BOOTSTRAP INFERENCE IN SEMIPARAMETRIC GENERALIZED ADDITIVE MODELS
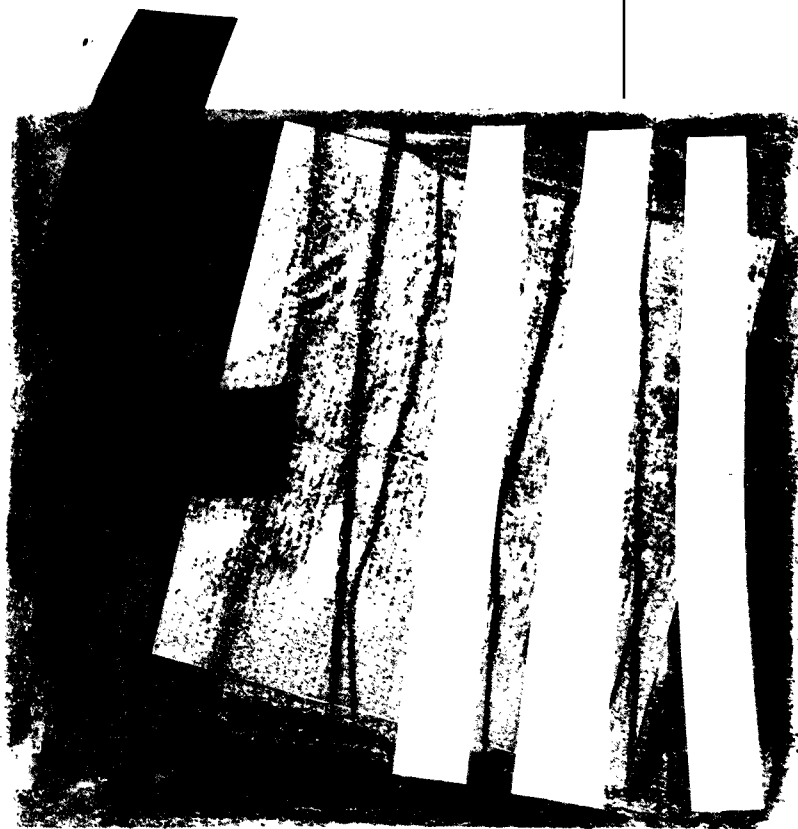
Wolfgang Härdle, Sylvie Huet
Enno Mammen, Stefan Sperlich

00-70

BOOTSTRAP INFERENCE IN SEMIPARAMETRIC GENERALIZED ADDITIVE MODELS

Wolfgang Härdle, Sylvie Huet, Enno Mammen and Stefan Sperlich *

Abstract

Semiparametric generalized additive models are a powerful tool in quantitative econometrics. With response Y, covariates X, T the model is

$$E(Y|X;T) = G \{X^T \beta + \alpha + m_1(T_1) + ... + m_d(T_d)\}.$$

Here, G is a known link, α, β are unknown parameters, and $m_1, ..., m_d$ are unknown (smooth) functions of possibly higher dimensional covariates $T_1, ..., T_d$. Estimates of $m_1, ..., m_d$, α and β are presented and asymptotic distribution theory for both the nonparametric and the parametric part is given. The main focus is the application of bootstrap methods. It is shown that bootstrap can be used for bias correction, hypothesis testing (e.g. component-wise analysis) and the construction of uniform confidence bands. Various bootstrap tests for model specification and parametrization are given, in particular for testing additivity and link function specification. The practical performance of our methods is illustrated in simulations and in an application to East-West German migration.

Keywords: Bootstrap; Specification tests; Generalized Additive models.

* Härdle, Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät, Humboldt-Universität zu Berlin, D 10178 Berlin, Germany; Huet, Institut de Recherche Agronomique, Centre de Recherches de Jouy-en-Josas F 78352 Jouy-en Josas Cedex, France; Mammen, Institut für Angewandte Mathematik, Ruprecht-Karls-Universität Heidelberg Im Neuenheimer Feld 294, D 69120 Heidelberg, Germany; Sperlich, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe, Madrid, Spain, e-mail: stefan@est-econ.uc3m.es. 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, and the Spanish "Dirección General de Enseñanza Superior", number PB98-0025. We thank Marlene Müller, Juan M. Rodríguez-Póo and two anonymous referees for helpful discussion.

1 Introduction

Many problems in applied econometrics and other fields require estimating and analyzing the conditional mean of a random response Y given random covariates. We will consider models where the covariate vector can be decomposed into two components X and T. The vector X enters linearly into the conditional mean; in particular X contains all dummy variables. The influence of T will be described by additive nonparametric functions of the components of T. The conditional mean $m(x,t) = E(Y|X = x; T = t)$ can be written as

$$(1.1) \qquad m(x; t) = G(x^T\beta + \alpha + m_1(T_1) + \ldots + m_d(T_d))$$

for a known link function G, unknown coefficients α, β and unknown nonparametric functions $m_1, \ldots, m_d$ of possibly higher dimensional covariates $T_1, \ldots, T_d$. In this paper we discuss several bootstrap procedures. In particular we use bootstrap for *bias reduction*, for *component testing*, for *testing the link function*, for *testing addidivity* and for *confidence bands* of the nonparametric part of the model.

A traditional estimation approach for $m(x; t)$ assumes that m belongs to a known finite-dimensional parametric family, e.g. a Generalized Linear Model which is specified by

$$(1.2) \qquad m(x; t) = G(x^T\beta + \alpha + t^T\gamma)$$

with a known link function G and unknown coefficients α, β, γ . This model and related parametric specifications are often motivated by economic theory, identifiability conditions or practical reasons. Parameters can be estimated with $O_P(1/\sqrt{n})$ rate of convergence. Clearly, an estimated model of the form (1.2) can be misleading if $m(x; t)$ is misspecified. Misspecifications however may be avoided by non- or semiparametric approaches. In Horowitz (1998) a detailed introduction into semiparametric modeling in econometrics is given. There it is explained that a full nonparametric specification of the influence of t on $m(x; t)$ will not work for high-dimensional t. The nonparametric rate of convergence decreases rapidly as the dimension of the covariables increases ("curse of dimensionality"), see e.g. Stone (1985, 1986). In addition, high-dimensional functions $t \to m(x, t)$ are difficult to graph and to interpret. A natural extension of the purely linear model is the popular Additive (Partially Linear) Model (see e.g. Fan, Härdle, Mammen, 1998)

$$(1.3) \qquad E(Y|X; T) = X^T\beta + \alpha + m_1(T_1) + \ldots + m_d(T_d).$$

The inclusion of a parametric component into the model is of high practical importance. It allows the introduction of discrete covariables, e.g. dummy variables. Furthermore linearly parameterized impact functions are used in several applications. Model (1.3) can be estimated at a rate typical for the lower dimensional explanatory variables $T_1, \ldots, T_d$ (Stone,1985). Projection smoothers using backfitting techniques have been considered in Hastie and Tibshirani (1990). Asymptotic theory for this iterative technique though is rather complicated, see Mammen, Linton and Nielsen (1999) or Opsomer and Ruppert (1997). Tjøstheim and Auestad (1994), and Linton and Nielsen (1995) introduced the marginal integration technique. The technical treatment of this method is easier and allows an asymptotic distribution theory. For a comparison of both approaches see e.g. Sperlich, Linton and Härdle (1999).

Additive models of the form (1.3) do not apply to e.g. binary responses, survival time data or growth data. For such situations regression models have been introduced with link function G that maps the real line into a bounded set. In the parametric case this is provided by

the generalized linear model (1.2), see McCullagh and Nelder (1989). A natural semiparametric extension is given by the Generalized Additive Model (1.1). Both, semiparametric additive and generalized additive models are important in economic theory as well as empirical economics. They have been introduced by Leontief (1947) and have been analyzed in the latent variable set up of Goldberger (1964), their importance is stressed in the standard works of Fuss, McFadden and Mundlak (1978) or Deaton and Muellbauer (1980). Apart from their statistical advantages they allow for economic or econometric analysis for subsets of regressors, permit decentralization in optimization and decision making.

Several semiparametric estimates for (1.1) are proposed in Hastie and Tibshirani (1990) (without asymptotic mathematical theory). For the marginal integration approach estimates of a pure nonparametric components model have been discussed in Linton and Härdle (1996). Horowitz (1999) provides an estimation technique for an additive index with unknown link. The special case of a Generalized Partially Linear Model (i.e. (1.1) with $d = 1$) is well studied. Severini and Staniswalis (1994), Ai (1997) discuss efficient estimation of the parametric component. Their approach is based on iterative application of smoothed local and un-smoothed global likelihood functions. In this paper we will combine and extend ideas of Severini and Staniswalis (1994) with marginal integration to estimate the model. For an approach using smoothing splines that also achieves efficient estimates of the parametric component but avoids an iterative procedure see Mammen and van de Geer (1997). Nonparametric bootstrap tests for Generalized Partially Linear Models can be found in Härdle, Mammen and Müller (1998). The related model $E[Y|X,T] = G\{\beta^T X + m(T^T \alpha)\}$ has been studied by Carroll, Fan, Gijbels and Wand (1997). The main emphasis of their model differs from our approach. Its aim is dimension reduction of the covariable T by projection and therefore the fitted nonparametric transformation m is typically more difficult to interpret than nonparametric curves in model (1.1).

The main subject of this paper is the introduction of bootstrap procedures for the Generalized Additive Model. Bootstrap will be used for bias correction, tests and confidence bands. The decomposition of m into an additive structure allows for component-wise statistical inference. Typically, 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 bias terms. Bootstrap tests will be considered for variable selection, for parametric specifications of impact functions, for testing additivity and for testing the specification of the link function. E.g. in the last testing problem (1.1) will be tested versus the alternative model that includes an interaction term of T_1 and T_2

$$(1.4) \qquad E(Y|X;T) = G\{X^T \beta + \alpha + m_{1,2}(T_1, T_2) + m_3(T_3) + \ldots + m_d(T_d)\}.$$

For tests of interaction terms in a purely additive model see also Gozalo and Linton (1999) and Sperlich, Tjøstheim and Yang (1999). Our approaches for testing make use of some ideas of Hastie and Tibshirani (1990). They proposed (without asymptotic mathematical justification) to use the likelihood ratio test and to take critical values of a χ^2 approximation. The test of this paper differs from this proposal by three modifications. Instead of comparing the nonparametric estimate with a linear fit we propose to compare the nonparametric fit with a bootstrap estimate under the hypothesis. Without this bias correction the test would not behave like an overall test, see Härdle and Mammen (1993) for similar discussions. Our second modification takes care of the fact that different likelihood functions [smoothed and un-smoothed likelihood functions] are used in the construction of the parametric and

nonparametric estimates. Furthermore, we propose using the bootstrap for the calculation of critical values. Consistency of bootstrap is shown by asymptotic theory.

The paper is organized as follows. In the next section we introduce estimates for parameters and nonparametric components of the semiparametric Generalized Additive Model (1.1). Asymptotic theory for these estimates is given in Appendix A2. Section 3 presents several applications of bootstrap for analyzing the nonparametric components. We start by discussion of bootstrap bias corrections for the nonparametric estimates. The largest part of Section 3 is concerned with bootstrap tests for different null hypotheses about the additive components. In the last part of Section 3 procedures and theory for uniform confidence bands are given for the nonparametric additive components. Again, their construction uses bootstrap. In Section 4 the presented methodology is studied in simulations and in an application. Assumptions, asymptotic theory for the estimators and proofs are postponed to the appendix.

2 Quasi-Likelihood Estimation in Generalized Additive Models

We will now describe the Quasi-Likelihood technique. For simplicity of notation we first concentrate on an additive binary model. We then proceed with the generalized additive model. Our estimation starts with the iterative algorithm of Severini and Staniswalis (1994), yielding a nonparametric estimate (that has no additive structure). We then obtain estimates of the additive components by marginal integration. Results on the asymptotic performance are postponed to Appendix A2.

2.1 Additive Binary Response Models

In an additive binary response model i.i.d. tuples (Y_i, X_i, T_i) are observed $(i = 1, \ldots, n)$, where $T_i = (T_{i,1}, \ldots, T_{i,d})$ is a random variable with components $T_{i,j}$ in $\mathbb{R}^{q_j}$, X_i is in $\mathbb{R}^p$ and Y_i is a binary response. Conditionally given (X_i, T_i) the variable Y_i is distributed as a Bernoulli variable with parameter $G\{X_i^T\beta + \alpha + m_1(T_{i,1}) + \ldots + m_d(T_{i,d})\}$ where G is a known (link) function, β is an unknown parameter in $\mathbb{R}^p$, and $m_1 : \mathbb{R}^{q_1} \to \mathbb{R}, \ldots, m_d : \mathbb{R}^{q_d} \to \mathbb{R}$ are unknown functions. The parameter α is in $\mathbb{R}$. For identifiability of this model it is assumed that $E\ w_1(T_{i,1})\ m_1(T_{i,1}) = 0, \ldots, E\ w_d(T_{i,d})\ m_d(T_{i,d}) = 0$ for weight functions $w_1, \ldots, w_d$. Given (X_i, T_i), the (conditional) likelihood of Y_i is

(2.1) $$Q(\mu_i; Y_i) = Y_i \log \mu_i + (1 - Y_i) \log(1 - \mu_i),$$

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

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

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

We now discuss how the additive components $m_1, \ldots, m_d$ can be estimated. Without loss of generality, we will explain this only for the first component m_1. We write $r = q_1$ and

$s = q_2 + \ldots + q_d$. Define the smoothed likelihood

$$(2.3) \qquad \mathcal{L}^S(m^+, \beta) = \int \sum_{i=1}^n K_h(t_1 - T_{i,1}) L_g(t_{-1} - T_{i,-1}) Q\left[G\{X_i^T \beta + m^+(t)\}; Y_i\right] dt,$$

where for 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 the vector $(u_2, \ldots, u_d)^T$ 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 $L_g(v) = (g_1 \cdot \ldots \cdot g_s)^{-1} L(g_1^{-1} v_1, \ldots, g_s^{-1} v_s)$ and for simplicity we assume that L is a product kernel $L = \prod_{j=1}^s L_j$. Similarly, we 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 below.

We now define an estimate of β and a preliminary estimate of m^+. Following Severini and Wong (1992), Severini and Staniswalis (1994) and Härdle, Mammen and Müller (1998) we base our estimates on an iterative application of smoothed local and un-smoothed global likelihood functions. We define for $\beta \in B$

$$(2.4) \qquad \widehat{m}_\beta(t) = \arg\max_\eta \sum_{i=1}^n K_h(t_1 - T_{i,1}) L_g(t_{-1} - T_{i,-1}) Q\left[G\{X_i^T \beta + \eta\}; Y_i\right],$$

$$(2.5) \qquad \widehat{\beta} = \arg\max_{\beta \in B} \mathcal{L}(\widehat{m}_\beta, \beta),$$

$$(2.6) \qquad \widehat{m} = \widehat{m}_{\widehat{\beta}}.$$

Equation (2.4) may be written as $\widehat{m}_\beta = \arg\max_m \mathcal{L}^S(m, \beta)$. The result $\widehat{m}$ is a multivariate kernel estimate of m^+ which does not use the additive structure of m^+. This $\widehat{m}$ will be used in an additional step in order to obtain estimates $\hat{\alpha}, \widehat{m}_1, \ldots, \widehat{m}_d$ of the additive components $\alpha, m_1, \ldots, m_d$. The final additive estimate of $m^+(t)$ will then be given by $\hat{\alpha} + \widehat{m}_1(t_1) + \ldots + \widehat{m}_d(t_d)$. 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 $\{\int w_{-1}(v) dv\}^{-1} \int w_{-1}(v) m^+(t_1, v) dv$ or $\{\frac{1}{n} \sum_{i=1}^n w_{-1}(T_{i,-1})\}^{-1} \frac{1}{n} \sum_{i=1}^n w_{-1}(T_{i,-1})$ $m^+(t_1, T_{i,-1})$ for a weight function w_{-1}. An estimate of m_1 is achieved by marginal integration or summation of an estimate of m. In particular, this method does not use iterations so that the explicit definition allows a detailed asymptotic analysis. A weight function w_{-1} is used here for two reasons. First, it may be useful to avoid problems at the boundary. Second, it can be chosen to minimize asymptotic variance, compare Fan, Härdle and Mammen (1998). For a weight function w_{-1} define

$$(2.7) \qquad \overline{m}_1(t_1) = \frac{\frac{1}{n} \sum_{i=1}^n w_{-1}(T_{i,-1}) \widehat{m}(t_1, T_{i,-1})}{\frac{1}{n} \sum_{i=1}^n w_{-1}(T_{i,-1})},$$

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

$$(2.8) \qquad \widehat{m}_1(t_1) = \overline{m}_1(t_1) - \frac{\frac{1}{n} \sum_{i=1}^n w_1(T_{i,1}) \overline{m}_1(T_{i,1})}{\frac{1}{n} \sum_{i=1}^n w_1(T_{i,1})}.$$

The additive constant α is estimated by

$$(2.9) \qquad \hat{\alpha} = \frac{\frac{1}{n} \sum_{i=1}^n w_0(T_i) \left[\widehat{m}(T_i) - \widehat{m}_1(T_{i,1}) - \ldots - \widehat{m}_d(T_{i,d})\right]}{\frac{1}{n} \sum_{i=1}^n w_0(T_i)}.$$

Again, the weight functions w_0 and w_1 may be useful to avoid problems at the boundary. The remaining nonparametric components are estimated analogously. The final additive estimate of m is given by

(2.10) $$\hat{m}^+(t) = \hat{\alpha} + \hat{m}_1(t_1) + \ldots + \hat{m}_d(t_d).$$

2.2 Generalized Additive Models

We now come to the discussion of the more general case of a generalized additive model. 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 below. For a positive function V the quasi-likelihood function is defined as

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

where μ is the (conditional) expectation of Y, i.e. $\mu = G\{X^T \beta + m(T)\}$. The quasi-likelihood 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. Maximum quasi-likelihood estimates can thus be interpreted as a modification of weighted least squares. Another motivation comes from the fact that for exponential families the maximum quasi-likelihood 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 $Var[Y_i | X_i, T_i] = \sigma^2 V(\mu_i)$ 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 quasi-likelihood function is well motivated for Models B and C. The more general Model A is included here for discussion of robustness issues. It allows to discuss the case of a wrongly specified [conditional] variance in Models B and C. If not otherwise stated all of the following remarks and results treat the most general Model A. The quasi-likelihood function and the smoothed quasi-likelihood function is now defined as in (2.2) and (2.3) with (2.1) replaced by (2.11). The estimates $\hat{m}_\beta$, $\hat{\beta}$, $\hat{m}$, $\overline{m}_1$, $\hat{m}_1$, $\hat{m}^+$ and $\hat{\alpha}$ are defined as in (2.4) - (2.8). Asymptotics for $\hat{m}_1$ are presented in Appendix A2. In particular, Theorem A2.1 shows that

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

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 $h_+ = \max_{1 \leq j \leq r} h_j$ and $g_+ = \max_{1 \leq j \leq s} g_j$. For a definition of A and B

see Theorem A2.1. Theorem A2.1 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 is asymptotically negligible and so asymptotics suggests the choice $g_+ = o(h_+)$. However, stochastic and numerical stability of the pre-estimator $\hat{m}$ demands that $h_1 \cdot \ldots \cdot h_r \cdot g_1 \cdot \ldots \cdot g_s$ is large. Otherwise not enough covariables would lie in the local support of the multidimensional kernel. Often even larger values for g_j than for h_l are needed for satisfactory performance of $\hat{m}$. The constant A 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.

The optimal rate of convergence for twice differentiable functions m_1 is $n^{-2/(4+r)}$. As long as second order kernels K and L are used this rate can be achieved under the assumptions of Theorem A2.1 only for dimension $s \leq 1$. For higher dimensions s, one can see from our expansions that the $n^{-2/(4+r)}$ rate can be achieved by using higher order kernels $L_1, \ldots, L_s$. Under higher order smoothness assumptions on m_1 higher order kernels K_j and L_i could be used. Also in this case one can show that the optimal rate for estimation of m_1 can be achieved under weaker conditions on s. Furthermore, it can be shown that Theorem A2.1 holds under weaker conditions on the bandwidths g and h. However, an essential generalization would require complex higher order stochastic expansions of the pilot estimate $\hat{m}$.

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 Appendix A2. Asymptotic normality of $\hat{\beta}$ is shown in Theorem A2.2. It turns out that for unbiasedness no undersmoothing is required in the nonparametric estimation part. The asymptotic performance of $\hat{\alpha}$ will be discussed after Theorem A2.2.

3   Bootstrap Applications in Generalized Additive Models

In this section we will propose bootstrap procedures for bias correction, testing and confidence bands. 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 which 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, variance 1 and which fulfill for a constant C that $|\varepsilon_i^*| \leq C$ (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

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

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 above 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 sub-exponential tails, i.e. for a constant C it holds that $E(\exp\{[|\varepsilon_i^*|/C]\}) \leq C$ for $i = 1, \ldots, n$ [compare (A2)].

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 $Y_1^*, \ldots, Y_n^*$ 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

Theorem A2.1 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 $m_2, \ldots, m_d$. This may lead to wrong interpretations of the estimate $\hat{m}_1$. Bootstrap bias estimates shall help to judge such effects.

In all three resampling schemes, one uses the data $(X_1, T_1, Y_1^*), \ldots, (X_n, T_n, Y_n^*)$ 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), $\ldots, (X_n, T_n, Y_n)$. The bias corrected estimate of $m_1(t_1)$ is defined by

$$\hat{m}_1^B(t_1) = \hat{m}_1(t_1) - \hat{\delta}_n^1(t_1),$$

where $\hat{\delta}_n^1(t_1) = E^* \hat{m}_1^*(t_1) - \hat{m}_1(t_1)$. The next 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 hold and that the corresponding version of bootstrap is used. Suppose further that assumptions (A1) - (A11) apply and that assumptions analogous to (A3) and (A4) hold for the estimation of the other additive components m_j for $j = 2, \ldots, d$ [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 $nh_1 \cdot \ldots \cdot h_r g_1^2 \cdot \ldots \cdot g_s^2 (\log n)^{-2}$ tends to infinity. Then it holds that

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

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

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 $G\{X_i^T \hat{\beta} + \hat{\alpha} + \hat{m}_1^O(T_{i,1}) + \hat{m}_2(T_{i,2}) + \ldots + \hat{m}_d(T_{i,d})\}$, where $\hat{m}_1^O$ is defined as $\hat{m}_1$ but with bandwidth vector h^O instead of h. Then if $h_j^O/h_+ \to \infty$ ($1 \leq j \leq r$) one can show that the left hand side of (3.1) is of order $O_p\{(h_+^O)^4 + g^4 + (nh_1^O \cdot \ldots \cdot h_r^O)^{-1/2}\}$, where h_+^O is the maximal element of h^O. For appropriate choices of h^O (e.g. for h^O with $h_+^O)^4$ and $(nh_1^O \cdot \ldots \cdot h_r^O)^{-1/2}$ of the same asymptotic order) this is of smaller order than the right hand side of (3.1) with resampling from $\hat{m}_1$.

8

3.2 Component-wise 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. A hypothesis test as well as uniform confidence bands are useful and necessary tools to analyze and interpret the estimated component functions. As discussed above, there are several reasons why it is interesting to test special parameterizations of the components. We concentrate here on the hypothesis that the components are linearly parameterized. This contains the case of polynomial regression and in particular it includes tests for linearity of a component

$$(3.2) \qquad H_0: \ m_1(t_1) = \gamma_1 \ t_1 \quad \text{for all } t_1 \text{ and a scalar } \gamma_1,$$

and for variable selection

$$(3.3) \qquad H_0: \ m_1(t_1) = c_1 \quad \text{for all } t_1 \text{ and a constant } c_1, \text{ respectively.}$$

Due to the condition $E[m_1] = 0$, in (3.3) we have $c_1 = 0$. The variable selection problem is easier to handle than the first testing problem. For this reason we only treat tests for hypothesis (3.2).

As indicated above, 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 give better approximations for the degrees of freedom. First we correct for the bias of the nonparametric estimate. Second, we correct the test statistic for the different likelihoods [smoothed or un-smoothed] that are used in the calculation of the nonparametric or parametric components. For this modified test statistic asymptotic normality is established in Theorem 3.2. The convergence to the normal limit is very slow as mathematical arguments and simulations indicate. Therefore we propose the bootstrap also for the calculation of critical values. Consistency of bootstrap is shown in Theorem 3.3.

The bias correction is used in the test because also on the hypothesis the estimate $\hat{m}_1(t_1)$ may have a non-negligible bias. For this reason, $\hat{m}_1(t_1)$ is compared with a bootstrap estimate of its expectation under the hypothesis. We therefore calculate semiparametric estimates in the hypothesis model (3.2)

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

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

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

For the bootstrap one proceeds now as follows: generate independent samples $(Y_1^*, \ldots, Y_n^*)$ as in the beginning of Section 3 but 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

9

mean of $\widehat{m}_1(t_1)$ is given by $E^*\widehat{m}_1^*(t_1)$, where E^* denotes the conditional expectation given the sample $(X_1, T_1, Y_1), \ldots, (X_n, T_n, Y_n)$. Define the following test statistic:

$$(3.4) \qquad R = \sum_{i=1}^{n} w(T_i) \frac{[G'\{X_i^T\widehat{\beta} + \widehat{m}^+(T_i)\}]^2}{V(G\{X_i^T\widehat{\beta} + \widehat{m}^+(T_i)\})} \, \{\widehat{m}_1(T_{i,1}) - E^*\widehat{m}_1^*(T_{i,1})\}^2 \, .$$

Here, $\widehat{m}^+(t) = \widehat{\alpha} + \widehat{m}_1(t_1) + \ldots + \widehat{m}_d(t_d)$. The weights $[G'\{\ldots\}]^2/V(G\{\ldots\})$ in the summation of the test statistic are motivated by likelihood considerations, see Härdle, Mammen and Müller (1998). They could be replaced by other weights. The following theorem states that the test statistic R has an asymptotic normal distribution.

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

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

with

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

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

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

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

The quantities e_n and v_n may be consistently estimated. Critical values for the test statistic can be therefore calculated using the normal approximation. In similar cases the normal approximation does not perform well (see e.g. Härdle, Mammen and Müller, 1998), we thus propose using the bootstrap for the calculation of critical values of the test statistic R. The bootstrap estimate of the distribution of R is given by the conditional distribution of the test statistic R^*, where R^* is defined as follows.

$$(3.5) \qquad R^* = \sum_{i=1}^{n} w(T_i) \frac{[G'\{X_i^T\widehat{\beta} + \widehat{m}^+(T_i)\}]^2}{V\{X_i^T\widehat{\beta} + \widehat{m}^+(T_i)\}} \, \{\widehat{m}_1^*(T_{i,1}) - E^*\widehat{m}_1^*(T_{i,1})\}^2 \, .$$

The conditional distribution $\mathcal{L}^*(R^*)$ (given the original data $(X_1, T_1, Y_1), \ldots, (X_n, T_n, Y_n)$) is our bootstrap estimate of the distribution $\mathcal{L}(R)$ of R (on the hypotheses (3.2)).

Consistency of bootstrap is the content of the next theorem.

Theorem 3.3
Under the assumptions of Theorem 3.2, 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

$$d_K(\mu,\nu) = \sup_{t \in \mathbb{R}} \left| \mu(X \leq t) - \nu(X \leq t) \right|.$$

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 $n^{-1/2}(h_1 \cdot \ldots \cdot h_r)^{-1/4}$. This means that the test does not reject alternatives that have a distance of order $n^{-1/2}$. However, the test detects also local deviations [of order $n^{-1/2}(h_1 \cdot \ldots \cdot h_r)^{-1/4}$] 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 sensitive the test reacts on local deviations. For smaller h the test detects deviations that are more locally concentrated, at the cost of a poorer power performance for more global deviations. In particular, as an extreme 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 this case R is a $n^{-1/2}$ consistent overall test.

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

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

where $\theta_1, \ldots, \theta_q$ are unknown parameters and where $f_1, \ldots, f_q$ are some given functions. In particular, this includes the hypothesis (3.3), i.e. testing for significant impact of T_1, as well as the model in which m_1 is a polynomial. Moreover, the results of this section can be extended to tests of other parametric hypotheses on m_1, e.g.

(3.6) $\qquad H_0: \ m_1(t_1) = m_\theta(t_1) \quad$ for all t_1 and a parameter θ,

where $\{m_\theta : \theta \in \Theta\}$ is a parametric family. This can be done similarly as in Härdle and Mammen (1993). In our case it requires an asymptotic study of parametric estimates in the model (1.1) with parametric specification (3.6) for m_1.

3.3   Testing Separability and Interactions

Our estimate of m_1 is robust against non-additivity of the other components. In the construction of the estimate it is only used that $m(x;t)$ is of the form

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

for an arbitrary function $m_{2,\ldots,d}$. It is not assumed that the function $m_{2,\ldots,d}$ is additive, i.e. $m_{2,\ldots,d}(T_2, \ldots, T_d) = m_2(T_2) + \ldots + m_d(T_d)$. Also even in case that $m(x;t)$ is not of the form (3.7) the estimate $\widehat{m}_1$ still makes sense and can be interpreted as an estimate of the average effect of T_1. Nevertheless the hypothesis of additivity is of its own interest and may be an important step in a model choice procedure. Following a similar idea as Sperlich, Tjøstheim and Yang (1999), we therefore consider a split of the first covariate T_1 into two components $T_{1:1}$ and $T_{1:2}$ and consider the hypothesis that

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

11

Estimates of $m_{1:1}$ and $m_{1:2}$ can be constructed by marginal integration

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

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

The remainder, i.e. $\widehat{m}_{1:1,2}(t_1) = \widehat{m}_1(t_1) - \widehat{m}_{1:1}(t_{1:1}) - \widehat{m}_{1:2}(t_{1:2})$ is an estimate for the (first order) interaction of $T_{1:1}$ and $T_{1:2}$. Note that model (1.4) fits into this framework. There the interaction between T_1 and T_2 has been considered.

We now discuss tests for the hypothesis (3.8). For this purpose we proceed similarly as in Section 2.2 and define

$$R_{inter} = \sum_{i=1}^{n} w(T_i)\frac{[G'\{X_i^T\widehat{\beta} + \widehat{m}^+(T_i)\}]^2}{V(G\{X_i^T\widehat{\beta} + \widehat{m}^+(T_i)\})}\left\{\widehat{m}_{1:1,2}(T_{i,1:1}, T_{i,1:2}) - E^*\widehat{m}_{1:1,2}^*(T_{i,1:1}, T_{i,1:2})\right\}^2,$$

where $m_{1:1,2}^*$ is an estimate based on a bootstrap sample. Bootstrap samples are generated as described in the beginning of Section 3 but now with $\widehat{\mu}_i$ replaced by

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

Asymptotic normality of the test statistic R_{inter} is stated in the next theorem.

Theorem 3.4
Under the assumptions of Theorem 3.2, on the hypotheses (3.8), it holds that

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

with e_n and v_n defined as in Theorem 3.2.

As in Section 3.2 we argue that the normal approximation is too inaccurate and we therefore use bootstrap for the calculation of critical values. The bootstrap estimate of the distribution of R_{inter} is given by the conditional distribution of the test statistic R_{inter}^*, where R_{inter}^* is defined as follows (compare also (3.5)).

$$(3.9) \qquad R_{inter}^* = \sum_{i=1}^{n} w(T_i)\frac{[G'\{X_i^T\widehat{\beta} + \widehat{m}^+(T_i)\}]^2}{V\{X_i^T\widehat{\beta} + \widehat{m}^+(T_i)\}} \times$$
$$\left\{\widehat{m}_{1:1,2}^*(T_{i,1:1}, T_{i,1:2}) - E^*\widehat{m}_{1:1,2}^*(T_{i,1:1}, T_{i,1:2})\right\}^2,$$

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

Consistency of this bootstrap is the content of the next theorem.

Theorem 3.5
Under the assumptions of Theorem 3.4, it holds that

$$d_K\{\mathcal{L}^*(R^*), \mathcal{L}(R)\} \overset{P}{\longrightarrow} 0.$$

The proofs of Theorems 3.4 and 3.5 are similar to the proofs of Theorems 3.2 and 3.3, respectively, and therefore skipped.

12

3.4 Testing the Link Function

We now discuss tests for the specification of the link function G. Horowitz and Härdle (1994) introduced the HH test for the null hypothesis of a fully parametric generalized linear model. Härdle, Spokoiny and Sperlich (1997) discussed rate optimality of those tests. Härdle, Mammen and Proenca (2000) introduced bootstrap modifications of the HH statistic. We use the latter approach for the null hypothesis of a generalized additive model.

$$(3.10) \qquad H_0 \quad : \quad E[Y|X,T] = G\{v(T,X,\beta)\} \quad \text{versus}$$

$$(3.11) \qquad H_1 \quad : \quad E[Y|X,T] = H\{v(T,X,\beta)\} \quad \text{where } H(\cdot) \text{ unknown.}$$

In our case we have $v(T,X,\beta) = \beta^T X + \alpha + m_1(T_1) + \ldots + m_d(T_d)$. The test statistic is defined as

$$(3.12) \qquad h_L^{1/2} \sum_{i=1}^{n} w(\widehat{v}_i) \frac{\sum_{j\neq i}^{n}[Y_j - G(\widehat{v}_j)]K(\{\widehat{v}_j - \widehat{v}_i\}/h_L)}{\sum_{j\neq i}^{n} K(\{\widehat{v}_j - \widehat{v}_i\}/h_L)}[Y_i - G(\widehat{v}_i)],$$

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

3.5 Uniform bootstrap confidence bands

In this section we apply the bootstrap to the construction of uniform confidence bands. We define

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

where $\widehat{\sigma}_1^2(t_1)$ is the estimate of the variance of $\widehat{m}_1(t_1)$, defined in (A2.1). In the simulations we also tried a bootstrap estimate of $\sigma_1(t_1)$, see Section 4. For the estimation of the distribution of S we again use bootstrap, as introduced in Section 2.2 for Model C. This defines the statistic $S^* = \sup_t w_1(t_1)|\widehat{m}_1^*(t_1) - E^* \widehat{m}_1^*(t_1)|\widehat{\sigma}_1^{-1}(t_1)$. In the definition of S^* the norming $\widehat{\sigma}(t_1)$ could be replaced by $\widehat{\sigma}_1^*(t_1)$. We write $S^{**} = \sup_t w_1(t_1)|\widehat{m}_1^*(t_1) - E^* \widehat{m}_1^*(t_1)|[\widehat{\sigma}_1^*]^{-1}(t_1)$. Here $\widehat{\sigma}_1^*(t_1)$ is an estimate of the variance of $\widehat{m}_1^*(t_1)$, that is defined similarly as $\widehat{\sigma}_1(t_1)$ but that uses a bootstrap resample instead of the original sample. The first norming may help to save computation time, for the second choice bootstrap theory from other set ups suggests higher order accuracy of bootstrap.

Both bootstrap procedures can be used to construct valid uniform confidence bands for additive components. This follows from the following theorem.

Theorem 3.6
Assume that Model A , Model B or Model C hold 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 $nh_1 \cdot \ldots \cdot h_r g_1^2 \cdot \ldots \cdot g_s^2 (\log n)^{-2} \to \infty$. Then it holds that

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

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

¿From Theorem 3.6 we see that critical values of S can be consistently estimated by bootstrap. This gives uniform confidence intervals for $m_1(t_1) - \delta_n^1(t_1)$. For confidence bands for m_1

we need a consistent estimate of $\delta_n^1(t_1)$. Estimation of $\delta_n^1(t_1)$ can be done by plug-in or bootstrap. Both approaches require oversmoothing, i.e. choice of a bandwidth vector h^O with $h_j^O/h_+ \to \infty$, see also the remark after Theorem 3.1. For related discussions in nonparametric density estimation and regression see Bickel and Rosenblatt (1973), Eubank and Speckman (1993) or Neumann and Polzehl (1998).

4   Simulations and Applications

In the first part of this section we present simulations for the performance of our methods in small samples. We give results for estimation, for 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 also power functions are calculated. Furthermore, coverage probabilities of confidence bands are checked. In the second part, we examine a real data example in an empirical economic analysis, based on studies of Burda (1993) and Burda, Härdle, Müller, Werwatz (1998). Throughout this section we set all weight functions w_{-1}, w_0 and w_1 equal to 1 for simplicity.

4.1   Simulation Results

Binary response data are simulated from

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

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 chosen as independent random variables, X_1 and X_2 are standard normal and T_1 and T_2 have 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 in this section the quartic kernel $\frac{15}{16}(1 - u^2)^2 I(|u| \le 1)$ is used. Further, in this section we use the following notation: h_1 is the bandwidth that is used for the estimation of β; the bandwidth that is applied for the estimation of the fitted components m_k [$k = 1, = 2,$] is denoted by h. Finally, g is the bandwidth used in the smoothing for the nuisance component(s) m_j [$j \ne k$]. Figure 1 shows plots of m_1, m_2 and of the estimates for one sample. The chosen bandwidths are $h_1 = g = (1.0, 1.0)^T$, $h = 0.9$.

We now consider the test problem (3.2) $H_0 : m_1(t_1)$ *is linear*. As discussed above the normal approximation of Theorem 3.2 is quite inaccurate for a small sample size like $n = 250$. This can be seen from Figure 2. There 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 $m_1(t_1) = t_1$, β, m_2 and α as above. 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 with 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.

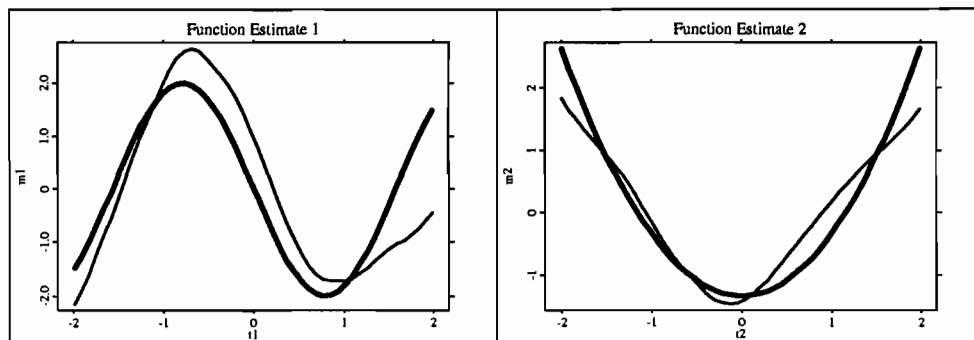Figure 3 plots the power of the test (thick line) for theoretical levels 0.05 and 0.1. The power

14

Figure 1: *Plots of the nonparametric components* $m_1(t_1) = 2\sin(-2t_1)$, $m_2(t_2) = t_2^2 - E[T_2^2]$ *and their estimates.*

is plotted 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 above. For comparison, we perform the same simulations for a parametric Likelihood Ratio Test (LRT) that tests the hypothesis $\gamma_1 = \ldots = \gamma_4 = 0$ 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 since for the parametric test the alternative as well as m_2 are assumed to be known. Nevertheless, the better performance of the parametric test [see Figure 3] is partly just due to the fact that the test R is conservative, see above. [Compare the power of R in the right plot with the power of the Likelihood Ratio Test in the left plot.]

For bootstrap confidence bands we investigate four questions; what is the coverage accuracy in a small sample, does it depend strongly on the bandwidth, how much vary the width of the band with the chosen coverage probability, and finally, does it really matter how we estimate $\sigma_1^2(t_1)$. In the simulations we use two estimates of $\sigma_1^2(t_1)$. The first estimate $\hat{\sigma}_1^2(t_1)$ was defined in (A2.1), see Section 3. The second estimate is the empirical variance of $m_1^*(t_1)$ in the bootstrap resamples. This estimate is denoted by $\tilde{\sigma}_1^2(t_1)$. The simulated model is again

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

where G is the Logit distribution function, $\beta = (0.3, -0.7)^T$, $m_1(t_1) = 2\sin(-2t_1)$, $m_2(t_2) = t_2^2 - E[T_2^2]$, $\alpha = 0$, and $n = 250$. Now T has a uniform distribution on $[-1, 1]$. The simulation is only done for the first component m_1. For simplicity we only use identical bandwidths $h_1 = h = g$.

For bandwidths from $h_1 = h = g = .3$ to .6 we obtain reasonable coverage accuracies. The results for $h_1 = h = g = .5$ are given in Table 1. The empirical coverage probabilities are close to the theoretical ones for all levels and for both variance estimates. For too small and too large bandwidths the accuracy is less accurate. This is caused by poorer bootstrap bias correction. The variance of the estimates is always well estimated by the bootstrap.

In Figures 4, 5 we compare 95% and 85% confidence bands. Despite their different levels the bands do hardly differ. So it is surprising how well the bootstrap fits the different coverage probabilities.
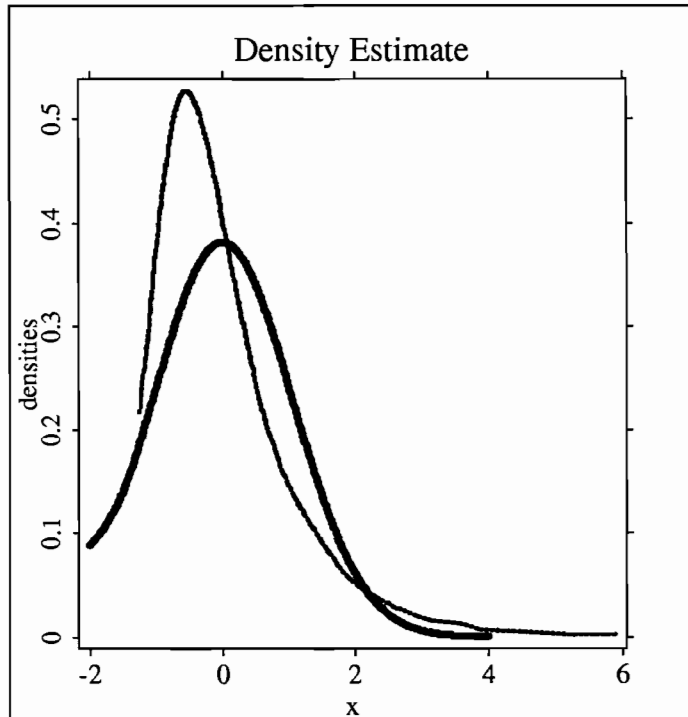
15

Figure 2: *Standardized density estimate of the test statistic (thin line) and standard normal density (thick line).*
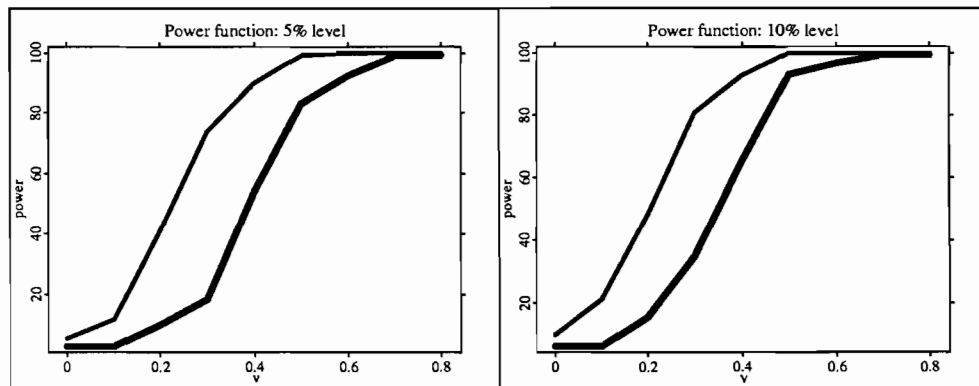

Figure 3: *Power functions for theoretical levels 0.05 and 0.1 , for the nonparametric bootstrap test (thick line) and the likelihood ratio test (thin line).*

Theoretical coverage	95%	90%	85%	80%
using $\hat{\sigma}_1^2(t_1)$	.963	.912	.846	.776
using $\tilde{\sigma}_1^2(t_1)$	.948	.904	.839	.776

Table 1: *Coverage probabilities for bootstrap confidence bands with $h_1 = h = g = 0.5$.*

16
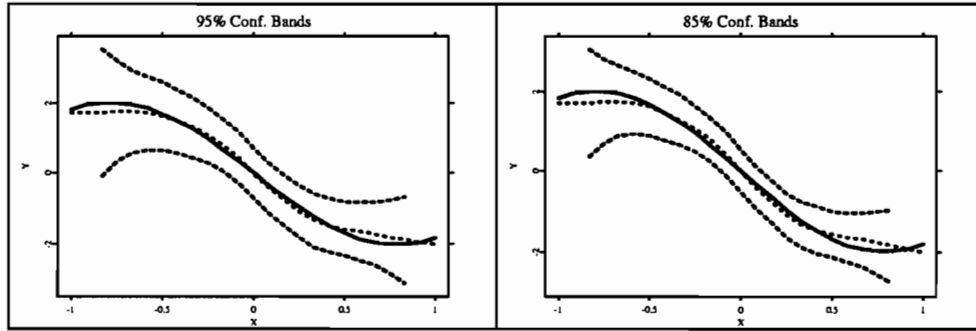
Figure 4: *95% and 85% Confidence Bands, using $\hat{\sigma}^2$. Dashed lines are the confidence bands and corresponding estimates, solid lines are the data generating functions.*
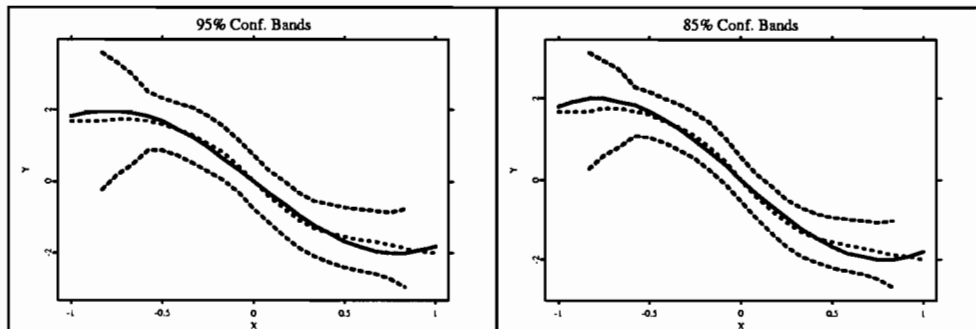

Figure 5: *95% and 85% Confidence Bands, using $\tilde{\sigma}$. Dashed lines are the confidence bands and corresponding estimates, solid lines are the data generating functions.*

Finally, we simulate the test for the link function that has been introduced in Section 3.4. In the simulations this test checks the null hypothesis of a logit link. The data are generated as in the simulations of confidence bands. The simulation results for level accuracy are reported in Table 2 for $h_1 = h = g = .35$. We also tried different bandwidths with no major difference in the results. An additional bandwidth h_L is used in the definition of the test statistic, see Section 3.4. This has been chosen as $h_L = 0.4 \cdot \hat{s}_I$ where $\hat{s}_I$ is an estimate of the standard deviation s_I of the index. As can be seen from Table 2 bootstrap critical values work quite well.

1%	5%	10%	15%
.014	.046	.090	0.13

Table 2: *Levels for testing the Link function using $h_L = \hat{s}_I * .4$, $n = 250$, $B = 250$.*

17

4.2 Application to East-West German migration data.

The following discussion of an application has two major intentions. On the one hand we want to demonstrate that our methods are feasible and robust in practice and do not break down when the model and data do not follow a simulated ideal situation but are drawn from a realistic complex structure. Those complex structures are typical in empirical research. Furthermore we discuss possible additional insights that may be gained by the more flexible structure of our semiparametric modeling. Additionally, robustness of our tools against choice of the bandwidths is shown and we will see that the findings of our different statistical tools are not contradictory but consistent in themselves.

For a 1991 GSOEP data set, Burda (1993), Müller (2000) and Burda, Härdle, Müller, Werwatz (1998) investigated the impact of various possible determinants on the intention of East-Germans to migrate to West Germany. The original data set contains 3710 East Germans who were surveyed in 1991 in the Socio-Economic Panel of Germany, see GSOEP (1991). Here we consider the data sets from two East German countries: the most northern country of East Germany, i.e. Mecklenburg-Vorpommern (M-V) with $n = 402$, and the most southern one, Sachsen (Sax) with $n = 955$ observations. We use the following explanatory variables: family/friend in West, unemployed/job loss certain, middle sized city (10000-100000 habitants) and female [dummies (= 1 if yes, = 0 if no), age (AGE) and household income (HHINCOME) [continuous variables]. The response is 1 if the person said he is willing to migrate and 0 otherwise. Figure 6 gives plots for the densities of AGE and HHINCOME for both countries. Tables 3 and 4 contain descriptive statistics. In the following, the variables AGE and HHINCOME are studentized.
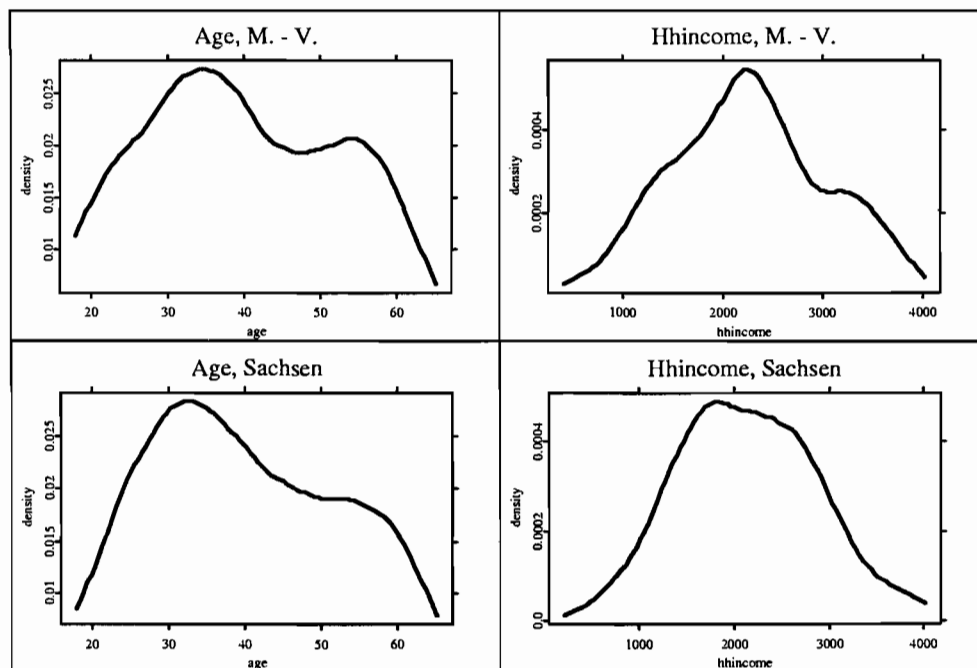

Figure 6: *Density plots for Mecklenburg-Vorpommern (top) and Sachsen (bottom),* AGE *on the left,* HHINCOME *on the right.*

In a first step we fit a parametric generalized linear regression model with logit link. The

18

sample size n			402	
	min.	max.	mean	stdev.
response y	0	1	0.390547	0.488481
family/friends in West x_1	0	1	0.88806	0.315686
unemployed/job loss certain x_2	0	1	0.211443	0.40884
city size (10000-100000) x_3	0	1	0.358209	0.480071
female x_4	0	1	0.502488	0.500617
age t_1	18	65	39.9353	12.8911
household income t_2	400	4000	2262.22	769.822

Table 3: *Descriptive statistics for the data from Mecklenburg-Vorpommern.*

sample size n			955	
	min.	max.	mean	stdev.
response y	0	1	0.395812	0.489281
family/friends in West x_1	0	1	0.824084	0.380948
unemployed/job loss certain x_2	0	1	0.183246	0.387071
city size (10000-100000) x_3	0	1	0.259686	0.438692
female x_4	0	1	0.51623	0.499998
age t_1	18	65	40.3675	12.6942
household income t_2	200	4000	2136.31	738.719

Table 4: *Descriptive statistics for the data from Sachsen.*

results are presented in Table 5 for both countries, M-V and Sax.

PARAMETRIC ESTIMATION RESULTS

	Mecklenburg-Vorpommern			Sachsen		
	Coeff.	stdev.	$P > |z|$	Coeff.	stdev.	$P > |z|$
family/friends West	0.5893	0.3820	0.124	0.7604	0.1972	<0.001
unemployed/...	0.7799	0.2779	0.005	0.1354	0.1783	0.447
middle sized city	0.8216	0.2421	0.001	0.2596	0.1556	0.085
female	-0.3884	0.2315	0.093	-0.1868	0.1382	0.178
age (standardized)	-0.9227	0.1330	<0.001	-0.5051	0.0728	<0.001
hh. income (stand.)	0.2318	0.1221	0.057	0.0936	0.0707	0.187
constant	-1.3673	0.2969	0.001	-1.0924	0.2003	<0.001

Table 5: *Results of a generalized linear regression.*

The variable AGE is by far the most significant variable. This holds true for both countries.

On the other hand people behave quite differently in the two countries. This concerns X_1 [relatives or friends in West Germany], X_2 [their status of employment] and X_3 [city size].

In a second step we fit a semiparametric generalized additive model for both data sets. We present the results for different smoothing parameters, choosing $h = 1.0$ and $h = 1.25$ for M-V and $h = 0.75$, $h = 1.0$ for Sax. The other bandwidths have always been $h_1 = g = 1.1 \cdot h$. In Figures 7 and 8 the additive components for AGE and HHINCOME are plotted. Table 6 gives the estimates for the parametric part.
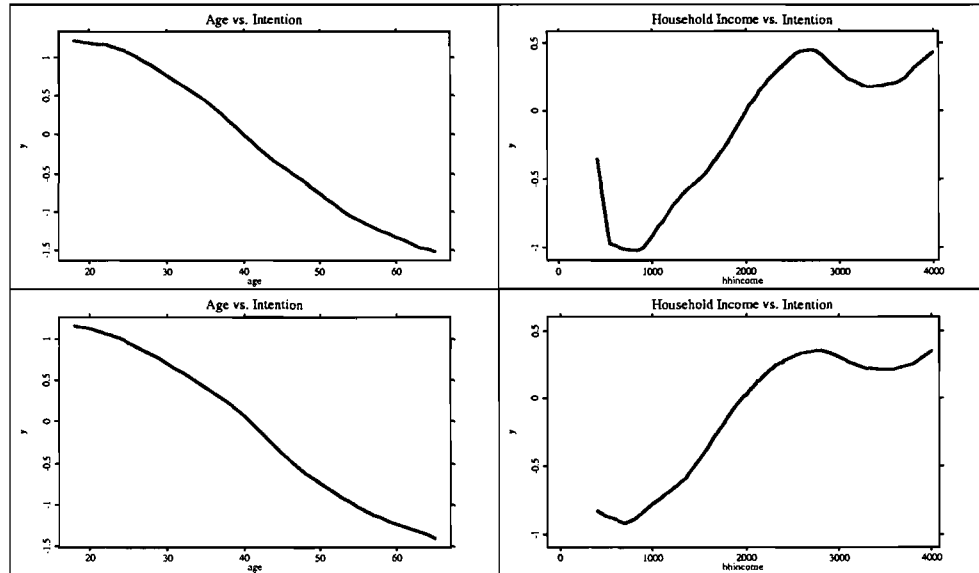
Figure 7: *The semiparametric estimates for the influence of* AGE *(left) and* HHINCOME *(right) in Mecklenburg-Vorpommern. The upper plots were estimated with* $h = 1.0$, *the lower ones with* $h = 1.25$.

COEFFICIENTS OF THE LINEAR PART

	M-V			Sax		
	par.	semi. a	semi. b	par.	semi. a	semi. b
family/friends West	.5893	.5920	.5809	.7604	.7137	.7289
unemployed/...	.7799	.7771	.7992	.1354	.1469	.1308
middle sized city	.8216	.7156	.7127	.2596	.3134	.2774
female	-.3884	-.3309	-.3485	-.1868	-.1898	-.1871
constant	-1.367	-1.462	-1.411	-1.092	-1.105	-1.101

Table 6: *Results of purely parametric estimates (par.) and of the parametric part of a generalized additive partially linear regression model: semi. a (with bandwidth* $h = 1.0$*), semi. b (*$h = 1.25$*) for M-V; semi. a (*$h = 0.75$*) and semi. b (*$h = 1.0$*) for Sax.*

All estimates do not depend strongly on the chosen bandwidth. Moreover, for the linear part of the model the results are similar to the values of the parametric model, see Table 6. So the qualitative interpretation of the parametric coefficients does not change. In the figures the influence of AGE in M-V does not differ strongly from the influence of AGE in Sax, except
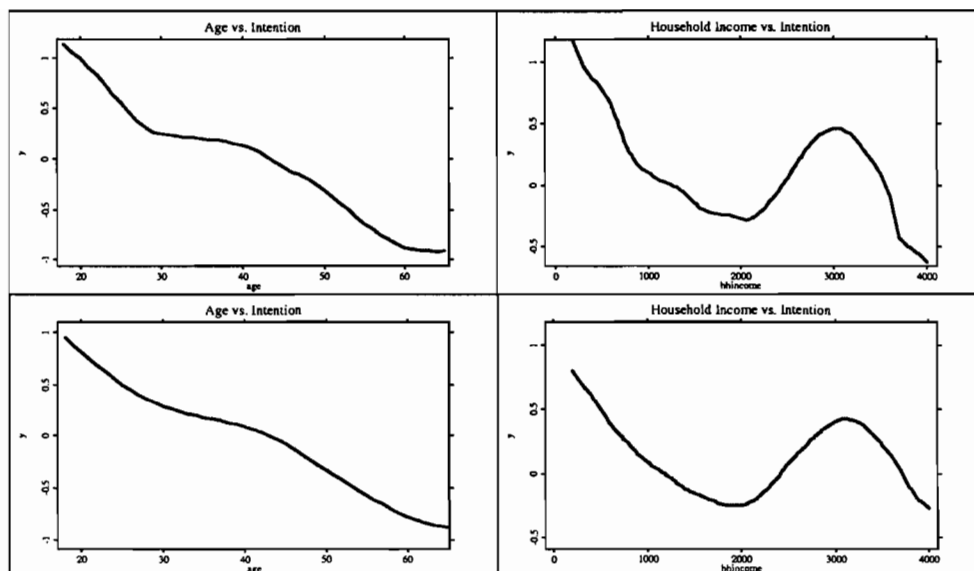
Figure 8: *The semiparametric estimates for the influence of* AGE *(left) and* HHINCOME *(right) in Sachsen. The upper plots were estimated with* $h = 0.75$, *the lower ones with* $h = 1.0$.

that the curve from Sax is more flat in the middle part. In contrast, for HHINCOME the curves from both countries have a totally different shape. On first glance one would guess that AGE could be modeled linearly, at least for M-V. This is less clear for HHINCOME.

In a third step we apply our bootstrap test for linearity to the variables AGE and HHINCOME. We always use 499 repetitions in the bootstrap resampling. The bandwidths are chosen as above. For the input AGE, linearity is always rejected for the 1 percent level, for all bandwidths in both countries. For the variable HHINCOME, the observed p-values are .16 [for $h = 1.0$, M-V], .14 [for $h = 1.25$, M-V], .02 [for $h = 0.75$, Sax], and .01 [for $h = 1.0$, Sax]. So the deviations for AGE from linearity are much more significant. At first sight, this seems to be surprising and even unreasonable because the plots for HHINCOME differ much more from linearity. Reasons are presumably that the estimates for HHINCOME have large variance and/or the model(s) is (are) misspecified, e.g. the link function $G(\cdot)$ is misspecified. The latter point could cause the somehow strange result for the impact of AGE in M-V since we do not know exactly what the bootstrap does when $G(\cdot)$ is (strongly) misspecified.

In a next step we construct uniform confidence bands. In Figure 9 95% uniform confidence bands are given for the impact functions for M-V. We use bandwidths $h = 1.25$, $h_1 = g = 1.1h$ and $B = 500$ bootstrap replications. In Figure 10 we give 95% confidence bands for Sax with $h = 0.75$, $h_1 = g = 1.1h$ and $B = 500$. All confidence bands contain a linear fit. Only for HHINCOME and country Sax the linear fit would lie on the boundary. So inference only based on uniform confidence bands would not lead to rejection of linear indices.

In a last step we test the specification of the link function. For the test statistic, see Section 3.4, we use $h = 0.75$, $h_1 = g = 1.25h$ for M-V and $h = 0.6$, $h_1 = g = 1.25h$ for Sax. For the bandwidth h_L we chose $h_L = c \cdot \hat{s}_I$ where $\hat{s}_I$ is an estimate of the standard deviation s_I of the index and where $c = 0.6$ to 0.75 for M-V and $c = 0.5$ to 0.6 for Sax. With $B = 499$ bootstrap
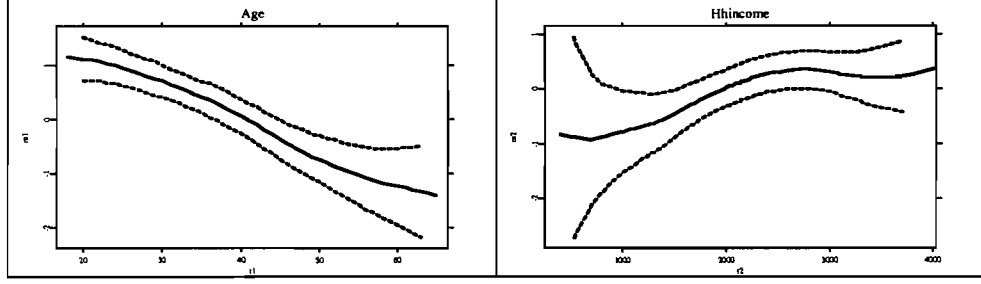
Figure 9: *Estimates and 95% uniform confidence bands for the impact of* AGE *(left) and* HHINCOME *(right) in M-V with* $h = 1.25$, $h_1 = g = 1.1h$.
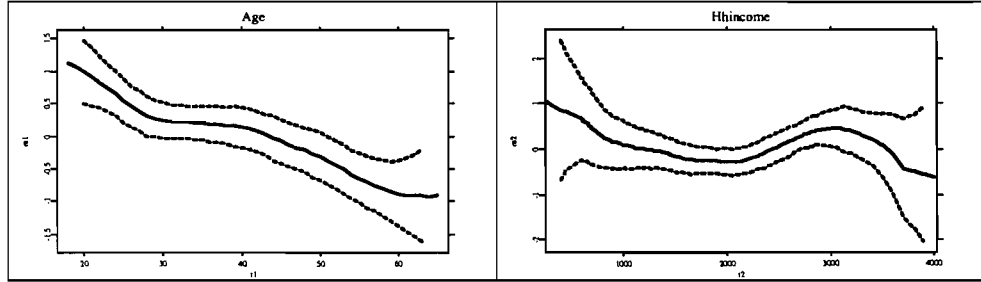

Figure 10: *Estimates and 95% uniform confidence bands for the impact of* AGE *(left) and* HHINCOME *(right) in Sax with* $h = 0.75$, $h_1 = g = 1.1h$.

replications we get p-values of about 7% for all bandwidths for M-V and p-values that are always larger than 15% for Sax. So we can indeed conclude that the slight inconsistency we found in the results for AGE in M-V might be caused by a misspecification of the link $G(\cdot)$.

A1   Assumptions

We now state the assumptions that are used in the results in Sections 2.1, 2.2 and Appendix A3. We use the notation

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

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

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

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

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

$$(A1.1) \qquad \lambda_i'(u) = \frac{Y_i - G(u)}{V[G(u)]} G'(u),$$

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

22

For our asymptotic expansions we use the following assumptions.

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

(A2) $E(Y|X,T) = G\{X^T\beta + m^+(T)\}$ with $\beta \in I\!R^p$. Here m^+ denotes the function $m^+(t) = \alpha + m_1(t_1) + \ldots + m_d(t_d)$, with $E\, m_j(T_{i,j}) = 0$ for $j = 1, \ldots, d$. The conditional variance $Var(Y_i|T_i = t)$ 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 $S_{T,1} \subset I\!R^r$ and $S_{T,-1} \subset I\!R^s$. T has a twice continuously differentiable density f_T with $\inf_{t \in S_T} f_T(t) > 0$.

(A4) For compact sets $B \subset I\!R^p$ and $H \subset I\!R$ we define

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

where, as above,

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

$\widehat{m}_\beta(t)$ is defined as

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

For $\beta \in B$ we put

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

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 $\eta \in H, t \in S_T, \beta \in B$

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

implies that

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

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

(A6) $m_1, \ldots, m_d$ are twice continuously differentiable functions from $I\!R^{q_j}$ to $I\!R$. 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 $t \in S_{T,-1}^- = \{s : \text{ there exists an } u \notin S_{T,-1} \text{ with } \|s - u\| \leq \delta\}$, $t \in S_{T,1}^- = \{s : \text{ there exists an } u \notin S_{T,1} \text{ with } \|s - u\| \leq \delta\}$ or $t \in S_T^- = \{s : \text{ there exists an } u \notin S_T \text{ with } \|s - u\| \leq \delta\}$, 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 $K(v) = K_1(v_1) \cdot \ldots \cdot K_r(v_r)$ and $L(v) = L_1(v_1) \cdot \ldots \cdot L_s(v_s)$. The kernels K_i and L_j are symmetric probability densities with compact support ($[-1,1]$, say).

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

(A9) The matrix $E\, Z^2 \tilde{X} \tilde{X}^T$ is strictly positive definite. The random vectors Z and $\tilde{X}$ have been defined in Theorems A2.1 and A2.2, respectively.

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

(A10) $m_1, \ldots, m_d$ 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)–(A6) contain boundedness conditions on covariables 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 estimate the minimization of the quasi likelihood 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.

A2 Asymptotic Theory for Estimation

This section contains asymptotic results on the marginal integration estimates $\widehat{m}_j$ and the parametric estimate $\widehat{\beta}$.

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

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

converges to a centered Gaussian variable with variance

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

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

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

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

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

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

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

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

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

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

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

where H_1 is a diagonal matrix with diagonal elements

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

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

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

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 $j = 2, \ldots, d$ can be done as the estimation of m_1 in Theorem 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 denotes always the bandwidth of the estimated component and g is chosen as bandwidth of the other components.] Then under these conditions the estimates $\widehat{m}_1(t_1), \ldots, \widehat{m}_d(t_d)$ are asymptotically independent. This leads to a multidimensional result. The random vector

$$\sqrt{nh}\begin{pmatrix} \widehat{m}_1(t_1) - m_1(t_1) - \delta_n^1(t_1) \\ \vdots \\ \widehat{m}_d(t_d) - m_d(t_d) - \delta_n^d(t_d) \end{pmatrix}$$

converges to a centered Gaussian variable with covariance matrix

$$\begin{bmatrix} \sigma_1(t_1) & 0 & \ldots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & \ldots & 0 & \sigma_d(t_d) \end{bmatrix}.$$

25

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

(A2.1)
$$\hat{\sigma}_1^2(t_1) = \sum_{i=1}^{n} \hat{\tau}_i^2,$$

where

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

(A2.2)
$$\kappa_i(t) = \frac{K_h(t_1 - T_{i,1})L_g(t_{-1} - T_{i,-1})}{\frac{1}{n}\sum_{j=1}^{n} K_h(t_1 - T_{j,1})L_g(t_{-1} - T_{j,-1})}$$

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

with

$$\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{[Y_i - \hat{\mu}_i]^2}{V(\hat{\mu}_i)}$$

and

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

The estimation of the nonparametric components yields also 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 un-smoothed global likelihood function in the definition of $\hat{\beta}$. Our conditions imply that $s + r \leq 3$. Again 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.

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

$$n^{1/2}\{\hat{\beta} - \beta\}$$

converges in distribution to $N(0, I^{-1})$ where Z^2 is defined as in Theorem A2.1 and where

$$
\begin{aligned}
I &= EZ^2\widetilde{X}\widetilde{X}^T \quad \text{with} \\
\widetilde{X} &= X - \{E(Z^2|T)\}^{-1}E(Z^2X|T).
\end{aligned}
$$

Our estimate of β does not make use of the additive structure of m. For this reason it is not efficient in our model (1.1). However it achieves the efficiency bound in the partial linear model $m(x;t) = G\{x^T\beta + \alpha + m(T_1, \ldots, T_d)\}$, see Mammen and van de Geer (1997) for calculation of the efficiency bound in this model. An estimate that takes care of additivity is given by

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

26

where $\widehat{m}_\beta^+(t)$ is defined as $\widehat{m}^+(t)$ with $\widehat{m}$ replaced by $\widehat{m}_\beta$ in equation (2.7). We expect that this estimate achieves higher efficiency. However this estimate has two drawbacks. Calculation of this estimate would need several nested iterative algorithm and is for this reason for larger data sets not feasible. Furthermore, the estimate is not robust against deviations from additivity.

Compared to $\widehat{\beta}$ root-n consistency of $\widehat{\alpha}$ requires additional conditions. The estimate $\widehat{\alpha}$ inherits by construction the biases of the nonparametric estimates $\widehat{m}$, $\widehat{m}_1$, ..., $\widehat{m}_d$. 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 $\widehat{\beta}$. On the other hand it can be checked that $\widehat{\alpha}$ has as $\widehat{\beta}$ asymptotic variance of order $O(n^{-1})$. Clearly, these remarks are not that essential. For most applications the parameter α has no direct interpretation.

A3 Proofs

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

A3.1 Proof of Theorem A2.1

We start by showing consistency of the estimate $\widehat{\beta}$:

(A3.1) $$\widehat{\beta} = \beta_0 + o_P(1).$$

For the proof of (A3.1) we show first that

(A3.2) $$\sup_{t,\beta} |\widehat{m}_\beta(t) - m_\beta(t)| = o_p(1).$$

Proof of (A3.2): For the proof of claim (A3.2) we show first that:

(A3.3) $$\sup_{\eta,t,\beta} |\Delta(m_\beta(t), t, \beta)| = O_p(\rho_2),$$

where the following notation has been used:

$$
\begin{aligned}
\Delta(\eta, t, \beta) &= \Delta_1(\eta, t, \beta) - \Delta_2(\eta, t, \beta), \\
\Delta_1(\eta, t, \beta) &= \frac{1}{n} \sum_i \lambda_i'(X_i^T \beta + \eta) \kappa_i(t), \\
\Delta_2(\eta, t, \beta) &= E\left[\lambda'(X^T \beta + \eta)|T = t\right], \\
\kappa_i(t) &= \frac{K_h(t_1 - T_{i,1}) L_g(t_{-1} - T_{i,-1})}{\frac{1}{n} \sum_{j=1}^{n} K_h(t_1 - T_{j,1}) L_g(t_{-1} - T_{j,-1})}.
\end{aligned}
$$

(A3.4)

For the proof of (A3.3) 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 (A2). By standard application of exponential inequalities we get for every $\nu_1 > 0$ that for C' large enough

(A3.5) $$P\{|\Delta(\eta, t, \beta)| > C'\rho_2\} = o(n^{-\nu_1}).$$

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 (A3.5), following the same argument as in Härdle and Mammen (1993), we obtain (A3.3).

For the proof of (A3.2), one can conclude from (A3.3) that, with probability tending to one, $\widehat{m}_\beta(t)$ lies in the interior of H, see (A4). This gives

(A3.6) $$\Delta_1(\widehat{m}_\beta(t), t, \beta) = 0.$$

With (A3.3) we obtain

$$\sup_{t,\beta} |\Delta_2(\widehat{m}_\beta(t), t, \beta)| = O_p(\rho_2).$$

With assumption (A4) this yields (A3.2).

We use (A3.2) to prove (A3.1) (consistency of $\widehat{\beta}$).

Proof of (A3.1). Let $k(\beta) = E[Q\{X^T\beta + m_\beta(T); Y\}]$. We will show that

(A3.7) $$\sup_{\beta \in B} \left| \frac{1}{n}\mathcal{L}(\widehat{m}_\beta, \beta) - k(\beta) \right| \to 0 \qquad \text{(in probability)}.$$

This implies claim (A3.1) since

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

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

It remains to prove (A3.7). This follows from

(A3.8) $$\sup_{\beta \in B} \left| \frac{1}{n}\mathcal{L}(m_\beta, \beta) - k(\beta) \right| \to 0 \qquad \text{(in probability)},$$

(A3.9) $$\sup_{\beta \in B} \left| \frac{1}{n}\mathcal{L}(\widehat{m}_\beta, \beta) - \frac{1}{n}\mathcal{L}(m_\beta, \beta) \right| \to 0 \qquad \text{(in probability)}.$$

Claim (A3.8) holds since $\mathcal{L}(m_\beta, \beta)/n$ 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

$$
\begin{aligned}
\left| \frac{1}{n}\mathcal{L}(m_{\beta_1}, \beta_1) - \frac{1}{n}\mathcal{L}(m_{\beta_2}, \beta_2) \right| &\leq T_{n,1}\|\beta_1 - \beta_2\| + T_{n,2}\sup_t |m_{\beta_1}(t) - m_{\beta_2}(t)| \\
&\leq T_{n,1}\|\beta_1 - \beta_2\| + T_{n,2}\sup_{t,\beta} \left\| \frac{\partial}{\partial \beta}m_\beta(t) \right\| \|\beta_1 - \beta_2\|,
\end{aligned}
$$

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 $\frac{\partial}{\partial\beta}m_\beta(t)$ is uniformly bounded in β and t note that

$$(A3.10) \qquad \frac{\partial m_\beta}{\partial\beta}(\beta,t) = -\frac{E[\lambda''\{\beta^T X + m_\beta(T)\}X|T = t]}{E[\lambda''\{\beta^T X + m_\beta(t)\}|T = t]}.$$

Equation (A3.10) follows by differentiation of $E\{\lambda'(\beta^T X + m_\beta(t))|T = t\} = 0$. This shows (A3.8). Claim (A3.9) follows from

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

Thus finally (A3.1) is shown.

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

$$(A3.11) \qquad \widehat{\beta} = \beta + \{E(Z^2 \widetilde{X}\widetilde{X}^T)\}^{-1}\frac{1}{n}\sum_{i=1}^n \widetilde{X}_i \,\lambda_i'\{X_i^T\beta + m^+(T_i)\} + O_p(\rho_2^2),$$

$$(A3.12) \qquad \sup_{t\in S_T^*}\left|\Delta(t)\right| = O_p(\rho_2^2),$$

with

$$
\begin{aligned}
\Delta(t) &= \widehat{m}(t) - \Bigg\{\overline{m}(t) \\
(A3.13) \qquad & \quad + \{E(Z^2|T = t)\}^{-1}\,E(Z^2 X^T|T = t)\,\{E(Z^2\widetilde{X}\widetilde{X}^T)\}^{-1} \\
& \quad \times \frac{1}{n}\sum_{i=1}^n \widetilde{X}_i\,\lambda_i'\{X_i^T\beta + m^+(T_i)\}\Bigg\},
\end{aligned}
$$

$$(A3.14) \qquad \overline{m}(t) = m^+(t) + \{E(Z^2|T = t)\}^{-1}\frac{1}{n}\sum_{i=1}^n \kappa_i(t)\,\lambda_i'\{X_i^T\beta + m^+(t)\},$$

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

$$(A3.15) \qquad \widetilde{X}_i = X_i - \{E[Z_i^2|T_i]\}^{-1}\,E[Z_i^2 X_i|T_i],$$

$$(A3.16) \qquad Z_i^2 = \frac{G'(X_i^T\beta + m^+(T_i))^2}{V[G(X_i^T\beta + m^+(T_i))]}.$$

Equations (A3.11) and (A3.12) follow from a slight modification of Lemma A3.3 and Corollary A3.4 in Härdle, Mammen and Müller (1998). There it has been assumed that the likelihood is maximized for β in a neighborhood of β_0 with radius ρ_1, see assumption (A7) in Härdle, Mammen and Müller (1998). In our set up we have that for a sequence δ_n' with $\delta_n' \to 0$ with probability tending to one

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

Using the same arguments as in Härdle, Mammen and Müller (1998), one can show that

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

This shows (A3.11). Equation (A3.12) can be shown similarly.

29

With the help of (A3.12) we arrive at

$$
\text{(A3.17)} \qquad \overline{m}_1(t_1) = \frac{\sum_{i=1}^{n} w_{-1}(T_{i,-1})\overline{m}(t_1, T_{i,-1})}{\sum_{i=1}^{n} w_{-1}(T_{i,-1})} + O_P(\rho_2^2 + n^{-1/2})
$$

$$
= m_1(t_1) + R_1 + \Delta_1(t_1) + O_P(\rho_2^2 + n^{-1/2}),
$$

where

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

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

where λ_j', κ_j and Z_i are defined by equations (A1.1), (A2.2) and (A3.16) 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

$$
nh\, Var(\Delta_1(t_1)|\mathcal{Z}_n)
$$

$$
\rightarrow \int L^2(u)\, du\, E\left[\frac{w^2(T_{-1})}{\{Ew_{-1}(T_{-1})\}^2} \frac{E(Z^2|T_1 = t_1)}{E(Z^2|T_1 = t_1, T_{-1})^2} \frac{f_{T_{-1}}^2(T_{-1})}{f_T^2(t_1, T_{-1})}\right].
$$

For this convergence, one uses for instance

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

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

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

$$
\text{(A3.18)} \qquad P\left(d_K\left(\mathcal{L}\left(\Delta_1(t_1) - E(\Delta_1(t_1)|\mathcal{Z}_n)\right), N(0, Var(\Delta_1(t_1)|\mathcal{Z}_n))\right) > \delta\right) \rightarrow 0
$$

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

$$
d_K(\mu, \nu) = \sup_{t \in \mathbb{R}} \left|\mu(X \leq t) - \nu(X \leq t)\right|.
$$

For the proof of (A3.18) 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

$$
\text{(A3.19)} \quad E(\Delta_1(t_1)|\mathcal{Z}_n) = \frac{1}{n} \sum_{i=1}^{n} \int K_h(t_1 - v_1)L_g(T_{i,-1} - v_{-1})
$$

$$
E\left[\{G(X^T \beta + m^+(v)) - G(X^T \beta + m^+(t_1, T_{i,-1}))\}\right.
$$

$$
\left. a^1(X, t_1, T_{i,-1})|T_{i,1} = t_1, T_{i,-1}\right] f_T(v)dv + r_n
$$

where the function a^1 is defined in Theorem A2.1, $r_n = O_P(\rho_2^2 + n^{-1/2}) + o_P(h^2 + g^2)$. Furthermore, $r_n = O_P(\rho_2^2 + n^{-1/2} + h^4 + g^4)$ under the additional assumption (A10). The proof of (A3.19) follows by standard, but tedious calculations. The asymptotic form of

30

$E(\Delta_1(t_1)|\mathcal{Z}_n)$ can be easily calculated from (A3.19). Note that the asymptotic bias of $\widehat{m}_1(t_1)$ 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, $\widehat{m}_1(t_1)$ and $\widetilde{m}_1(t_1)$ have the same asymptotic variance.

A3.2   Proof of Theorem A2.2

The conditions on h and g imply $\rho_2^2 = o(n^{-1/2})$. Therefore the statement of Theorem A2.2 can be followed from (A3.11).

A3.3   Proof of Theorem 3.1

The statement of the theorem follows from

(A3.20) $$2\widehat{m}_1(t_1) - E^*\widehat{m}_1^*(t_1) - m_1(t_1) = O_P(h^4 + g^4 + (nh)^{-1/2}).$$

Claim (A3.20) follows from

(A3.21) $$2\overline{m}_1(t_1) - E^*\,\overline{m}_1^*(t_1) - m_1(t_1) = R_1 - \hat{R}_1 + O_P(h^4 + g^4 + (nh)^{-1/2}),$$

(A3.22) $$\frac{1}{n}\sum_{i=1}^{n} w_1(T_{i,1})\left[2\overline{m}_1(T_{i,1}) - E^*\,\overline{m}_1^*(T_{i,1}) - m_1(T_{i,1})\right]$$

$$= \left[R_1 - \hat{R}_1\right]\frac{1}{n}\sum_{i=1}^{n} w_1(T_{i,1}) + O_P(h^4 + g^4 + (nh)^{-1/2}),$$

where

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

and where R_1 has been defined after (A3.17).

We give only the proof of (A3.21). Claim (A3.22) follows similarly. By (A3.17) 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

$$D_1(t_1) = \frac{1}{\sum_{i=1}^{n} w_{-1}(T_{i,-1})}\frac{1}{n}\sum_{i,j=1}^{n} \frac{w_{-1}(T_{i,-1})\kappa_j(t_1, T_{i,-1})}{E(Z_i^2|T_{i,1} = t_1, T_{i,-1})}\frac{G'\{X_j^T\beta + m^+(t_1, T_{i,-1})\}}{V(G\{X_j^T\beta + m^+(t_1, T_{i,-1})\})}$$

$$\left[G\{X_j^T\beta + m^+(T_j)\} - G\{X_j^T\beta + m^+(t_1, T_{i,-1})\}\right].$$

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

$$\hat{D}_1(t_1) = \frac{1}{\sum_{i=1}^{n} w_{-1}(T_{i,-1})}\frac{1}{n}\sum_{i,j=1}^{n} \frac{w_{-1}(T_{i,-1})\kappa_j(t_1, T_{i,-1})}{E(Z_i^2|T_{i,1} = t_1, T_{i,-1})}\frac{G'\{X_j^T\hat{\beta} + \widehat{m}^+(t_1, T_{i,-1})\}}{V(G\{X_j^T\hat{\beta} + \widehat{m}^+(t_1, T_{i,-1})\})}$$

$$\left[G\{X_j^T\hat{\beta} + \widehat{m}^+(T_j)\} - G\{X_j^T\hat{\beta} + \widehat{m}^+(t_1, T_{i,-1})\}\right].$$

For claim (A3.21) it suffices to show

(A3.23)
$$D_1(t_1) - \hat{D}_1(t_1) = O_P(h^4 + g^4 + (nh)^{-1/2}).$$

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

$$D_1(t_1) - \hat{D}_1(t_1) = \sum_{i,j=1}^{n} W_{i,j} \left[G\{X_j^T \beta + m^+(T_j)\} - G\{X_j^T \beta + m^+(t_1, T_{i,-1})\} \right.$$

(A3.24)
$$\left. -G\{X_j^T \hat{\beta} + \hat{m}^+(T_j)\} + G\{X_j^T \hat{\beta} + \hat{m}^+(t_1, T_{i,-1})\} \right]$$
$$+ O_P(h^4 + g^4 + (nh)^{-1/2}),$$

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 (A3.24) can be treated by using Taylor expansions of G and the stochastic
expansions of $\hat{m}_j$ given in (A3.17). Consider e.g. for $k \neq 1$

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

Then by using the expansions of $\hat{m}_k$ given in (A3.17) and the expansion of the bias of $\hat{m}_k$
[see Theorem 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

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

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 $\omega_{i,n}(\mathcal{Z}_n, t_1)$:

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

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.

A3.4   Proof of Theorem 3.2

We consider the statistic

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

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

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

We will show that

(A3.25) $$U = V + o_p(h^{-1/2}),$$
(A3.26) $$R = U + o_p(h^{-1/2}),$$

where

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

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

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

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

The function a^1 has been defined in the statement of Theorem 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]

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

Because v_n^2 is of order h^{-1} for the proof of the theorem it remains to show (A3.25) and (A3.26).

Proof of (A3.25). Because $\rho_2^2 = o(n^{-1/2})$, it follows from (A3.12) [compare (A3.17)] that uniformly for t_1 in $S_{T,1}^-$:

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

where

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

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

33

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 Theorem A2.1 one can show that this implies the following uniform expansion of $\widehat{m}_1$.

(A3.27) $\qquad \widehat{m}_1(t_1) = \gamma_1 t_1 + \widehat{m}_1^{APPR1}(t_1) + \widehat{m}_1^{APPR2}(t_1) + \delta_n^1(t_1) + o_P(n^{-1/2}),$

where

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

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 Theorem A2.1.

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

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

with

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

for some uniformly bounded functions $\omega_{i,n,3}$.

Together with (A3.27) and a stochastic expansion of $\tilde{\gamma}$ this gives that uniformly for t_1 in $S_{T,1}^-$:

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

with

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

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

Claim (A3.25) follows from

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

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

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

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

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

Proof of (A3.26). Because of Theorem A2.2, we have that $\widehat{\beta} - \beta = O_P(n^{-1/2})$ and $\widehat{\alpha} - \alpha = O_P(n^{-1/2})$. Moreover we can easily show that

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

It follows that

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

Now,

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

This proves (A3.26).

A3.5   Proof of Theorem 3.3

This theorem follows by replication of the arguments in the proof of the last theorem for the "Bootstrap world".

A3.6   Proof of Theorem 3.6

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

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

Here, for $\delta > 0$ small enough we have put $S_{T,1}^- = \{s : \text{there exists an } u \notin S_{T,1} \text{ with } |s - u| \le \delta\}$. [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

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

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

We will give now 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 $\mathcal{Z}_n = ((X_1, T_{1,1}, \ldots, T_{1,d}), \ldots, (X_n, T_{n,1}, \ldots, T_{n,d}))$. 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 $Y_1^+, Y_1^{++}, \ldots, Y_1^+, Y_1^{++}, \ldots, Y_n^+, Y_n^{++}$ by the following construction: choose an i.i.d. sample $U_1, \ldots, U_n$ 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 $(X_1, T_1, Y_1), \ldots, (X_n, T_n, Y_n)$, $(Y_1^+, Y_1^{++}), \ldots, (Y_n^+, Y_n^{++})$ 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

$$(A3.28) \qquad \max_{1 \le i \le n} E^* |Y_i^{++} - Y_i^+| = O_P(\rho_2).$$

Here E^* denotes the conditional expectation given the original data $(X_1, T_1, Y_1), \ldots, (X_n, T_n, Y_n)$. 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 (A3.28) follows from

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

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 $Y_1^+, \ldots, Y_n^+$ is denoted by $\hat{m}_1^+(t_1)$. The estimate that is based on $Y_1^{++}, \ldots, Y_n^{++}$ is denoted by $\hat{m}_1^{++}(t_1)$.

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

$$(A3.29) \qquad \max_{1 \le i \le n} \sup_{0 \le t \le \tau} E^* |\varepsilon_i^{++} - \varepsilon_i^+|^2 \left\{ 1 + \exp(t|\varepsilon_i^+|) + \exp(t|\varepsilon_i^{++}|) \right\} = O_P(\rho_2).$$

This can be seen by straight forward calculations using (A3.28) 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 (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 3.6 follow from

$$(A3.30) \qquad \sup_{t_1 \in S_{T,1}^-} |\hat{\sigma}_1(t_1) - \sigma_1(t_1)| = o_P(1),$$

$$(A3.31) \qquad \sup_{t_1 \in S_{T,1}^-} |\hat{\sigma}_1^*(t_1) - \sigma_1(t_1)| = o_P([\log n]^{-1}),$$

$$(A3.32) \qquad \sup_{t_1 \in S_{T,1}^-} \left| \left[\hat{m}_1^{++}(t_1) - \hat{m}_1(t_1)\right] \right.$$
$$\left. - \left[\hat{m}_1^+(t_1) - m_1(t_1)\right] \right| = o_P((nh)^{-1/2}[\log n]^{-1/2}).$$

We give here only the proof of (A3.32). One shows first that

$$\sup_{t_1 \in S_{T,1}^-} |\hat{m}_1^+(t_1) - m_1(t_1) - \frac{1}{n} \sum_{i=1}^n a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1}) \varepsilon_i^+|$$
$$= o_P((nh)^{-1/2}[\log n]^{-1/2}),$$

$$\sup_{t_1 \in S_{T,1}^-} |\hat{m}_1^{++}(t_1) - \hat{m}_1(t_1) - \frac{1}{n} \sum_{i=1}^n a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1}) \varepsilon_i^{++}|$$
$$= o_P((nh)^{-1/2}[\log n]^{-1/2}).$$

This can be done by using expansions of the type (A3.12). 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 (A3.32) it remains to show

(A3.33)
$$\sup_{t_1 \in S_{T,1}^-} |\frac{1}{n} \sum_{i=1}^n a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1})[\varepsilon_i^+ - \varepsilon_i^{++}]|$$
$$= o_P((nh)^{-1/2}[\log n]^{-1/2}).$$

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

(A3.34)
$$\sup_{t_1 \in S_{T,1}^-} P^* \left\{ |\frac{1}{n} \sum_{i=1}^n a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1})[\varepsilon_i^+ - \varepsilon_i^{++}]| > C_2 \kappa_n \right\}$$
$$= o_P(n^{-C_1}),$$

where $\kappa_n[nh/\rho_1]^{-1/2}[\log n]^{3/2}$ and where P^* denotes the conditional distribution given 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 (A3.34) shows that (A3.33) if the supremum runs over a finite set with $O(n^{C_1})$ elements. This implies (A3.33) 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 (A3.33). Note that

$$P^* \left\{ \frac{1}{n} \sum_{i=1}^n a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1})[\varepsilon_i^+ - \varepsilon_i^{++}] > C_2 \kappa_n \right\}$$

$$\leq E^* \exp \left[\log n \kappa_n^{-1} \frac{1}{n} \sum_{i=1}^n a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1})[\varepsilon_i^+ - \varepsilon_i^{++}] \right] \exp[\log n \kappa_n^{-1} C_2 \kappa_n]$$

$$\leq n^{-C_2} \prod_{i=1}^n E^* \exp \left[\frac{\log n}{\kappa_n n} a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1})[\varepsilon_i^+ - \varepsilon_i^{++}] \right].$$

We use now the expansion $\exp[x] \leq 1 + x + x^2/2 \{1 + \exp[x]\}$. Because of $E^* \varepsilon_i^+ - \varepsilon_i^{++} = 0$ and because of (A3.29) this gives that the last term is bounded by

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

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

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

With another constant C' this can be bounded by

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

For C_2 large enough, this is of order $o(n^{C_1})$. This shows (A3.33).

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.

Bickel, P. and Rosenblatt, M. (1973) On some global measures of the deviations of density function estimates. *Ann. Statist.* 1, 1071 - 1095.

Burda, M. (1993) The determinants of East–West German migration. *European Economic Review* 37, 452 - 461.

Burda, M., Härdle, W., Müller, M. and Werwatz, A. (1998) Semiparametric Analysis of German East-West Migration Intentions: Facts and Theory. *Journal of Applied Econometrics* 13, 545 - 541.

Carroll, R.J., Fan, J., Gijbels, I, and Wand, M.P. (1997) Generalized partially linear single-index models. The University of New South Wales, Australian Graduate School of Management Working Paper Series, No. 95-010.

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

Eubank, R. L. and Speckman, P. L. (1993) Confidence bands in nonparametric regression. *J. Amer. Statist. Assoc.* 88, 1287 - 1301.

Fan, J., Härdle, W. and Mammen, E. (1998) Direct estimation of low dimensional components in additive models. *Ann. Statist.* 26, 943 - 971.

Fuss, M., McFadden, D. and Mundlak, Y. (1978) A survey of functional forms in the economic analysis of production. *In Fuss, M. and McFadden, D. (eds.)*, Production Economics: A Dual Approach to Theory and Applications, 1 219 - 268.

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

Gozalo, P.L. and Linton, O.B. (1999) Testing additivity in generalized nonparametric regression models. *Preprint, Yale University, USA*

Härdle, W. and Mammen, E. (1993) Testing parametric versus nonparametric regression. *Ann. Statist.* 21, 1926 -1947.

Härdle, W., Mammen, E. and Müller, M. (1998) Testing parametric versus semiparametric modelling in generalized linear models. *J. Amer. Statist. Assoc.* 93, 1461 - 1474.

Härdle, W., Mammen, E. and Proenca, I. (2000) A bootstrap test for single index models. *Preprint, University of Heidelberg.*

Härdle, W., Spokoiny, V. and Sperlich, S. (1997) Semiparametric single index versus fixed link function modelling. *Ann. Statist.* 25, 212 - 243.

Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized additive models.* Chapman and Hall, London.

Horowitz, J. (1998) Semiparametric methods in econometrics. Lecture Notes in Statistics **131**, Springer, Heidelberg, Berlin, New York.

Horowitz, J. (1999) Nonparametric estimation of a generalized additive model with an unknown link function. Preprint, Iowa State University.

Horowitz, J. and Härdle, W. (1994) Testing a parametric model versus a semiparametric alternative. *Econometric Theory.* **10**, 821 - 848.

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

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

Linton, O. B. and Nielsen, J.P. (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.* Lectures Notes in Statist. **77**, Springer, Heidelberg, Berlin, New York.

Mammen, E. and van de Geer, S. (1997) Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.* **25**, 1014 - 1035.

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

McCullagh, P. and Nelder, J.A. (1989) *Generalized linear models.* Chapman and Hall, London.

Müller, M. (2000) Estimation and Testing in Generalized Partial Linear Models - A Comparative Study. *to appear in Statistics and Computing*

Neumann, M. and Polzehl, J. (1998) Simultaneous bootstrap confidence bands in nonparametric regression. *J. Nonpar. Statist.* to appear

Opsomer, J. D. and Ruppert, D. (1997) Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* **25**, 186 - 211.

Severini, T. A. and Staniswalis, J. G. (1994) Quasi-Likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.* **89**, 501 - 511.

Severini, T. A. and Wong, (1992) Generalized profile likelihood and conditionally parametric models. *Ann. Statist.* **20**, 1768 - 1802.

Sperlich, S., Linton, O. B. and Härdle, W. (1999) Integration and Backfitting Methods in Additive Models - Finite Sample Properties and Comparison. *Test* **8**, 419 - 458.

Sperlich, S., Tjøstheim, D. and Yang, L. (1999) Nonparametric Estimation and Testing of Interaction in Additive Models. *Working Paper 99-85 (34), Carlos III de Madrid*

Stone, C.J. (1985) Additive regression and other nonparametric models. *Ann. Statist.* **13**, 685 - 705.

Stone, C.J. (1986) The dimensionality reduction principle for generalized additive models. *The Annals of Statistics* **14**, 592 - 606.

Tjøstheim, D. J. and Auestadt, B. H. (1994) Nonparametric identification of nonlinear time series: projections. *J. Amer. Statist. Assoc.* **89**, 1398 - 1409.

Wu, C.F.G. (1986) Jackknife, bootstrap and other resampling methods in regression analysis. (with discussion) *Ann. Statist.* **14**, 1291 - 1380.