Semiparametric additive indices for binary response and generalized additive models

Wolfgang HÄRDLE

Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät,

Humboldt-Universität zu Berlin, D 10178 Berlin, Germany

Sylvie HUET

Institut de Recherche Agronomique, Centre de Recherches de Jouy-en-Josas

F 78352 Jouy-en-Josas Cedex, France

Enno MAMMEN

Institut für Angewandte Mathematik, Ruprecht-Karls-Universität Heidelberg

Im Neuenheimer Feld 294, D 69120 Heidelberg, Germany

Stefan SPERLICH

Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät,

Humboldt-Universität zu Berlin, D 10178 Berlin, Germany

November 9, 1998

Abstract

Models are studied where the response Y and covariates X, T are assumed to fulfill $E(Y|X;T) = G\{X^T\beta + \alpha + m_1(T_1) + \ldots + m_d(T_d)\}$. Here G is a known (link) function, β is an unknown parameter, and $m_1, \ldots, m_d$ are unknown functions. In particular, we consider additive binary response models where the response Y is binary. In these models, given X and T, the response Y has a Bernoulli distribution with parameter $G\{X^T\beta + \alpha + m_1(T_1) + \ldots + m_d(T_d)\}$. The paper discusses estimation of β and $m_1, \ldots, m_d$. Procedures are proposed for testing linearity of the additive components $m_1, \ldots, m_d$. Furthermore, bootstrap uniform confidence intervals for the additive components are introduced. The practical performance of the proposed methods is discussed in simulations and in two economic applications. [1]

1 Introduction

Many problems in applied econometrics and other fields require estimating the conditional mean of a random response Y given random covariates. Assume that the covariate vector is decomposed in two components (X, T). This paper is concerned with estimating the conditional mean $m(x, t) = E(Y | X = x; T = t)$. We will assume that the influence of X is linked linearly to $m(x, t)$. The influence of T will be described by additive nonparametric functions of the components of the vector T (Generalized Additive Regression). We will discuss construction of tests and confidence bands for these nonparametric functions.

A traditional estimation approach for $m(x; t)$ begins by assuming that m belongs to a known finite-dimensional parametric family in the class of generalized linear models. That is, $m(x; t) = G(x^T \beta + \alpha + t^T \gamma)$ for a known link function G and a linear parametric index $(x^T \beta + \alpha + t^T \gamma)$. If the true relationship between (X, T) and Y is given by such a generalized linear model then the parameters can be estimated with $O_P(n^{-1/2})$ rates of convergence. The estimated parameter, though, can be misleading if $m(x; t)$ is misspecified. The possibility of misspecification may be eliminated by a non- or semiparametric approach at the cost of less precise statistical estimation and additional numerical burden. Bierens (1987), Härdle (1990) provide overviews over the nonparametric estimation methods and discuss the issue of rates of convergence. An excellent introduction into semiparametrics in econometrics is given in Horowitz (1998). The nonparametric rate of convergence decreases rapidly as the dimension of the covariables increases (Stone (1983), Silverman (1986, Table 4.20)). The rate of convergence may be improved through the use of dimension reducing methods. One popular method is the assumption of additivity for the nonparametric components. The subject of this paper are tests and confidence bands in generalized additive regression where the influence of the X variable is kept linearly and the influence of T is modelled in an additive nonparametric way. In these models the response Y and covariates X, T are assumed to fulfill

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

Here G is a known (link) function, β is an unknown parameter, and $m_1, \ldots, m_d$ are unknown functions. This model is a semiparametric generalisation of the generalized linear model where the conditional expectation of the response depends on all covariates via the link function G in a linear way, i.e. $E(Y | X; T) = G\{X^T \beta + \alpha + T^T \gamma\}$ with an additional parameter γ. Models of this type are logit and probit models that are widely used in mobility analysis, employment studies, marketing analysis, credit scoring and many other fields. They are often applied because they allow a simple interpretation of a "linear index" and software is routinely widely accessible.

Appropriateness of linearity in these index models has been questioned in recent applications. Burda (1993) analysed East - West migration in Germany; Fahrmeir and

Hamerle (1984), Fahrmeir and Tutz (1994) used logit models in credit scoring and found nonlinear influences in the predictor variables. Bertschek (1996) and Horowitz and Härdle (1996) analysed innovative behavior of firms and proposed non- and semi-parametric approaches which are shown to be a valuable alternative to linear index modelling. Severini and Staniswalis (1994), Ai (1997), Ai and McFadden (1997) demonstrated how parametric and nonparametric components can be estimated efficiently in case of one nonparametric component. Their approach is based on an iterative application of smoothed local and un-smoothed global likelihood functions. For a related model with semiparametric index see Carroll, Fan, Gijbels and Wand (1995). A nonparametric bootstrap test for the parametric index can be found in Härdle, Mammen and Müller (1998). In this paper we improve upon this earlier work by considering several additive nonparametric components and by constructing confidence bands for these components.

The additive modelling has been analysed theoretically for high-dimensional regression data, see Stone (1985, 1986), Andrews and Whang (1990), Newey (1994). It helps to circumvent the curse of high dimension: (i) The model can be estimated at a rate typical for one dimensional explanatory variables. (ii) The resulting curves are one-dimensional and can be inspected graphically with the aid, e.g. of uniform confidence bands. Two practical proposals exist for the estimation of additive components in regression models. Projection smoothers using backfitting techniques have been considered in Buja, Hastie and Tibshirani (1989). Asymptotic theory for this iterative technique is rather complicated, see Linton, Mammen and Nielsen (1998), Opsomer (1997) and Opsomer and Ruppert (1997). Tools (e.g. tests and confidence bands) for statistical inference based on the estimates are rare and there is no complete mathematical knowledge on the choice of the bandwidth. Recently, an "integration" technique of additive components has been introduced by Tjøstheim and Auestad (1994), Linton and Nielsen (1995). The technical treatment of this method is simple and allows an asymptotic distribution theory. This approach has been applied in regression by Fan, Härdle and Mammen (1998), Severance-Lossin and Sperlich (1997) and in time series analysis by Masry and Tjøstheim (1995,1997). For generalized additive models this method has been discussed in Linton and Härdle (1996). Linton (1997) proposed a modification that achieves certain oracle bounds. For a simulation comparison of both approaches see Sperlich, Linton and Härdle (1997). Horowitz (1997) provides an estimation technique for a purly additive index with unknown link.

In this paper we study bootstrap tests and confidence bands that are based on integration estimates. The paper is organised as follows. In the next section we introduce integration estimates for additive binary choice models. Section 3 generalizes this discussion to generalized additive models and it states asymptotics for integration estimates. Typically, the bias of the integration estimate depends on the shape of *all* additive components. This complicates the data analytic interpretation of estimated nonparametric components. We will show how bootstrap can be used to correct for

the bias. Section 4 introduces bootstrap tests for testing linearity of additive compo-
nents. The tests are modifications of an approach of Hastie and Tibshirani (1990).
They proposed to use the likelihood ratio test and to take critical values of a χ^2 ap-
proximation. 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 com-
pare the nonparametric fit with an bootstrap estimate of its expectation [under the
hypothesis of linearity]. Without this bias correction the test does not behave like an
overall test, see Härdle and Mammen (1993) for a similiar discussion in a simple re-
gression model. Our second modification takes care of the fact that different likelihood
functions [smoothed and unsmoothed 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. Section 5 presents theory for uniform confidence bands of nonpara-
metric additive components. Again, their construction uses bootstrap. In Section 6
the presented methodology is applied to a migration problem and to a labour market
problem. This section also includes a small simulation study. Assumptions and proofs
are postponed to the appendix.

2 Estimation in 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 is a random variable in $I\!R^d$, X_i is in $I\!R^p$ and Y_i is a binary response. Condition-
ally 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 un-
known parameter in $I\!R^p$, and $m_1, \ldots, m_d$ are unknown functions $I\!R \to I\!R$. The param-
eter α is in $I\!R$. 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 $w_1, .., w_d$. Given (X_i, T_i), the (con-
ditional) 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 discuss now how the additive components $m_1, \ldots, m_d$ can be estimated. Without
loss of generality, we will do this only for the first component m_1. Define the smoothed
likelihood

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

where for a vector $u \in I\!\!R^d$ we denote the vector $(u_2, \ldots, u_d)^T$ by u_{-1}. Similarly, $T_{i,-1} = (T_{i,2}, \ldots, T_{i,d})^T$. For a kernel function L defined on $I\!\!R^{d-1}$ put $L_g(v) = g^{-(d-1)}L(g^{-1}v)$ and for a kernel function K defined on $I\!\!R$ put $K_h(v) = h^{-1}K(h^{-1}v)$, for L take the product kernel $L = \prod_{j=1}^{d-1} L_j$. The bandwidth g is related to smoothing in direction of the "nuisance" covariates. The relative speed of g to h and the choice of these bandwidths will be presented later. We define now 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) these estimates are based on an iterative application of smoothed local and un-smoothed global likelihood functions. We define for $\beta \in B$

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

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

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

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 makes no use of the additive structure of m^+. This $\widehat{m}$ will be used in an additional step as an auxiliary quantity for obtaining 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. Firstly, it may be useful to avoid problems at the boundary. Secondly, it can be chosen to minimize the asymptotic variance. In particular, for a regression model (without link function) it has been shown in Fan, Härdle and Mammen (1998) that after appropriate choice of w_{-1} a component m_1 can be estimated with the same asymptotic bias and variance as if the other components $m_2, \ldots, m_d$ were known. For a weight function w_{-1} define

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

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

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

The additive constant α is estimated by

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

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

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

Asymptotics of $\widehat{m}_1$ will be discussed in the next section for the general case of generalized additive models. We come back to binary choice models in Section 6 where some simulations will be presented and where the methods will be applied to economic data.

3 Estimation in generalized additive models: asymptotics, bootstrap bias correction

We come now 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

$$Q(\mu; y) = \int_\mu^y \frac{(s-y)}{V(s)}\, ds \tag{3.1}$$

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 $Y_1, ..., Y_n$ 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 because we want 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 (3.1). The estimates $\widehat{m}_\beta$, $\hat\beta$, $\widehat{m}$, $\overline{m}_1$, $\widehat{m}_1$, $\widehat{m}^+$ and $\hat\alpha$ are defined as in (2.4) - (2.8). Asymptotics for $\widehat{m}_1$ are presented in the following theorem. The assumptions can be found in Appendix A1.

Theorem 3.1
Suppose that the assumptions (A1) - (A9) apply. Then if h and g tend to zero and $nhg^{2(d-1)}(\log n)^{-2}$ tends to infinity,

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

converges to a centered Gaussian variable with variance

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

where $f_{T_{-1}}$ and f_T are the densities of T_{-1} or $T = (T_1, T_{-1})$, respectively. [For a vector $(v_1, \ldots, v_d$ 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:

$$
\begin{aligned}
Z_1 &= w_1^2(T_{-1})\frac{Z^2}{V[G\{X^T\beta + m^+(T)\}]}f_{T_{-1}}^2(T_{-1})Var(Y|X,T),\\
Z_2 &= E\left[Z^2\,\middle|\, T_1 = t_1, T_{-1}\right]^2 f_T^2(t_1, T_{-1}),\\
Z^2 &= \frac{G'(X^T\beta + m^+(T))^2}{V[G\{X^T\beta + m^+(T)\}]}.
\end{aligned}
$$

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

$$
\begin{aligned}
d_n^1(t_1) &= g^2 \int_{I\!\!R^{d-1}} E\left[a^1(X, t_1, u)\sum_{j=2}^d \sigma_{L,j}^2 b_j(X, t_1, u)\,|T = (t_1, u)\right] f_{T_{-1}}(u)du\\
&\quad + h^2 \int_{I\!\!R^{d-1}} E\left[a^1(X, t_1, u)\sigma_K^2 b_1(X, t_1, u)\,|T = (t_1, u)\right] f_{T_{-1}}(u)du.
\end{aligned}
$$

7

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

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

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

Under the additional assumption of (A10) the rest term $o_P(h^2 + g^2)$ in the expansion of $\delta^1_n(t_1)$ can be replaced by $O_P(h^4 + g^4)$.

The optimal rate of convergence for twice differentiable functions m_1 is $n^{-2/5}$. As long as second order kernels K and L are used this rate can be achieved under the assumptions of Theorem 3.1 only for $d \leq 2$. For higher dimensions d, one can see from our expansions that the $n^{-2/5}$ rate can be achieved by using higher order kernels $L_1, \ldots, L_{d-1}$. Furthermore, it can be shown that Theorem 3.1 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 $\widehat{m}$.

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 3.1.If assumptions analogous to (A1) - (A9) [(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.] One sees that 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

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

converges to a centered Gaussian variable with covariance matrix

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

The variance of the estimate $\widehat{m}_1(t_1)$ can be estimated by

$$(3.2) \qquad \hat{\sigma}^2_1(t_1) = nh \sum_{i=1}^n \hat{\tau}^2_i,$$

where

$$\hat{\tau}_i = \left[\frac{1}{n}\sum_{j=1}^n w_1(T_{j,-1}) \right]^{-1} \frac{1}{n}\sum_{j=1}^n w_1(T_{j,-1})\kappa_j(t_1, T_{i,-1})$$

8

$$\left[\frac{1}{n}\sum_{l=1}^{n}\frac{G'(X_l^T\hat{\beta}+\widehat{m}^+(T_l))}{V[G\{X_l^T\hat{\beta}+\widehat{m}^+(T_l)\}]}\kappa_l(t_1,T_{j,-1})\right]$$

$$\frac{G'(X_i^T\hat{\beta}+\widehat{m}^+(t_1,T_{j,-1}))}{V[G\{X_i^T\hat{\beta}+\widehat{m}^+(t_1,T_{j,-1})\}]}\hat{s}_i,$$

$$(3.3)\qquad \kappa_j(t) = \frac{K_h(t_1-T_{i,1})L_g(t_{-1}-T_{i,-1})}{\frac{1}{n}\sum_{j=1}^{n}K_h(t_1-T_{j,1})L_g(t_{-1}-T_{j,-1})}$$

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

with

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

and

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

Theorem 3.1 shows that if the bandwidths h and g are of the same order, the bias of $\widehat{m}_1(t_1)$ depends on the shape of the other additive components $m_2,\ldots,m_d$. This may lead to wrong interpretations of the estimate $\widehat{m}_1$. The bootstrap bias estimates help here to judge such effects.

Three versions of bootstrap will be considered here [see also Mammen and van de Geer (1997), Härdle, Mammen and Müller (1998)]. 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 setups. 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 $i=1,\ldots,n$.

Under the additional model assumption

$$Var(Y|X,T) = \sigma^2 V(G(X^T\beta_0+m^+(T)))$$

(Model B) one may use a resampling scheme that takes care of this relation. For this reason, we propose to 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 $E(e^{[|\varepsilon_i^*|/C]}) \leq C$ 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 propose to 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 it is reasonable to resample from the Bernoulli distribution with parameter $\hat{\mu}_i$.

In all three resampling schemes, one uses the data (X_1, T_1, Y_1^*), ..., (X_n, T_n, Y_n^*) to calculate the estimate $\widehat{m}_1^*$. 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 $\widehat{m}_1(t_1)$ is given by $E^* \widehat{m}_1^*(t_1)$, where E^* denotes the conditional expectation given the sample (X_1, T_1, Y_1), ..., (X_n, T_n, Y_n). The bias corrected estimate of $m_1(t_1)$ is defined by

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

where $\widehat{\delta}_n^1(t_1) = E^* \widehat{m}_1^*(t_1) - \widehat{m}_1(t_1)$. The next theorem shows that the bias terms of order g^2 are removed by this construction.

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) 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 h and g tend to zero and that $nhg^{2(d-1)}(\log n)^{-2}$ tends to infinity. Then it holds that

(3.4) $$\widehat{m}_1^B(t_1) - m_1(t_1) = O_p\{h^4 + g^4 + (nh)^{-1/2}\}.$$

For application of bootstrap in nonparametric regression it has been proposed to generate the bootstrap samples from another estimate of the regression function. Suppose e.g. that in the third step of the bootstrap algorithm $\hat{\mu}_i$ is replaced by $G\{X_i^T \hat{\beta} + \hat{\alpha} + \widehat{m}_1^O(T_{i,1}) + \widehat{m}_2(T_{i,2}) + \ldots + \widehat{m}_d(T_{i,d})\}$, where $\widehat{m}_1^O$ is defined as $\widehat{m}_1$ but with bandwidth h^O instead of h. Then if $h^O/h \to \infty$ one can show that the left hand side of (3.4) is of order $O_p\{(h^O)^4 + g^4 + (nh^O)^{-1/2}\}$. Under weak conditions on h^O and g this is of order $o_P\{(nh)^{-1/2}\}$, i.e. $\widehat{m}_1^B(t_1)$ has no bias of first order. Using this fact it can be shown that under the assumptions of Theorem 3.2 the unconditional distribution of $\widehat{m}_1(t_1) - m_1(t_1)$ and the conditional distribution of $\widehat{m}_1^*(t_1) - \widehat{m}_1^O(t_1)$ have the same normal limit, i.e. the distribution of $\widehat{m}_1(t_1) - m_1(t_1)$ is consistently estimated by the bootstrap.

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 $\widehat{\beta}$. Our conditions imply that $d \leq 3$. Again this constraint can be weakened by assumption of higher order smoothness of $m_1, \ldots, m_d$ and by use of higher order kernels.

Theorem 3.3

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:

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

converges in distribution to $N(0, I^{-1})$ where Z^2 is defined as in Theorem 3.1 and where

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

4 Bootstrap tests for linearity of additive components.

Interesting shape characteristics may be visible in plots of estimates of additive components. The complicated nature of the model may make it difficult to judge the statistical significance of such findings. A first test would be a comparison of the nonparametric estimates with linear functions. Deviance of the estimates from linear functions may give an indication on the significance of appearing shape characteristics. The hypothesis of interest is therefore:

$$(4.1) \qquad\qquad m_1(t_1) = \gamma_1\, t_1 \quad \text{for all } t_1 \text{ and a scalar } \gamma_1.$$

Our test is a modification of a general test approach described in 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 derived by calculating the expectation of asymptotic expansions of the test statistic under the null hypothesis. For this approach only heuristic justification has been given. Here we propose modifications of this approach that give better approximations for degrees of freedom. First we correct for the bias of the nonparametric estimate. Secondly, we modify the test statistic for the reason that different likelihoods [smoothed or unsmoothed likelihood, respectively] have been used in the calculation of the nonparametric or parametric components. For this modified test statistic asymptotic normality [see (Theorem 4.1)] is established. The convergence to the normal limit is very slow. Therefore we propose using the bootstrap for the calculation of critical values. Consistency of bootstrap is shown in Theorem 4.2.

The bias correction is used because also on the hypothesis the estimate $\widehat{m}_1(t_1)$ may have a non-negligible bias. For this reason in our test, $\widehat{m}_1(t_1)$ is compared with a bootstrap estimate of its expectation under the hypothesis. For this purpose we calculate semiparametric estimates in the hypothesis model (4.1)

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

The α occurring in the preceeding equation is different from the α defined in Section 2, 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. This defines estimates $\tilde{\beta}, \tilde{\alpha}, \tilde{\gamma}_1, \widetilde{m}_2, \ldots, \widetilde{m}_d$. Put

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

For the bootstrap proceed now as follows: generate independent samples $(Y_1^*, \ldots, Y_n^*)$ as in the last section but with μ_i replaced by $\tilde{\mu}_i$. Furthermore, using the data (X_1, T_1, Y_1^*), $\ldots$, (X_n, T_n, Y_n^*) calculate our estimate $\widehat{m}_1^*$. The bootstrap estimate of the 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:

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

Here, $\widehat{m}^+(t) = \hat{\alpha} + \widehat{m}_1(t_1) + \ldots + \widehat{m}_d(t_d)$. 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). It should be remarked that in the definition of the test statistic R the bootstrap estimate $E^* \widehat{m}_1^*$ should not be replaced by a semiparametric estimate of the function m_1, say $\widetilde{m}_1(T_{i,1}) = \tilde{\gamma}_1 T_{i,1}$. This can be deduced from the discussion in Härdle and Mammen (1993) and Härdle, Mammen and Müller (1998) who considered a similar test in another setup.

The following theorem states that the test statistic R has an asymptotic normal distribution.

Theorem 4.1
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, $hg^{d-1}n^{1/2}(\log n)^{-1} \to \infty$ and h and $g = o(n^{-1/8})$, on the hypotheses (4.1), it holds that

$$v_n^{-1}(R - e_n) \overset{D}{\longrightarrow} N(0,1)$$

with

$$e_n = h^{-1} \int K(u)^2 \, du E[A f_{T_1}(T_1)],$$

12

$$v_n^2 = h^{-1} \int K^{(2)}(u)^2 \, du \, E\left\{ E[A|T_1]^2 f_{T_1}(T_1)^3 \right\},$$

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

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

The quantities e_n and v_n can be consistently estimated. So, critical values for the test statistic can be calculated using the normal approximation. Because in similar cases the normal approximation does not perform well (see Härdle, Mammen and Müller, 1996) we 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.

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

The quantities $\widehat{\beta}$ and $\widehat{m}^+$ are not recalculated in the resampling (using the bootstrap samples). This has been done to save computation time. The conditional distribution $\mathcal{L}^*(R^*)$ of 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 (4.1)).

Consistency of bootstrap is the content of the next theorem.

Theorem 4.2
Under the assumptions of Theorem 4.1, 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

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

The results of this section can be easily extended to tests of other parametric hypotheses on m_1, e.g.

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

where $\{m_\theta : \theta \in \Theta\}$ is a parametric family. In particular, one could consider the simple hypothesis that $m_1 \equiv 0$.

With similar arguments as in Härdle and Mammen (1993) one can show that the test R has nontrivial asymptotic power for deviations from the linear hypothesis of order $n^{-1/2}h^{-1/4}$. 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 $n^{-1/2}h^{-1/4}$] that are concentrated on shrinking intervals with length of order h. The test may be compared with overall tests that achieves 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 was not covered by our theory. It can be shown that in this case R is a $n^{-1/2}$ consistent overall test.

5 Uniform bootstrap confidence bands.

In this section we propose using the bootstrap for the construction of uniform confidence bands. We define

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

where $\hat{\sigma}_1^2(t_1)$ is the estimate of the variance of $\widehat{m}_1(t_1)$, defined in (3.2). For the estimation of the distribution of S we use again bootstrap, as introduced in Section 3 for Model C. This defines the statistic $S^* = \sup_t w_1(t_1)|\widehat{m}_1^*(t_1) - E^* \widehat{m}_1^*(t_1)|\hat{\sigma}_1^{-1}(t_1)$. In the definition of S^* the norming $\hat{\sigma}(t_1)$ could be replaced by $\hat{\sigma}_1^*(t_1)$. We write $S^{**} = \sup_t w_1(t_1)|\widehat{m}_1^*(t_1) - E^* \widehat{m}_1^*(t_1)|[\hat{\sigma}_1^*]^{-1}(t_1)$. Here $\hat{\sigma}_1^*(t_1)$ is an estimate of the variance of $\widehat{m}_1^*(t_1)$, that is defined similarly as $\hat{\sigma}(t_1)$ 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 5.1
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 h and g are of order $o(n^{-1/8})$ and that $ng^{2(d-1)}h(\log n)^{-2} \to \infty$. 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 5.1 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 h^O with $h^O/h \to \infty$, see also the remark after Theorem 3.2. For related discussions in nonparametric density estimation and regression see Bickel and Rosenblatt (1973), Eubank and Speckman (1993), Neumann and Polzehl (1998).

6   Simulations and applications

The following model was used to simulate data from a binary response model

$$(6.1) \qquad E(Y|X = x, T = t) = P(Y = 1|x, t) = G\{\beta^T x + m^+(t)\},$$

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 independent. The variables X_1 and X_2 are standard normal and T_1 and T_2 have a uniform distribution on $[-2, 2]$. The sample size was $n = 250$, the number of replications in the bootstrap resampling was $B = 249$. For all computations in this section the quartic kernel $K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \leq 1)$ was used. Figure 1 shows plots of m_1, m_2 and of their estimates. This is done for $\beta = (0.3, -0.7)^T$, $m_1(t_1) = 2\sin(-2t_1)$, $m_2(t_2) = t_2^2 - E[T_2^2]$ and $\alpha = 0$. The chosen bandwidths are $h_1 = (1.0, 1.0)^T$, $h = 0.9$ and $g = 1.0$. Here, h_1 was used for the estimation of β. For the estimation of m_α [$\alpha = 1, = 2,$] the bandwidth h was applied for m_α and g for the other nonparametric component m_j [$j \neq \alpha$]. In Figure 1 the estimates reflect well the shape of the functions m_1 and m_2.
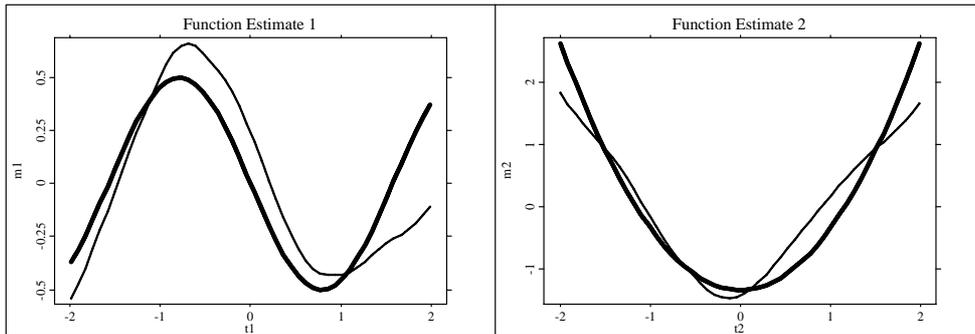

Figure 1: *Plots of the nonparametric components* $m_1(t_1) = 2\sin(-2t_1)$, $m_2(t_2) = t_2^2 - E[T_2^2]$ *and their estimates.*

Consider now the testing problem (4.1) H_0 : $m_1(t_1)$ *is linear.* As discussed above the normal approximation of Theorem 2.1 is quite inaccurate for a small sample size of $n = 250$. 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 as $\beta = (0.3, -0.7)^T$, $m_1(t_1) = t_1$, $m_2(t_2) = t_2^2 - E[T_2^2]$ and $\alpha = 0$. This distribution lies on the hypothesis. 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 the quartic kernel. For better comparison, the normal density has been convoluted with the quartic kernel [with the same bandwidth]. In a simulation with 500 replications the level of the bootstrap test was estimated. The result was a relative number of rejections of 0.03 for $\alpha = 0.05$ and 0.06 for $\alpha = 0.1$, i.e. the bootstrap test keeps its level. Figure 3 plots the power of the test (thick line) for the levels 0.05 and 0.1. The power has been plotted for the alternatives $m_1(t_1) = (1 - v)t_1 + v\{2\sin(-2t_1)\}$, $0 \leq v \leq 1$. The other parameters were chosen as above. For comparison, we made the same simulations for a parametric Likelihood Ratio Test (LRT) of H_0 versus

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

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. The better performance of the parametric test [see Figure 3] is mainly 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.]
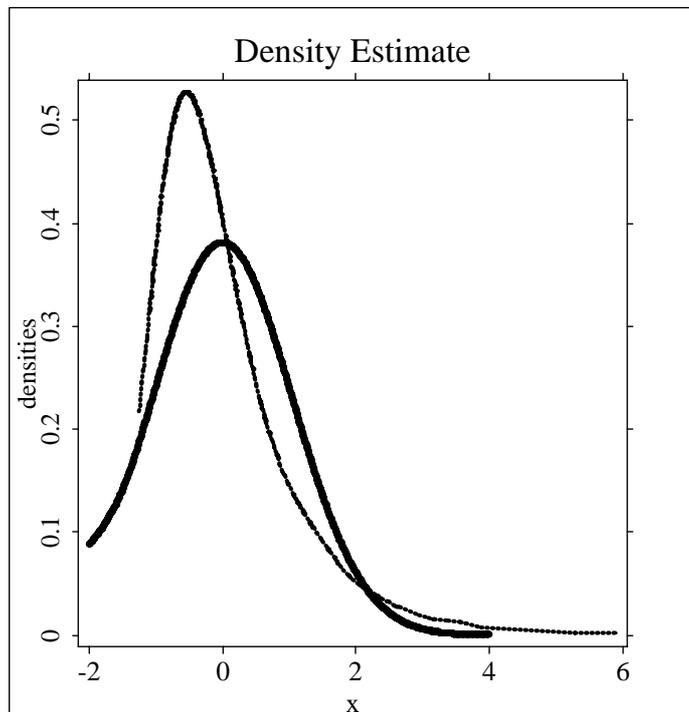

Figure 2: *Standardized density estimate of the test statistic (thin line) and standard normal density (thick line).*
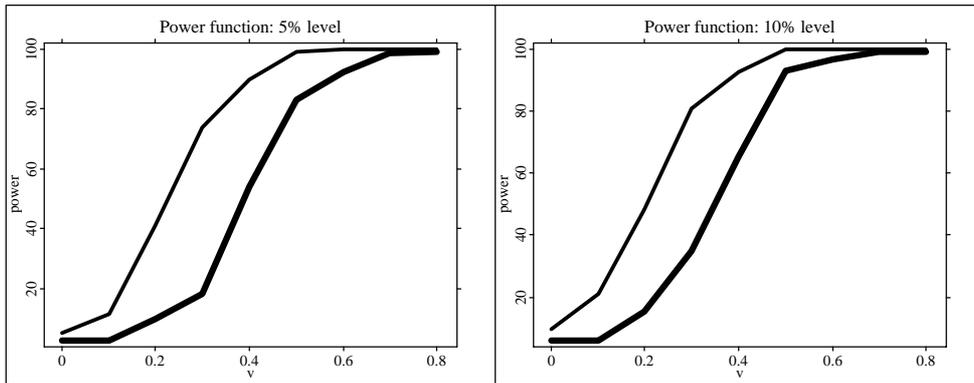
Figure 3: *Power functions for theoretical levels 0.05 and 0.1 , for the non-parametric bootstrap test (thick line) and the likelihood ratio test (thin line).*

We have considered two applications of our methods. For 1991, one year after the unification of East and West Germany, Burda (1993) investigated the impact of various possible determinants on the intention of East-Germans to migrate from East to West Germany. The original data set contains 3710 East Germans who have been surveyed in 1991 in the Socio-Economic Panel of Germany, see GSOEP (1991). Here we consider the datasets from two East German countries: the most northern (Mecklenburg-Vorpommern, $n = 402$) and the most southern (Sachsen, $n = 955$) country of East Germany. 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 is willing to migrate and 0 otherwise. Figures 4 and 5 give plots for the densities of AGE and HHINCOME for both countries. Tables 1 and 2 contain descriptive statistics.
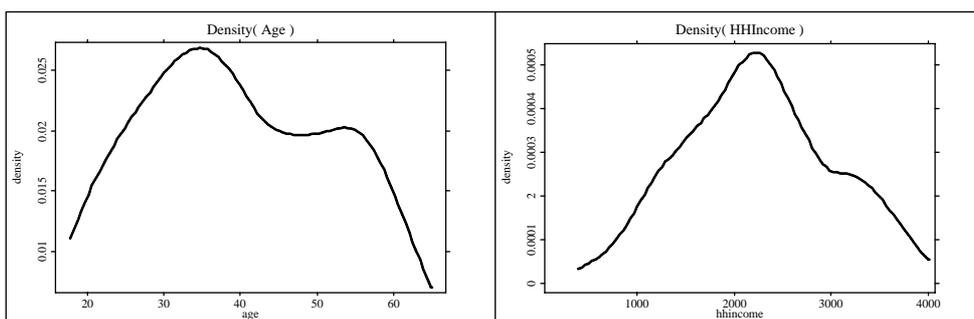

Figure 4: *Density plots for Mecklenburg-Vorpommern, AGE on the left, HHINCOME on the right.*
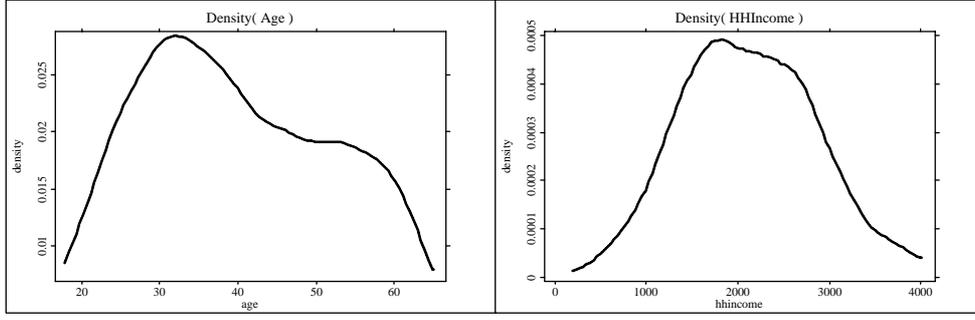
Figure 5: *Density plots for Sachsen,* AGE *on the left,* HHINCOME *on the right.*

<div style="text-align:center">MECKLENBURG-VORPOMMERN</div>

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 1: *Descriptive statistic for our data of Mecklenburg-Vorpommern.*

<div style="text-align:center">SACHSEN</div>

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 2: *Descriptive statistic for our data of Sachsen.*

In the following, the variables AGE and HHINCOME have been standardized to [0, 1]. In a first step we fitted a parametric generalized linear regression model with logit link. The results are presented in Table 3 for both countries, Mecklenburg-Vorpommern and Sachsen.

<div align="center">PARAMETRIC ESTIMATION RESULTS</div>

	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 3: *Results of a generalized linear regression.*

The variable AGE is by far the most significant variable. This holds true for both countries. Obviously people behave quite differently in the two countries, especially concerning X_1 [relatives or friends in West Germany] and for X_2 [their status of employment] and X_3 [city size].

In a second step we fitted a semiparametric generalized additive model for both data sets. We present the results for different smoothing parameters, see the captions of Figures 6 and 7. We choose $h = 1.0$ and $h = 1.25$ for Mecklenburg-Vorpommern and $h = 0.75$, $h = 1.0$ for Sachsen. The other bandwidths have always been $h_1 = g = 1.1 \cdot h$. In Figures 6 and 7 the additive components for AGE and HHINCOME are plotted. Table 4 gives the parametric estimates of the semiparametric model for both choices of the bandwidth. The estimates do not seem to depend strongly on the bandwidth. Furthermore they are similar to the values of the parametric model, compare Table 4. So the qualitative interpretation of these coefficients does not change. In the figures the influence of AGE in Mecklenburg-Vorpommern does not differ strongly from the influence of AGE in Sachsen, except that the curve from Sachsen is more flat in the middle part. For HHINCOME the curves from both countries have a totally different shape.

	Mecklenburg-Vorpommern		Sachsen	
	semi. a	semi. b	semi. a	semi. b
family/friends West	0.5920	0.5809	0.7137	0.7289
unemployed/...	0.7771	0.7992	0.1469	0.1308
middle sized city	0.7156	0.7127	0.3134	0.2774
female	-0.3309	-0.3485	-0.1898	-0.1871
constant	-1.4616	-1.4111	-1.1045	-1.1007

Table 4: *Results for the pure parametric estimation (par.) and for the parametric part of the generalized additive partially linear regression model: semi. a (with bandwidth $h = 1.0$), semi. b ($h = 1.25$) for Mecklenburg-Vorpommern; semi. a ($h = 0.75$) and semi. b ($h = 1.0$) for Sachsen.*

In a third step we applied the bootstrap test procedure to the variables AGE and HHINCOME. We always used 499 replications in the bootstrap resampling. The bandwidths have been chosen as above. For the input AGE, linearity has always been rejected for the 1 percent level, for all bandwidths in both countries. For the variable HHINCOME, the observed p-values are .16 [for $h = 1.0$, Mecklenburg-Vorpommern], .14 [for $h = 1.25$, Mecklenburg-Vorpommern], .02 [for $h = 0.75$, Sachsen], and .01 [for $h = 1.0$, Sachsen]. So the deviations of curves for AGE from linearity are much more significant. At first sight, this seems to be surprising because the plots for HHINCOME differ more from linearity. The reason is that the estimates for HHINCOME have a larger variance.
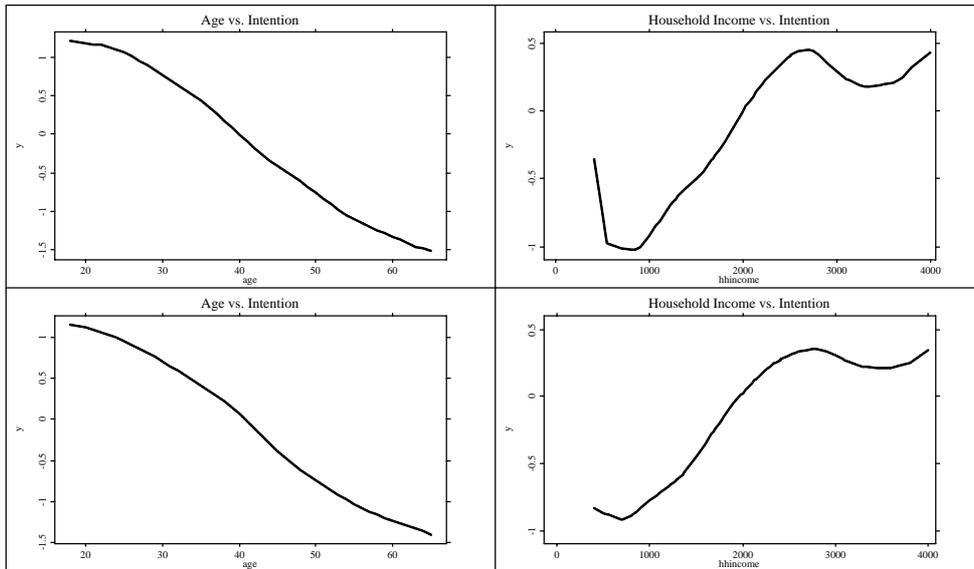
Figure 6: *The semiparametric estimates for the influence of* Age *(left) and* HhIncome *(right) in Mecklenburg-Vorpommern. The upper plots were estimated with* $h = 1.0$, *the lower plots with* $h = 1.25$.
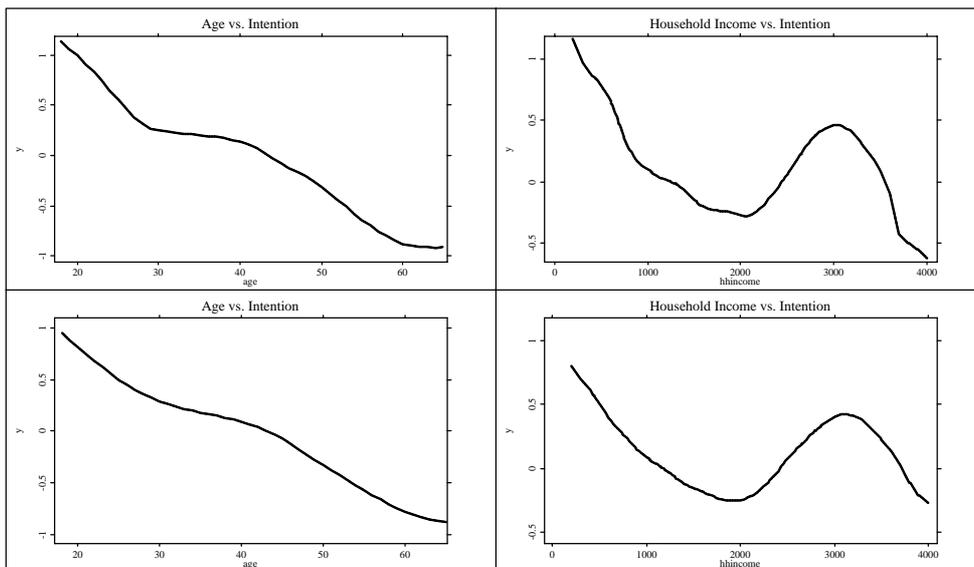

Figure 7: *The semiparametric estimates for the influence of* Age *(left) and* HhIncome *(right) in Sachsen. The upper plots were estimated with* $h = 0.75$, *the lower plots with* $h = 1.0$.

As a second example we considered a data set on the probability that an apprentice becomes unemployed directly after finishing his apprenticeship. The data set has already been discussed by Proença and Werwatz (1995). They considered a sample of 462 individuals from the first nine waves (1984 to 1992) of the GSOEP (German socio

21

economic panel, only West Germany). All people who had completed an apprentice-ship between 1984 and 1991 were included in the sample. We give a brief description of the data. The dependent variable takes on the value "1" if an individual is regis-tered as unemployed in the year following the completion of the apprenticeship. The explanatory variables are summarized in Table 5.

Variable	Definition/Comments
SEX	Sex of the respondent. It takes the value "1" if the respon-dent is female, "0" if male.
AGE	Age of the respondent in the year the apprenticeship was completed.
SCHOOLING	Years of schooling (9 - 13).
EARNINGS	Gross monthly earnings as an apprentice.
BIG CITY	"1" if the city where the respondent lived at end of his ap-prenticeship has between 250 000 and 500 000 inhabitants.
HUGE CITY	"1" if the city has more than 500 000 inhabitants.
DEGREE	Percentage of people apprenticed in a certain occupation, divided by the percentage of people employed in this occu-pation in the entire economy.
U-RATE	Unemployment rate in the state the respondent lived in dur-ing the year the apprenticeship was completed.

Table 5: *Explanatory variables.*

In Figure 8 we present the nonparametric regression curves for AGE, EARNINGS and DEGREE using bandwidths $h = \sigma_T .* (1.5, 1.7, 0.6)^T$, $g = \sigma_T .* (2.0, 2.0, 1.0)^T$. Here .* means elementwise multiplication and σ_T is the vector of standard deviations of T. In the parametric logit fit of Proença and Werwatz (1995) all variables except the constant and URATE have been not significant. We wanted to check if the reason for insignificance could be the assumption of linearity in their model. The plots in Figure 8 show very strong nonlinearities. However, the "jumps" in the plots could be caused by boundary effects and data sparseness. So we applied our bootstrap linearity test for these three covariates. All observed critical levels were more than 20 percent. Therefore the nonlinearities in the plots are not significant. A plausible explanation for the nonlinearities is data sparseness. We conclude that our test safeguards against an overinterpretation of observed shapes of nonparametric smoothers.
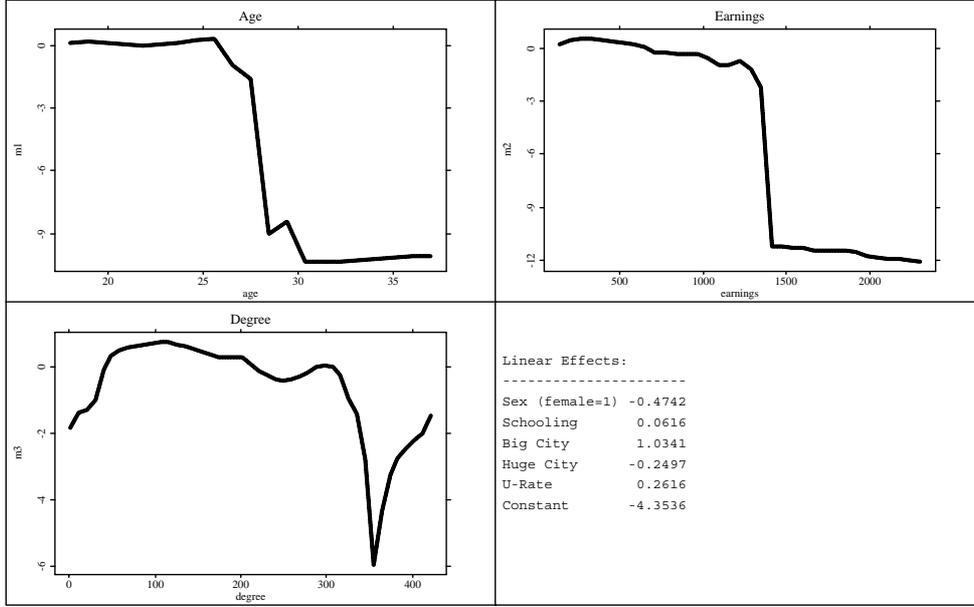
Figure 8: *Estimates of the additive components in the reduced model and the coefficients for the linear part.*

A1 Assumptions

We state now the assumptions used in the results in Sections 2 and 3. We use the notation

$$h_{max} = max\{h, g\},$$

$$h_{prod} = hg^{d-1},$$

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

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

Furthermore, we put

$$\lambda_i(u) = Q\{G(u); Y_i\},$$
$$\lambda(u) = Q\{G(u); Y\}.$$

Then we have

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

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

23

For the asymptotic expansions we make the following assumptions.

(A1) $(X_1, T_1, Y_1), \ldots, (X_n, T_n, Y_n)$ are i.i.d. tuples. T_i takes values in $I\!\!R^d$, X_i is $I\!\!R^p$ valued, and Y_i is $I\!\!R$ valued.

(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 $S_{T,1} \subset I\!\!R$ and $S_{T,-1} \subset I\!\!R^{d-1}$. T has a twice continuously differentiable density f_T with $\inf\limits_{t \in S_T} f_T(t) > 0$.

(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 on $\mathbb{R}$. The weight functions w, w_{-1} and w_1 are positive and twice continuously differentiable. To avoid problems on the boundary, we assume that for a $\delta > 0$ we have that $w_{-1}(t) = 0$, $w_1(t) = 0$, and $w(t) = 0$ for $t \in S_{T,-1}^- = \{s : \text{ there exists an } u \notin S_{T,-1} \text{ with } \|s - u\| \leq \delta\}$, $t \in S_{T,1}^- = \{s : \text{ there exists an } u \notin S_{T,1} \text{ with } \|s - u\| \leq \delta\}$ 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 kernel L is a product kernel $L(v) = L_1(v_1) \cdot \ldots \cdot L_{d-1}(v_{d-1})$. The kernels L_j are symmetric probability densities with compact support ($[-1, 1]$, say), $j = 1, \ldots, d - 1$. The kernel K is a symmetric probability density with compact support (e.g.$[-1, 1]$),too.

(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 $E\, Z^2 \widetilde{X}\widetilde{X}^T$ is strictly positive definite. The random vectors Z and $\widetilde{X}$ have been defined in Theorems 3.1 and 3.3, 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 $\widetilde{X}_{i1} \equiv 0$.

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

(A11) The kernels K and L are twice continuously differentiable.

A2 Proof of Theorem 3.1

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

(A2.1)
$$\widehat{\beta} = \beta_0 + o_P(1).$$

For the proof of (A2.1) we show first that

(A2.2)
$$\sup_{t,\beta} |\widehat{m}_\beta(t) - m_\beta(t)| = o_p(1).$$

Proof of (A2.2): For the proof of claim (A2.2) we show first that:

(A2.3)
$$\sup_{\eta,t,\beta} |\Delta(m_\beta(t), t, \beta)| = O_p(\rho_2),$$

where the following notation has been used:

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

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

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

(A2.4)
$$\kappa_i(t) = \frac{K_h(t_1 - T_{i,1}) L_g(t_{-1} - T_{i,-1})}{\frac{1}{n} \sum_{j=1}^n K_h(t_1 - T_{j,1}) L_g(t_{-1} - T_{j,-1})}.$$

For the proof of (A2.3) 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

(A2.5)
$$P\{|\Delta(\eta, t, \beta)| > C' \rho_2\} = o(n^{-\nu_1}).$$

We consider now 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 (A2.5), following the same argument as for example in Härdle and Mammen (1993), this shows (A2.3).

For the proof of (A2.2), one can conclude from (A2.3) that, with probability tending to one, $\widehat{m}_\beta(t)$ lies in the interior of H, see (A4). This gives

(A2.6)
$$\Delta_1(\widehat{m}_\beta(t), t, \beta) = 0.$$

Because of (A2.3) this shows

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

Because of assumption (A4) this implies (A2.2).

We apply now (A2.2) to prove (A2.1) i.e. that $\widehat{\beta}$ is a consistent estimator of β_0. We proceed similarly as in the proof of Proposition 1 in Severini and Wong (1992).

Proof of (A2.1). Let $k(\beta) = E[Q\{X^T\beta + m_\beta(T); Y\}]$. We will show that

(A2.7)
$$\sup_{\beta \in B} \left| \frac{1}{n} \mathcal{L}(\widehat{m}_\beta, \beta) - k(\beta) \right| \to 0 \qquad \text{(in probability)}.$$

This implies claim (A2.1) because

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

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

It remains to prove (A2.7). This follows from the following two properties:

$$(A2.8) \qquad \sup_{\beta \in B} \left| \frac{1}{n} \mathcal{L}(m_\beta, \beta) - k(\beta) \right| \to 0 \qquad \text{(in probability)},$$

$$(A2.9) \qquad \sup_{\beta \in B} \left| \frac{1}{n} \mathcal{L}(\widehat{m}_\beta, \beta) - \frac{1}{n} \mathcal{L}(m_\beta, \beta) \right| \to 0 \qquad \text{(in probability)}.$$

Claim (A2.8) holds because $\mathcal{L}(m_\beta, \beta)/n$ converges to $k(\beta)$ by the law of large numbers and because $\{\mathcal{L}(m_\beta, \beta)/n, \beta \in B\}$ is tight. For the proof of tightness note first that

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

where

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

It is easy to see that, 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

$$(A2.10) \qquad \frac{\partial m_\beta}{\partial \beta}(\beta, t) = -\frac{E[\lambda''\{\beta^T X + m_\beta(T)\} X | T = t]}{E[\lambda''\{\beta^T X + m_\beta(t)\} | T = t]}.$$

Equation (A2.10) follows by differentiation of $E\{\lambda'(\beta^T X + m_\beta(t)) | T = t\} = 0$. This shows (A2.8).

Claim (A2.9) follows from

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

Thus claim (A2.1) is shown.

Now, we show the following uniform stochastic expansions of $\widehat{\beta}$ and $\widehat{m}(t)$.

$$(A2.11) \qquad \widehat{\beta} = \beta + \{E(Z^2 \widetilde{X} \widetilde{X}^T)\}^{-1} \frac{1}{n} \sum_{i=1}^n \widetilde{X}_i \, \lambda_i'\{X_i^T \beta + m^+(T_i)\} + O_p(\rho_2^2),$$

$$(A2.12) \qquad \sup_{t \in S_T^*} \left| \Delta(t) \right| = O_p(\rho_2^2),$$

27

with

$$
\begin{aligned}
\Delta(t) \;=\; & \widehat{m}(t) - \Big\{ \overline{m}(t) \\
(\text{A2.13}) \qquad & + \{ E(Z^2 | T = t) \}^{-1}\, E(Z^2 X^T | T = t)\, \{ E(Z^2 \widetilde{X} \widetilde{X}^T) \}^{-1} \\
& \times \frac{1}{n} \sum_{i=1}^{n} \widetilde{X}_i\, \lambda_i' \{ X_i^T \beta + m^+(T_i) \} \Big\},
\end{aligned}
$$

$$
(\text{A2.14}) \quad \overline{m}(t) \;=\; m^+(t) + \{ E(Z^2 | T = t) \}^{-1} \frac{1}{n} \sum_{i=1}^{n} \kappa_i(t)\, \lambda_i' \{ X_i^T \beta + m^+(t) \},
$$

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

$$
(\text{A2.15}) \quad \widetilde{X}_i \;=\; X_i - \{ E[Z_i^2 | T_i] \}^{-1}\, E[Z_i^2 X_i | T_i],
$$

$$
(\text{A2.16}) \quad Z_i^2 \;=\; \frac{G'(X_i^T \beta + m^+(T_i))^2}{V[G(X_i^T \beta + m^+(T_i))]}.
$$

Equations (A2.11) and (A2.12) follow from a slight modification of Lemma A3.3 and Corollary A3.4 in Härdle, Mammen and Müller (1996). 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 (1996). 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 (1996), 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 (A2.11). Equation (A2.12) can be shown similarly.

With the help of (A2.12) we arrive at

$$
\begin{aligned}
(\text{A2.17}) \qquad \overline{m}_1(t_1) \;=\; & \frac{\sum_{i=1}^{n} w_{-1}(T_{i,-1}) \overline{m}(t_1, T_{i,-1})}{\sum_{i=1}^{n} w_{-1}(T_{i,-1})} + O_P(\rho_2^2 + n^{-1/2}) \\
\;=\; & m_1(t_1) + R_1 + \Delta_1(t_1) + O_P(\rho_2^2 + n^{-1/2}),
\end{aligned}
$$

where

$$
\begin{aligned}
R_1 \;=\; & \frac{1}{\sum_{i=1}^{n} w_{-1}(T_{i,-1})} \sum_{i=1}^{n} w_{-1}(T_{i,-1}) \left[m_2(T_{i,2}) + \dots + m_d(T_{i,d}) \right] \\
\Delta_1(t_1) \;=\; & \frac{1}{\sum_{i=1}^{n} w_{-1}(T_{i,-1})} \frac{1}{n} \sum_{i,j=1}^{n} \frac{w_{-1}(T_{i,-1}) \kappa_j(t_1, T_{i,-1})}{E(Z_i^2 | T_{i,1} = t_1, T_{i,-1})} \lambda_j' \{ X_j^T \beta + m^+(t_1, T_{i,-1}) \},
\end{aligned}
$$

where λ'_j, κ_j and Z_i are defined by equations (A1.1), (3.3) and (A2.16) 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

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

For this convergence, one uses for instance

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

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

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

$$(A2.18)\quad P(d_K\Big(\mathcal{L}\big(\Delta_1(t_1) - E(\Delta_1(t_1)|\mathcal{Z}_n)\big), N\big(0, Var(\Delta_1(t_1)|\mathcal{Z}_n)\big)\Big) > \delta) \to 0$$

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

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

For the proof of (A2.18) 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

$$(A2.19)\, E(\Delta_1(t_1)|\mathcal{Z}_n) \;=\; \frac{1}{n}\sum_{i=1}^n \int K_h(t_1 - v_1)L_g(T_{i,-1} - v_{-1})$$
$$E\left[\left\{G(X^T\beta + m^+(v)) - G(X^T\beta + m^+(t_1, T_{i,-1}))\right\}\right.$$
$$\left. a^1(X, t_1, T_{i,-1})|T_{i,1} = t_1, T_{i,-1}\right] f_T(v)dv + r_n$$

where the function a^1 is defined in Theorem 3.1, $r_n = O_P(\rho_2^2 + n^{-1/2}) + o_P(h^2 + g^2)$. Furthermore, $r_n = O_P(\rho_2^2 + n^{-1/2} + h^4 + g^4)$ under the additional assumption (A10). The proof of (A2.19) follows by standard, but tedious calculations. The asymptotic form of $E(\Delta_1(t_1)|\mathcal{Z}_n)$ can be easily calculated from (A2.19). 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 Proof of Theorem 3.2

The statement of the theorem follows from

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

Claim (A3.1) follows from

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

(A3.3) $\qquad \dfrac{1}{n}\displaystyle\sum_{i=1}^n w_1(T_{i,1})\left[2\overline{m}_1(T_{i,1}) - E^*\,\overline{m}_1^{\,*}(T_{i,1}) - m_1(T_{i,1})\right]$

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

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 (A2.17).

We give only the proof of (A3.2). Claim (A3.3) follows similarly. Because of (A2.17) 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 gets

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

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 (A3.2) it suffices to show

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

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

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

(A3.5) $\qquad\qquad\qquad\qquad \left. -G\{X_j^T\hat{\beta} + \widehat{m}^+(T_j)\} + G\{X_j^T\hat{\beta} + \widehat{m}^+(t_1, T_{i,-1})\}\right]$

$$+O_P(h^4 + g^4 + (nh)^{-1/2}),$$

30

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 (A3.5) can be treated by using Taylor expansions of G and the stochastic expansions of $\widehat{m}_j$ given in (A2.17). Consider e.g. for $k \neq 1$

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

Then by using the expansions of $\widehat{m}_k$ given in (A2.17) and the expansion of the bias of $\widehat{m}_k$ [see Theorem 3.1] one can show

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

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 $\widehat{m}_1^B(t_1) - m_1(t_1)$. By similar calculations for the other terms one can show the theorem.

A4 Proof of Theorem 3.3

The conditions on h and g imply $\rho_2^2 = o(n^{-1/2})$. Therefore the statement of Theorem 3.3 can be followed from (A2.11).

A5   Proof of Theorem 4.1

We consider the statistic

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

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

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

with

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

We will show that

$$\text{(A5.1)} \qquad\qquad U = V + o_p(h^{-1/2}),$$

$$\text{(A5.2)} \qquad\qquad R = U + o_p(h^{-1/2}),$$

where

$$
\begin{aligned}
V &= \sum_{i=1}^{n} W_i \left\{ \widehat{m}_1^{APPR1}(T_{i,1}) \right\}^2 , \\
\widehat{m}_1^{APPR1}(t_1) &= \frac{1}{n} \sum_{i=1}^{n} a^1(X_i, t_1, T_{i,-1}) f_{T_{-1}}(T_{i,-1}) K_h(t_1 - T_{i,1}) \varepsilon_i, \\
\varepsilon_i &= Y_i - \mu(X_i, T_i), \\
\mu(x, t) &= G\left[x^T \beta + \alpha + \gamma_1 t_1 + m_2(t_2) + \ldots + m_d(t_d) \right].
\end{aligned}
$$

The function a^1 has been defined in the statement of Theorem 3.1. Asymptotic normality of V can be shown as in Härdle and Mammen (1992). In particular, one gets [with pairwise different indices i, j, k and l]

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

32

Because v_n^2 is of order h^{-1} for the proof of the theorem it remains to show (A5.1) and (A5.2).

Proof of (A5.1). Because $\rho_2^2 = o(n^{-1/2})$, it follows from (A2.12) [compare (A2.17)] that uniformly for t_1 in $S_{T,1}^-$:

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

where

$$
\begin{aligned}
M(t) &= \frac{1}{E\left[Z^2|T=t\right]}E\left[Z^2 X^T|T=t\right]E\left[\tilde{X}\tilde{X}^T|T=t\right]^{-1}, \\
B_n &= \frac{1}{n}\sum_{i=1}^n \tilde{X}_i \lambda_i'[X_j^T \beta + m^+(T_j)].
\end{aligned}
$$

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

(A5.3) $\qquad \widehat{m}_1(t_1) = \gamma_1 t_1 + \widehat{m}_1^{APPR1}(t_1) + \widehat{m}_1^{APPR2}(t_1) + \delta_n^1(t_1) + o_P(n^{-1/2}),$

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 $\omega_{i,n,2}$:

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

The function δ_n^1 has been defined in Theorem 3.1.

Furthermore, using similar arguments as in the proof of Theorem 3.2 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 (A5.3) 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 (A5.1) 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 (A5.2). Because of Theorem 3.3, 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

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

Now,

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

This shows (A5.2).

A6   Proof of Theorem 4.2

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

A7 Proof of Theorem 5.1

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 3.1, 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 $S_{T,1}^- = \{s : \text{there exists an } u \notin S_{T,1} \text{ with } |s-u| \leq \delta\}$. [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 we have, that 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

(A7.1) $$\max_{1 \leq i \leq n} E^*|Y_i^{++} - Y_i^+| = O_P(\rho_2).$$

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 (A7.1) 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 $\widehat{m}_1^+(t_1)$. The estimate that is based on $Y_1^{++}, \ldots, Y_n^{++}$ is denoted by $\widehat{m}_1^{++}(t_1)$.

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

$$(A7.2) \qquad \max_{1 \leq i \leq n} \sup_{0 \leq t \leq \tau} E^* |\varepsilon_i^{++} - \varepsilon_i^+|^2 \left\{ 1 + \exp(t|\varepsilon_i^+|) + \exp(t|\varepsilon_i^{++}|) \right\} = O_P(\rho_2).$$

This can be seen by straight forward calculations using (A7.1) 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 $b_n - a_n \leq c_n \log n \, (nh)^{-1/2}$ 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 5.1 follow from

$$(A7.3) \qquad \sup_{t_1 \in S_{T,1}^-} |\hat{\sigma}_1(t_1) - \sigma_1(t_1)| \;=\; o_P(1),$$

$$(A7.4) \qquad \sup_{t_1 \in S_{T,1}^-} |\hat{\sigma}_1^*(t_1) - \sigma_1(t_1)| \;=\; o_P([\log n]^{-1}),$$

$$(A7.5) \qquad \sup_{t_1 \in S_{T,1}^-} \left| \left[\widehat{m}_1^{++}(t_1) - \widehat{m}_1(t_1) \right] \right.$$
$$\left. - \left[\widehat{m}_1^+(t_1) - m_1(t_1) \right] \right| \;=\; o_P((nh)^{-1/2}[\log n]^{-1/2}).$$

We give here only the proof of (A7.5). One shows first that

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

This can be done by using expansions of the type (A2.12). Note that the bias of $\widehat{m}_1^+(t_1)$ and $\widehat{m}_1^{++}(t_1)$ is of order $o_P((nh)^{-1/2}[\log n]^{-1/2})$. So, for (A7.5) it remains to show

$$(A7.6) \qquad \sup_{t_1 \in S_{T,1}^-} |\frac{1}{n} \sum_{i=1}^n a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1})[\varepsilon_i^+ - \varepsilon_i^{++}]|$$
$$= o_P((nh)^{-1/2}[\log n]^{-1/2}).$$

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

$$(A7.7) \quad \sup_{t_1 \in S_{T,1}^-} P^* \left\{ |\frac{1}{n} \sum_{i=1}^{n} a^1(X_i, t_1, T_{i,-1}) K_h(t_1 - T_{i,1})[\varepsilon_i^+ - \varepsilon_i^{++}]| > C_2 \kappa_n \right\}$$

$$= o_P(n^{-C_1}),$$

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 (A7.7) shows that (A7.6) if the supremum runs over a finite set with $O(n^{C_1})$ elements. This implies (A7.6) 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 (A7.6). 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 (A7.2) 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 (A7.6).

References

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

Ai, C. and McFadden, D. (1997) Estimation of some partially identified nonlinear models. *Econometrics* **176** 4 - 37.

Andrews, D. W. K. and Whang, Y. J. (1990) Additive interactive regression models: Circumvention of the curse of dimensionality *Econometric Theory* **6**, 466 - 479.

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.

Bertschek, I. (1996) Semiparametric analysis of innovative behavior. *Ph.D. thesis.* Université Catholique Louvain la Neuve.

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

Bierens, H. (1987) Kernel estimators of regression functions. *Advances in Econometrics: 5th World Congress vol 1.* Bewley, T. F., ed. Cambridge University Press, Cambridge

Buja, A., Hastie, T.J. and Tibshirani, R.J. (1989) Linear smoothers and additive models (with discussion). *Ann. Statist.* **17**, 453 - 510.

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

Carroll, R.J., Fan, J., Gijbels, I, and Wand, M.P. (1995) Generalized partially linear single-index models. The University of New South Wales, Australian Graduate School of Management Working Paper Series, No. 95-010.

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

Fahrmeir, L. and Hamerle, A. (1984) *Multivariate statistische Verfahren*, De Gruyter, Berlin.

Fahrmeir, L. and Tutz, G. (1994) *Multivariate statistical modelling based on generalized linear models*, Springer.

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

Härdle, W. (1990) *Applied nonparametric regression* Cambridge University Press, Cambridge.

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 semi-parametric modelling in generalized linear models. *SFB373 Discussion paper, Humboldt-Universität zu Berlin. Available via http://sfb.wiwi.hu-berlin.de, J. Amer. Statist. Assoc.* to appear.

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.(1997) Nonparametric estimation of a generalized additive model with an unknown link function. Manuscript.

Horowitz, J. and Härdle, W. (1996) Direct semiparametric estimation of single-index models with discrete covariates *J. Amer. Statist. Assoc.* **91**, 1632 - 1640.

Linton, O. B. (1997) Efficient estimation of generalized additive nonparametric regression models. *Technical Report.*

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.

Linton, O. B., Mammen, E. and Nielsen, J. P. (1998) The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Preprint SFB 373.*

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.

Masry, E. and Tjøstheim, D. (1995) Nonparametric estimation and identification of nonlinear ARCH time series: Strong convergence properties and asymptotic normality. *Econometric Theory* **11**, 258-289.

Masry, E. and Tjøstheim, D. (1997) Additive nonlinear ARX time series and projection estimates. *Econometric Theory* **13**, 214-252.

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

Newey, W. K. (1994) Kernel estimation of partial means and a general variance estimator. *Econometric Theory* **10**, 233 - 253.

Opsomer, J. D. (1997) On the existence and asymptotic properties of backfitting estimators. *Preprint.*

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

Proenca, I. and Werwatz, A. (1995) Comparing parametric and semiparametric binary response models. in *XpoRe: An Interactive Statistical Environment,* Springer, Heidelberg, Berlin, New York.

Severance-Lossin, E. and Sperlich, S. (1997) Estimation of derivatives for additive separable models. *SFB 373 Discussion paper 30, Humboldt-Universitt zu Berlin. Available via http://sfb.wiwi.hu-berlin.de*

Severini, T. A. and Staniswalis, J. G. (1994) Quasi-Likelihood estimation in semi-parametric 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.

Silverman, B. W. (1986) *Density estimation for statistics and data analysis.* Chapman and Hall, London.

Sperlich, S., Linton, O. B. and Härdle, W. (1997) A simulation comparison between integration and backfitting methods of estimating separable nonparametric regression models. *SFB 373 Discussion paper 66, Humboldt-Universitt zu Berlin. Available via http://sfb.wiwi.hu-berlin.de*

Stone, C.J. (1983) Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. *Recent Advances in Statistics: Papers presented in Honor of Herman Chernoff's Sixtieth Birthday,* M.H. Rizvi, J.S. Rustagi, and D. Siegmund (eds.), Academic Press, New York.

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.