-1}) \\ &\qquad \times K_h(T_{i,1} - T_{j,1}) K_h(T_{l,1} - T_{j,1}) K_h(T_{i,1} - T_{k,1}) \\ &\qquad \times K_h(T_{l,1} - T_{k,1}) \operatorname{Var}[Y_j | X_j, T_j] \operatorname{Var}[Y_k | X_k, T_k])\} \\ &\quad + O(n^{-1}h^{-2}) \\ &= v_n^2 + O(h + n^{-1}h^{-2}). \end{aligned}$$

Because v_n^2 is of order h^{-1} for the proof of the theorem it remains to show (A.28) and (A.29).

Proof of (A.28). Because $\rho_2^2 = o(n^{-1/2})$, it follows from (A.15) (compare (A.20)) that uniformly for t_1 in $S_{T,1}^{-}$

$$\bar{m}_1(t_1) = m_1(t_1) + R_1 + \Delta_1(t_1) + \frac{E[w_{-1}(T_{-1})M(t_1, T_{-1})]}{E[w_{-1}(T_{-1})]} B_n + o_P(n^{-1/2}),$$

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

where

$$M(t) = \frac{1}{E[Z^2|T=t]} E[Z^2 X^T|T=t] E[\widetilde{X}\widetilde{X}^T|T=t]^{-1},$$

$$B_n = \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_i \lambda_i'[X_j^T\beta + m^+(T_j)].$$

Furthermore, for $\Delta_1(t_1)$ one can show the following uniform expansion:

$$\Delta_1(t_1) = \frac{1}{n}\sum_{i=1}^{n} a^1(X_i, t_1, T_{i,1}) K_h(t_1 - T_{i,1})[Y_i - \mu(X_i, t_1, T_{i,-1})] + o_P(n^{-1/2}).$$

By similar expansions as in the proof of Lemma A2.1 one can show that this implies the following uniform expansion of $\hat{m}_1$:

$$\hat{m}_1(t_1) = \gamma_1 t_1 + \hat{m}_1^{APPR1}(t_1) + \hat{m}_1^{APPR2}(t_1) + \delta_n^1(t_1) + o_P(n^{-1/2}), \qquad \textbf{(A.30)}$$

where

$$\hat{m}_1^{APPR2}(t_1) = \frac{1}{n}\sum_{i=1}^{n} \omega_{i,n,2}(t_1)\varepsilon_i$$

with some uniformly bounded functions $\omega_{i,n,2}$:

$$\sup_{1\le i\le n} \sup_{t_1 \in S_{T,1}^-} \omega_{i,n,2}(t_1) = O(1).$$

The function δ_n^1 has been defined in Lemma A2.1.

Furthermore, using similar arguments as in the proof of Theorem 3.1 one can show that

$$E^*\hat{m}_1^*(t_1) = \tilde{\gamma}_1 t_1 + \delta_n^1(t_1) + \hat{m}_1^{APPR3}(t_1) + o_P(n^{-1/2})$$

with

$$\hat{m}_1^{APPR3}(t_1) = \frac{1}{n}\sum_{i=1}^{n} \omega_{i,n,3}(t_1)\varepsilon_i$$

for some uniformly bounded functions $\omega_{i,n,3}$. Together with (A.30) and a stochastic expansion of $\tilde{\gamma}$ this gives that uniformly for t_1 in $S_{T,1}^-$

$$\hat{m}_1(t_1) - E^*\hat{m}_1^*(t_1) = \hat{m}_1^{APPR1}(t_1) + \hat{m}_1^{APPR4}(t_1) + o_P(n^{-1/2})$$

with

$$\hat{m}_1^{APPR4}(t_1) = \frac{1}{n}\sum_{i=1}^{n} \omega_{i,n,4}(t_1)\varepsilon_i$$

for some uniformly bounded functions $\omega_{i,n,4}$.

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

Claim (A.28) follows from

$$\sum_{i=1}^{n} W_i \{\hat{m}_1^{APPR4}(T_{i,1})\}^2 = o_P(h^{-1/2}),$$

$$\sum_{i=1}^{n} W_i \hat{m}_1^{APPR1}(T_{i,1}) \hat{m}_1^{APPR4}(T_{i,1}) = o_P(h^{-1/2}),$$

$$\sum_{i=1}^{n} |W_i \hat{m}_1^{APPR4}(T_{i,1})| = o_P(n^{1/2}h^{-1/2}),$$

$$\sum_{i=1}^{n} |W_i \hat{m}_1^{APPR1}(T_{i,1})| = o_P(n^{1/2}h^{-1/2}).$$

These bounds can be shown by calculation of expectations of the terms on the left-hand side. ■

Proof of (A.29). Because of Lemma A2.2, we have that $\hat{\beta} - \beta = O_P(n^{-1/2})$ and $\hat{\alpha} - \alpha = O_P(n^{-1/2})$. Moreover we can easily show that

$$\sup_{t_1} \left| \Delta_1(t_1) - \frac{1}{n} \sum_i \Delta_1(T_{i,1}) \right| = O_P(\rho_2).$$

It follows that

$$\sup_{1 \le i \le n} |\hat{W}_i - W_i| = O_P(\rho_2 + n^{-1/2}).$$

Now,

$$\begin{aligned} |U - R| &\le \sup_{1 \le i \le n} |\hat{W}_i - W_i| \sum_{i=1}^{n} \{\hat{m}_1(T_{i,1}) - E^* \, \hat{m}_1^*(T_{i,1})\}^2 \\ &= O_P(\rho_2 + n^{-1/2}) \, O_P(h^{-1}) \\ &= o_P(h^{-1/2}). \end{aligned}$$

This proves (A.29). ■■

Proof of Theorem 3.3. The proof follows the lines of the proof of Theorem 3.2. In a first step one again shows asymptotic normality of the test statistic.

LEMMA A3.2. *Under the assumptions of Theorem 3.3, it holds that*

$$v_n^{-1}(R_{inter} - e_n) \xrightarrow{\mathcal{L}} N(0,1)$$

with e_n and v_n defined as in Lemma A3.1. ■

Proof of Theorem 3.4. The proofs for Models A and B can be done as in Neumann and Polzehl (1998), where wild bootstrap of one-dimensional regression functions has been considered. In this paper it has been shown that the regression estimates in the

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press

bootstrap world and in the real world can be approximated by the same Gaussian process. For this purpose one shows that $\hat{m}_1(t_1) - E[\hat{m}_1(t_1)|\mathcal{Z}_n]$ and $\hat{m}_1^*(t_1) - E^*[\hat{m}_1^*(t_1)]$ have linear stochastic expansions. In particular, using the expansions given in the proof of Lemma A2.1, one shows that

$$\sup_{t_1 \in S_{T,1}^-} \left| \hat{m}_1(t_1) - E[\hat{m}_1(t_1)|\mathcal{Z}_n] - \frac{1}{n}\sum_{i=1}^n a^1(X_i, t_1, T_{i,-1}) f_{T_{-1}}(T_{i,-1}) K_h(t_1 - T_{i,1})\varepsilon_i \right|$$
$$= O_P(n^{-1/2}\sqrt{\log n}).$$

Here, for $\delta > 0$ small enough we have put $S_{T,1}^- = \{s\colon \text{there exists a } u \notin S_{T,1} \text{ with } |s-u| \le \delta\}$. (Then, if δ is small enough we have that $w_1(t_1) = 0$ for $s \notin S_{T,1}^-$.) Similarly one can see that

$$\sup_{t_1 \in S_{T,1}^-} \left| \hat{m}_1^*(t_1) - E^*[\hat{m}_1^*(t_1)] - \frac{1}{n}\sum_{i=1}^n a^1(X_i, t_1, T_{i,-1}) f_{T_{-1}}(T_{i,-1}) K_h(t_1 - T_{i,1})\varepsilon_i^* \right|$$
$$= O_P(n^{-1/2}\sqrt{\log n}).$$

By small modifications of the arguments of Neumann and Polzehl (1998) one can see that their approach carries over to our estimates.

We now will give a sketch of the proof for Model C. First note that $d_K\{\mathcal{L}^+(S), \mathcal{L}(S)\} \to 0$ in probability where $\mathcal{L}^+$ denotes the conditional distribution given $\mathcal{Z}_n = ((X_1, T_{1,1}, \dots, T_{1,d}), \dots, (X_n, T_{n,1}, \dots, T_{n,d}))$. This can be seen as in Neumann and Polzehl (1998). The proof of the theorem will be based on strong approximations. For this purpose we introduce random variables $Y_1^+, Y_1^{++}, \dots, Y_1^+, Y_1^{++}, \dots, Y_n^+, Y_n^{++}$ by the following construction: choose an i.i.d. sample $U_1, \dots, U_n$ that is independent of $\mathcal{Z}_n$. We put $Y_i^+ = F_i^{-1}(U_i)$ and $Y_i^{++} = G_i^{-1}(U_i)$, where F_i and G_i are the distribution functions of $\mathcal{L}^+(Y_i)$ and $\mathcal{L}^*(Y_i^*)$, respectively. Then given the original data $(X_1, T_1, Y_1), \dots, (X_n, T_n, Y_n)$, $(Y_1^+, Y_1^{++}), \dots, (Y_n^+, Y_n^{++})$ are conditionally i.i.d., $\mathcal{L}^*(Y_i^+) = \mathcal{L}^+(Y_i)$ and $\mathcal{L}^*(Y_i^{++}) = \mathcal{L}^*(Y_i^*)$. Furthermore we have that

$$\max_{1\le i\le n} E^*|Y_i^{++} - Y_i^+| = O_P(\rho_2). \tag{A.31}$$

Here E^* denotes the conditional expectation given the original data $(X_1, T_1, Y_1), \dots, (X_n, T_n, Y_n)$. Note that $\mathcal{L}^*(Y_i^+)$ and $\mathcal{L}^*(Y_i^{++})$ belong to the same exponential family with expectation μ_i or $\hat{\mu}_i$, respectively. Property (A.31) follows from

$$E^*|Y_i^{++} - Y_i^+| = \int_0^1 |F_i^{-1}(u) - G_i^{-1}(u)|\, du$$
$$= \int_{-\infty}^{\infty} |F_i(v) - G_i(v)|\, dv$$
$$= O(\mu_i - \hat{\mu}_i) = O_P(\rho_2).$$

Put $\varepsilon_i^+ = Y_i^+ - \mu_i$ and $\varepsilon_i^{++} = Y_i^{++} - \hat{\mu}_i$. The estimate of the first component that is based on the sample $Y_1^+, \dots, Y_n^+$ is denoted by $\hat{m}_1^+(t_1)$. The estimate that is based on $Y_1^{++}, \dots, Y_n^{++}$ is denoted by $\hat{m}_1^{++}(t_1)$.

We argue now that for $\tau > 0$ small enough

$$\max_{1\le i\le n} \sup_{0\le t\le \tau} E^*|\varepsilon_i^{++} - \varepsilon_i^{+}|^2\{1 + \exp(t|\varepsilon_i^{+}|) + \exp(t|\varepsilon_i^{++}|)\} = O_P(\rho_2). \qquad \textbf{(A.32)}$$

This can be seen by straightforward calculations using (A.31) and the fact that the natural parameter of $\mathcal{L}^*(Y_i^{+})$ and $\mathcal{L}^*(Y_i^{++})$ is bounded away from the boundary of the natural parameter space of the exponential family (see Assumption (A2)).

It can be shown that for a sequence $c_n = o(1)$ and for all $a_n < b_n$ with $b_n - a_n \le c_n \log n\,(nh)^{-1/2}$ one has that $P(S \notin [a_n, b_n])$ converges to 0. This can be seen similarly as for kernel smoothers in one-dimensional regression (see, e.g., Neumann and Polzehl, 1998). The statements of Theorem 3.4 follow from

$$\sup_{t_1\in S_{T,1}^-} |\hat\sigma_1(t_1) - \sigma_1(t_1)| = o_P(1), \qquad \textbf{(A.33)}$$

$$\sup_{t_1\in S_{T,1}^-} |\hat\sigma_1^*(t_1) - \sigma_1(t_1)| = o_P([\log n]^{-1}), \qquad \textbf{(A.34)}$$

$$\sup_{t_1\in S_{T,1}^-} |[\hat m_1^{++}(t_1) - \hat m_1(t_1)] - [\hat m_1^{+}(t_1) - m_1(t_1)]| = o_P((nh)^{-1/2}[\log n]^{-1/2}). \qquad \textbf{(A.35)}$$

We give here only the proof of (A.35). One shows first that

$$\sup_{t_1\in S_{T,1}^-} \left| \hat m_1^{+}(t_1) - m_1(t_1) - \frac{1}{n}\sum_{i=1}^{n} a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1})\varepsilon_i^{+} \right|$$
$$= o_P((nh)^{-1/2}[\log n]^{-1/2}),$$
$$\sup_{t_1\in S_{T,1}^-} \left| \hat m_1^{++}(t_1) - \hat m_1(t_1) - \frac{1}{n}\sum_{i=1}^{n} a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1})\varepsilon_i^{++} \right|$$
$$= o_P((nh)^{-1/2}[\log n]^{-1/2}).$$

This can be done by using expansions of the type (A.15). Note that the bias of $\hat m_1^{+}(t_1)$ and $\hat m_1^{++}(t_1)$ is of order $o_P((nh)^{-1/2}[\log n]^{-1/2})$. So, for (A.35) it remains to show

$$\sup_{t_1\in S_{T,1}^-} \left| \frac{1}{n}\sum_{i=1}^{n} a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1})[\varepsilon_i^{+} - \varepsilon_i^{++}] \right|$$
$$= o_P((nh)^{-1/2}[\log n]^{-1/2}). \qquad \textbf{(A.36)}$$

For the proof of this claim we use a standard method that has been applied for calculation of the sup-norm of linear smoothers. We show first that for all constants $C_1 > 0$ there exists a constant C_2 such that

$$\sup_{t_1\in S_{T,1}^-} P^*\left\{ \left| \frac{1}{n}\sum_{i=1}^{n} a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1})[\varepsilon_i^{+} - \varepsilon_i^{++}] \right| > C_2\kappa_n \right\}$$
$$= o_P(n^{-C_1}), \qquad \textbf{(A.37)}$$

where $\kappa_n [nh/\rho_1]^{-1/2} [\log n]^{3/2}$ and where P^* denotes the conditional distribution given the original data $(X_1, T_1, Y_1), \ldots, (X_n, T_n, Y_n)$. Note that $\kappa_n = o((nh)^{-1/2} [\log n]^{-1/2})$. Equation (A.37) implies a modification of claim (A.36) where the supremum runs only over a finite grid of $O(n^{C_1})$ elements. The unmodified claim (A.36) follows by taking a crude bound on

$$\sup_{t_1 \in S_{T,1}^-} \left| \frac{\partial}{\partial t_1} \frac{1}{n} \sum_{i=1}^n a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1}) [\varepsilon_i^+ - \varepsilon_i^{++}] \right|.$$

It remains to show (A.36). Note that

$$\begin{aligned} &P^* \left\{ \frac{1}{n} \sum_{i=1}^n a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1}) [\varepsilon_i^+ - \varepsilon_i^{++}] > C_2 \kappa_n \right\} \\ &\quad \le E^* \exp\left[\log n \kappa_n^{-1} \frac{1}{n} \sum_{i=1}^n a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1}) [\varepsilon_i^+ - \varepsilon_i^{++}] \right] \\ &\qquad \times \exp[\log n \kappa_n^{-1} C_2 \kappa_n] \\ &\quad \le n^{-C_2} \prod_{i=1}^n E^* \exp\left[\frac{\log n}{\kappa_n n} a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1}) [\varepsilon_i^+ - \varepsilon_i^{++}] \right]. \end{aligned}$$

We use now the expansion $\exp[x] \le 1 + x + x^2/2 \, \{1 + \exp[x]\}$. Because of $E^* \varepsilon_i^+ - \varepsilon_i^{++} = 0$ and because of (A.32) this gives that the last term is bounded by

$$\le n^{-C_2} \prod_{i=1}^n \left[1 + C \frac{(\log n)^2}{\kappa_n^2 n^2} a^2(X_i, t_1, T_{i,-1}) K_h^2(t_1 - T_{i,1}) \rho_2 \right],$$

where C is a constant. We use now $1 + x \le \exp[x]$. This gives the bound

$$\le n^{-C_2} \exp\left[\sum_{i=1}^n C \frac{(\log n)^2}{\kappa_n^2 n^2} a^2(X_i, t_1, T_{i,-1}) K_h^2(t_1 - T_{i,1}) \rho_2 \right].$$

With another constant C' this can be bounded by

$$\begin{aligned} &\le n^{-C_2} \exp\left[C' \frac{(\log n)^2}{\kappa_n^2 nh} \rho_2 \right] \\ &\le n^{C' - C_2}. \end{aligned}$$

For C_2 large enough, this is of order $o(n^{C_1})$. This shows (A.36).

https://doi.org/10.1017/S026646660420202X Published online by Cambridge University Press