# Testing for additivity in partially linear regression with possibly missing responses

Ursula U. Müller, Anton Schick and Wolfgang Wefelmeyer

Abstract. We consider a partially linear regression model with multivariate covariates and with responses that are allowed to be missing at random. This covers the usual settings with fully observed data and the nonparametric regression model as special cases. We first develop a test for additivity of the nonparametric part in the complete data model. The test statistic is based on the difference between two empirical estimators that estimate the errors in two ways: the first uses a local polynomial smoother for the nonparametric part; the second estimates the additive components by a marginal integration estimator derived from the local polynomial smoother. We present a uniform stochastic expansion of the empirical estimator based on the marginal integration estimator, and we derive the asymptotic distribution of the test statistic. The transfer principle of Koul, Müller and Schick (2012) then allows a direct adaptation of the results to the case when responses are missing at random. We examine the performance of the tests in a small simulation study.

*Key words:* Partially linear regression, additive regression, local polynomial smoother, marginal integration estimator, uniform stochastic expansion, responses missing at random.

## 1. Introduction

Data sets with a large number of covariates are commonly observed in applications, in particular in biological studies. It is well known that many nonparametric methods do not perform well in this situation, which is often referred to as the 'curse of dimensionality'. A popular semiparametric model which is used to cope with this difficulty is the partially linear model. It combines the flexible nonparametric regression model with the basic linear regression model. In this article we consider a partially linear regression model of the form

$$Y = \vartheta^\top U + \varrho(X) + \varepsilon,$$

where $\vartheta$ is an unknown vector in $\mathbb{R}^p$ and $\varrho$ is an unknown smooth function. The error $\varepsilon$ has mean zero and is assumed to be independent of the pair $(U, X)$, where $U$ and $X$ are (random) covariate vectors. In the ideal situation one observes the triplet $(U, X, Y)$. However, in almost all real life data sets there are missing values. This is an important

problem which needs to be handled with care, since the presence of missing data can easily distort statistical inferences if the wrong method is used. In this article we are specifically interested in the case when some responses $Y$ are missing. Then one observes $(\delta, U, X, \delta Y)$ with $\delta$ an indicator random variable, with the interpretation that for $\delta = 1$ one observes the full triplet $(U, X, Y)$, while for $\delta = 0$ one observes only the covariates $(U, X)$. We make the common assumption that the responses are *missing at random*, which means that the conditional distribution of $\delta$ given $(U, X, Y)$ depends only on the covariates $(U, X)$,

$$P(\delta = 1|U, X, Y) = P(\delta = 1|U, X).$$

Monographs on missing data are Little and Rubin (2002) and Tsiatis (2006).

The partially linear regression model considered here has by definition a partially *additive* structure. We want to go one step further and test the hypothesis that the regression function is completely additive, i.e. even the smooth function $\varrho$ is actually additive,

$$\varrho(x) = \varrho_1(x_1) + \cdots + \varrho_q(x_q), \quad x = (x_1, \ldots, x_q) \in \mathbb{R}^q.$$

It is important to have a diagnostic tool to assess additivity. As shown by Stone (1985), additive models avoid the curse of dimensionality and are easy to interpret.

We will first develop a test procedure for the model with fully observed data, which we describe next. Then we will apply a method by Koul, Müller and Schick (2012), which they call the *transfer principle*, to derive a corresponding procedure for the model with missing responses. The *transfer principle* is a novel approach that makes it easy to derive procedures for certain missing data problems from those with fully observed data.

Assume that we observe $n$ independent copies $(U_1, X_1, Y_1), \ldots, (U_n, X_n, Y_n)$ of $(U, X, Y)$. Our test statistic for additivity will be of the form

$$T = n^{1/2}\|\hat{\mathbb{F}} - \tilde{\mathbb{F}}\| = n^{1/2} \sup_{t\in\mathbb{R}} |\hat{\mathbb{F}}(t) - \tilde{\mathbb{F}}(t)|$$

with two different residual-based empirical distribution functions $\hat{\mathbb{F}}$ and $\tilde{\mathbb{F}}$.

The first uses residuals of the form $\hat{\varepsilon}_j = Y_j - \hat{\vartheta}^\top U_j - \hat{\varrho}(X_j)$ with $\hat{\varrho}$ a local polynomial smoother based on the covariates $X_j$ and the "observations" $Y_j - \hat{\vartheta}^\top U_j$. The second exploits the additivity assumption and works with residuals of the form $\tilde{\varepsilon}_j = Y_j - \hat{\vartheta}^\top U_j - \tilde{\varrho}(X_j)$ with $\tilde{\varrho}$ the marginal integration estimator derived from $\hat{\varrho}$. In both cases, $\hat{\vartheta}$ is some $\sqrt{n}$-consistent estimator of $\vartheta$. Efficient estimators of $\vartheta$ for additive $\varrho$ are constructed in Schick (1996b). Our test statistic $T$ is a variant of the test statistic in Neumeyer and Van Keilegom (2010), who test for additivity in a nonparametric regression model with heteroscedastic errors. Those authors study a bootstrap test based on their test statistic. Here we use the asymptotic distribution to develop our test. We show that, under additivity, $T$ converges in distribution to $\kappa|Z|$, where $Z$ is standard normal and $\kappa$ is a constant depending on the underlying distribution. This leads us to the test $\mathbf{1}[T > \hat{\kappa} z_{\alpha/2}]$ which rejects the null hypothesis if $T$ exceeds $\hat{\kappa} z_{\alpha/2}$ with $z_{\alpha/2}$ the $(1 - \alpha/2)$-quantile of the standard normal distribution and $\hat{\kappa}$ a consistent estimator of $\kappa$.

Our test for missing data uses the *complete case* version of the above test, which is constructed using only the observations with observed responses. More precisely, we reject the null hypothesis if $T_c$ exceeds $\hat\kappa_c z_{\alpha/2}$, where $T_c$ and $\hat\kappa_c$ are the complete case versions of $T$ and $\hat\kappa$. The complete case version of a statistic $S_n = s_n((U_1, X_1, Y_1), \dots, (U_n, X_n, Y_n))$ is of the form $S_c = s_N((U_{i_1}, X_{i_1}, Y_{i_1}), \dots, (U_{i_N}, X_{i_N}, Y_{i_N}))$, where $(U_{i_1}, X_{i_1}, Y_{i_1}), \dots, (U_{i_N}, X_{i_N}, Y_{i_N})$ are the $N = \sum_{j=1}^n \delta_j$ observations with observed responses. An implementation of the test is straightforward since it suffices to write a program for the model with fully observed data. This program then can be used for applications with responses missing at random: just delete all cases where only the covariates are available and work with the remaining $N$ cases that are complete. Since we assume that the covariates and the errors are independent it is clear that the covariates *alone* do not carry information about the error distribution: the complete cases are *sufficient* for inference about functionals of the error distribution function $F$; see also the discussion in Koul et al. (2012).

The reason for using the marginal integration estimator in $\tilde{\mathbb{F}}$ is that the stochastic expansion of $\tilde{\mathbb{F}}$ is then *different* from that of $\hat{\mathbb{F}}$ even under the hypothesis of additivity of $\varrho$, as will be shown in Section 2. This is necessary for the test based on $T$ to have power under *contiguous* alternatives of the form $\varrho(x) = \varrho_1(x_1) + \cdots + \varrho_q(x_q) + n^{-1/2}s(x)$. The two stochastic expansions of $\hat{\mathbb{F}}$ and $\tilde{\mathbb{F}}$ imply in particular an expansion of our test statistic $T$ under the hypothesis of additivity. From this we obtain the asymptotic distribution of $T$ and hence an asymptotic critical value for the test.

We note that the marginal integration estimator is not particularly well suited for estimating the error distribution function. A better estimator would be the series estimator studied in Section 4 of Müller, Schick and Wefelmeyer (2012). The empirical distribution function of this estimator would however be stochastically equivalent to $\hat{\mathbb{F}}$ and therefore lead to a test with local asymptotic power equal to the significance level. The estimator $\hat{\mathbb{F}}$ was studied in Müller et al. (2012), generalizing results by Müller, Schick and Wefelmeyer (2007) who estimate the error distribution function in the *partially linear regression model* but only for one-dimensional $X$. The case $\vartheta = 0$ was studied by Müller, Schick and Wefelmeyer (2009), and by Neumeyer and Van Keilegom (2010), who assume heteroscedastic errors.

The components of the regression function in *additive* regression models can be estimated in several ways. Stone (1985) uses an additive spline estimator. The backfitting method of Breiman and Friedman (1985), and Buja, Hastie and Tibshirani (1989), estimates the additive components one by one and iterates this procedure. Orthogonal series estimators for semiparametric regression models are studied by Eubank, Hart and Speckman (1990), Andrews (1991), Donald and Newey (1994), Eubank (1999), Li (2000), and Delecroix and Protopopescu (2001); for partially linear additive regression models see Müller et al. (2012). Here we use the marginal integration method of Newey (1994), Tjøstheim and Auestad (1994), and Linton and Nielsen (1995). The method starts with an estimator $\hat\varrho$ for

a multivariate nonparametric regression function and obtains estimators for the additive components by integrating out all but one of the variables, usually with empirical estimators based on the remaining components of the covariates. Linton (1997) uses marginal integration to provide an initial estimator, and then a single backfitting step. See also Fan, Härdle and Mammen (1998), and Mammen, Linton and Nielsen (1999). The estimators are compared by Sperlich, Linton and Härdle (1999), Delecroix and Protopopescu (2000), and Dette, Von Lieres und Wilkau and Sperlich (2005).

Residual-based empirical distribution functions can be used to test various other hypotheses about regression models. Tests for parametric hypotheses about the regression function are considered in nonparametric regression by Stute (1997), Khmaladze and Koul (2004, 2009), and Stute, Xu and Zhu (2008). Tests for a parametric regression function in heteroscedastic nonparametric regression are studied in Van Keilegom, González Manteiga and Sánchez Sellero (2008).

The paper is organized as follows. In Section 2 we derive a uniform stochastic expansion for $\tilde{\mathbb{F}}$. The proof is in Section 6. We apply the result to testing $\varrho(x) = \varrho_1(x_1) + \cdots + \varrho_q(x_q)$ in Section 3. Section 4 shows how the results carry over to the situation with responses missing at random. In Section 5 we discuss the finite sample performance of the test and summarize some simulation results.

## 2. Residual-based empirical distribution functions

First we consider the general partially linear model $Y = \vartheta^\top U + \varrho(X) + \varepsilon$, where the error $\varepsilon$ has mean zero, finite variance $\sigma^2$ and a density $f$, and is independent of the covariate pair $(U, X)$, with $U$ a $p$-dimensional random vector and $X$ a $q$-dimensional random vector. We make the following standard assumptions on $U$ and $X$.

(G) The distribution $G$ of $X$ is quasi-uniform on $\mathscr{C} = [0,1]^q$ in the sense that $G(\mathscr{C}) = 1$ and has a density $g$ that is bounded and bounded away from zero on $\mathscr{C}$.

(H) The covariate vector $U$ satisfies $E[|U|^2] < \infty$ and the matrix

$$W = E[(U - E(U|X))(U - E(U|X))^\top]$$

is positive definite.

For a non-negative integer $m$ and a $\gamma \in (0,1]$ we introduce the Hölder space $\mathcal{H}_q(m,\gamma)$ as follows. We say that a function $h$ from $\mathscr{C}$ to $\mathbb{R}$ belongs to $\mathcal{H}_q(m,\gamma)$ if it has continuous partial derivatives up to order $m$ and the partial derivatives of order $m$ are Hölder with exponent $\gamma$. We assume that the function $\varrho$ belongs to $\mathcal{H}_q(m,\gamma)$, and estimate it by a local polynomial smoother of degree $m$; see Stone (1980, 1982), and Ruppert and Wand (1994) for general results on multivariate local polynomial smoothers. Such estimators were used in Müller et al. (2009) for estimating the error distribution function in the case $\vartheta = 0$, i.e., for the nonparametric regression model. Since $\vartheta$ is not zero here, we need a $\sqrt{n}$-consistent

estimator $\hat{\vartheta}$ of $\vartheta$. Such estimators exist, see e.g. Schick (1996a). We then work with the difference $Y_j - \hat{\vartheta}^\top U_j$ instead of the response variable $Y_j$.

In order to define the local polynomial smoother, we introduce some notation. By a *multi-index* we mean a $q$-dimensional vector $i = (i_1, \dots, i_q)$ whose components are non-negative integers. For a multi-index $i$ let $\psi_i$ denote the function on $\mathbb{R}^q$ defined by

$$\psi_i(x) = \frac{x_1^{i_1}}{i_1!} \cdots \frac{x_q^{i_q}}{i_q!}, \quad x = (x_1, \dots, x_q) \in \mathbb{R}^q.$$

Set $i_\bullet = i_1 + \cdots + i_q$. Let $I(m)$ denote the set of multi-indices $i$ with $i_\bullet \le m$, and $J(m)$ the set of multi-indices $i$ with $i_\bullet = m$. Now fix densities $w_1, \dots, w_q$ and set

$$w(x) = w_1(x_1) \cdots w_q(x_q), \quad x = (x_1, \dots, x_q) \in \mathbb{R}^q.$$

Let $c_n$ be a bandwidth. Then the *local polynomial smoother $\hat{\varrho}$ (of degree $m$)* is defined as follows. For a fixed $x$ in $\mathscr{C}$, the estimator $\hat{\varrho}(x)$ is the component $\hat{\beta}_0(x)$ corresponding to the multi-index $0 = (0, \dots, 0)$ of a minimizer

$$\hat{\beta}(x) = \operatorname*{arg\,min}_{\beta=(\beta_i)_{i\in I(m)}} \sum_{j=1}^n w\Big(\frac{X_j - x}{c_n}\Big) \bigg( Y_j - \hat{\vartheta}^\top U_j - \sum_{i \in I(m)} \beta_i \psi_i\Big(\frac{X_j - x}{c_n}\Big) \bigg)^2 .$$

We estimate the errors $\varepsilon_j$ by the residuals

$$\hat{\varepsilon}_j = Y_j - \hat{\vartheta}^\top U_j - \hat{\varrho}(X_j).$$

The empirical distribution functions for $F$ based on the errors $\varepsilon_j$ and on the residuals $\hat{\varepsilon}_j$, respectively, are denoted by

$$\mathbb{F}(t) = \frac{1}{n} \sum_{j=1}^n \mathbf{1}[\varepsilon_j \le t], \qquad \hat{\mathbb{F}}(t) = \frac{1}{n} \sum_{j=1}^n \mathbf{1}[\hat{\varepsilon}_j \le t].$$

Let us write

$$\mu(X) = E(U|X), \qquad \tau(X) = E(|U|^2|X).$$

Müller et al. (2012) have shown the following uniform stochastic expansion for $\hat{\mathbb{F}}$.

THEOREM 1. *Suppose (G) and (H) hold, $\|U\|$ has a moment greater than 2, $\mu$ is continuous and $\tau g$ is bounded. Suppose that $\varrho$ belongs to $\mathcal{H}_q(m, \gamma)$ with $s = m + \gamma > 3q/2$. Let the error density $f$ have mean zero, a finite moment of order greater than $4s/(2s - q)$, and be Hölder with exponent greater than $q/(2s - q)$. Let the densities $w_1, \dots, w_q$ be $(q + 2)$ times continuously differentiable with compact support $[-1, 1]$. Choose a bandwidth $c_n \sim (n \log n)^{-1/(2s)}$. Then we have the uniform stochastic expansion*

$$\sup_{t\in\mathbb{R}} \Big| \hat{\mathbb{F}}(t) - \mathbb{F}(t) - f(t) \frac{1}{n} \sum_{j=1}^n \varepsilon_j \Big| = o_p(n^{-1/2}).$$

The smoothness parameter $s = m + \gamma$ is assumed to be greater than $3q/2$. This means that the higher the dimension $q$ of the covariate vector $X$, the more partial derivatives for $\varrho$ we need. We also point out that $4s/(2s-q) < 3$ and $q/(2s-q) < 1/2$ if $s > 3q/2$. Thus the assumptions on the error density $f$ are satisfied if $f$ has mean zero, a finite third moment, and is Hölder with exponent $1/2$. The Hölder condition is met by all densities with finite Fisher information for location.

Suppose now that the regression function $\varrho$ is *additive*, $\varrho(x) = \varrho_1(x_1) + \cdots + \varrho_q(x_q)$. For this model we introduce an estimator for $\varrho$ such that the corresponding empirical distribution function has a stochastic expansion that is different from that of the empirical distribution function $\hat{\mathbb{F}}$ based on the above local polynomial smoother $\hat{\varrho}$, even in this submodel. Specifically, we take the *marginal integration estimator* $\tilde{\varrho}$ of $\varrho$,

$$\tilde{\varrho}(x) = (1-q)\bar{Y}_* + \sum_{l=1}^{q} \frac{1}{n} \sum_{j=1}^{n} \hat{\varrho}(X_{j,-l}(x_l)),$$

where the random vector $X_{j,-l}(x_l)$ is obtained from $X_j$ by replacing its $l$-th coordinate by $x_l$, and where $\bar{Y}_*$ is the average

$$\bar{Y}_* = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{\vartheta}^\top U_j) = \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j + \frac{1}{n}\sum_{j=1}^{n}\varrho(X_j) - (\hat{\vartheta}-\vartheta)^\top \frac{1}{n}\sum_{j=1}^{n} U_j. \tag{2.1}$$

This leads to the residuals $\tilde{\varepsilon}_j = Y_j - \hat{\vartheta}^\top U_j - \tilde{\varrho}(X_j)$ and to the residual-based empirical distribution function

$$\tilde{\mathbb{F}}(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}[\tilde{\varepsilon}_j \le t]$$

for $F$. The asymptotic behavior of $\hat{\mathbb{F}}$ differs from that of $\tilde{\mathbb{F}}$, as shown next. To state the result, we need some notation.

For $l = 1, \ldots, q$, let $g_l$ denote the density of the $l$-th coordinate of $X$, let $g_{-l}$ denote the density of the vector obtained from $X$ by deleting its $l$-th coordinate, and set

$$g_{(l)}(x) = g_l(x_l) g_{-l}(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_q), \quad x = (x_1, \ldots, x_q) \in \mathbb{R}^q.$$

Note that $g_{(l)}$ is the density of $X_{1,-l}(X_{2,l})$, which is the random vector $X_1$ with its $l$-th coordinate replaced by the $l$-th coordinate $X_{2,l}$ of $X_2$. Let us write

$$h(x) = \sum_{l=1}^{q} \frac{g_{(l)}(x) - g(x)}{g(x)}, \quad x \in \mathscr{C},$$

and

$$\nu = \int \mu(x) h(x) g(x)\, dx.$$

THEOREM 2. *Suppose that the assumption of Theorem 1 are satisfied, now for the partially linear additive model with $\varrho_1, \dots, \varrho_q$ belonging to $\mathcal{H}_1(m, \gamma)$, where $s = m + \gamma > 3q/2$. Then we have the uniform stochastic expansion*

$$\sup_{t\in\mathbb{R}} \Big|\tilde{\mathbb{F}}(t) - \mathbb{F}(t) - f(t)\Big(\frac{1}{n}\sum_{j=1}^{n} \varepsilon_j(1 + h(X_j)) - (\hat{\vartheta} - \vartheta)^\top \nu\Big)\Big| = o_p(n^{-1/2}).$$

The proof of Theorem 2 is in Section 6.

## 3. Testing for additivity

In this section we test the hypothesis $\varrho(x) = \varrho_1(x_1) + \cdots + \varrho_q(x_q)$ in the partially linear regression model $Y = \vartheta^\top U + \varrho(X) + \varepsilon$. As in Section 2, let $\hat{\mathbb{F}}$ and $\tilde{\mathbb{F}}$ denote the residual-based empirical distribution functions based on residuals $\hat{\varepsilon}_j = Y_j - \hat{\vartheta}^\top U_j - \hat{\varrho}(X_j)$ and $\tilde{\varepsilon}_j = Y_j - \hat{\vartheta}^\top U_j - \tilde{\varrho}(X_j)$, respectively.

It follows from Theorems 1 and 2 that, under the hypothesis of additivity, the test statistic

$$T = n^{1/2}\|\hat{\mathbb{F}} - \tilde{\mathbb{F}}\|$$

satisfies the stochastic expansion

$$T = \|f\|\Big|n^{-1/2}\sum_{j=1}^{n} \varepsilon_j h(X_j) - n^{1/2}(\hat{\vartheta} - \vartheta)^\top \nu\Big| + o_p(1). \tag{3.1}$$

Now assume that $\hat{\vartheta}$ satisfies the stochastic expansion

$$\hat{\vartheta} = \vartheta + \frac{1}{n}\sum_{j=1}^{n} W^{-1}(U_j - \mu(X_j))\varepsilon_j + o_p(n^{-1/2}). \tag{3.2}$$

Many authors have constructed such estimators; see e.g. Chen (1988) and Schick (1996a), and the references therein. For estimators satisfying (3.2) the expansion (3.1) becomes

$$T = \|f\| \Big|n^{-1/2}\sum_{j=1}^{n} \varepsilon_j\Big(h(X_j) - \nu^\top W^{-1}(U_j - \mu(X_j))\Big)\Big| + o_p(1),$$

and the test statistic $T$ converges in distribution to $\|f\|\sigma\gamma|Z|$, where $Z$ is a standard normal random variable and

$$\gamma = \big(E[h^2(X)] + \nu^\top W^{-1}\nu\big)^{1/2}.$$

The implementation of this test requires estimators of $\|f\|$, $\sigma$ and $\gamma$. We estimate $\sigma$ by the sample standard deviation $\hat{\sigma}$ based on the residuals $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$. An estimator for $\|f\|$ is $\|\hat{f}\|$, with $\hat{f}$ a kernel density estimator based on these residuals. We estimate $\gamma$ by

$$\hat{\gamma} = \Big(\frac{1}{n}\sum_{j=1}^{n} \hat{h}^2(X_j) + \hat{\nu}^\top \hat{W}^{-1}\hat{\nu}\Big)^{1/2},$$

where $\hat{h}$ is a plug-in estimator of $h$ using kernel estimators of $g_l$, $g_{-l}$ and $g$, and where

$$\hat{\nu} = \frac{1}{n}\sum_{j=1}^{n} \hat{\mu}(X_j)\hat{h}(X_j) \quad \text{and} \quad \hat{W} = \frac{1}{n}\sum_{j=1}^{n} (U_j - \hat{\mu}(X_j))(U_j - \hat{\mu}(X_j))^\top$$

with $\hat{\mu}$ a nonparametric estimator of $\mu$ such as a Nadaraya–Watson estimator. For properly chosen kernels and bandwidths, these estimators are consistent under the assumptions of Theorem 1. The resulting test is $\mathbf{1}[T > \|\hat{f}\|\hat{\sigma}\hat{\gamma}z_{\alpha/2}]$. If $\gamma$ is positive, it will have asymptotic size $\alpha$.

**Local asymptotic behavior.** The local asymptotic power of our test can be derived under a local alternative with $\varrho$ replaced by $\varrho + n^{-1/2}\Delta$, where $\Delta$ belongs to $L_2(G)$ and is orthogonal to the subspace of additive functions. Let us briefly sketch this. Suppose that $f$ has finite Fisher information for location and set $\ell = -f'/f$. Under the local alternative, $T$ converges in distribution to $\|f\|\big|\sigma\gamma Z + \int h\Delta\, dG\big|$. This follows from Le Cam's third lemma which says that the shift must be

$$\begin{aligned} &E\big[\varepsilon(h(X) - \nu^\top W^{-1}(U - \mu(X)))\ell(\varepsilon)\Delta(X)\big] \\ &\qquad = E\big[\varepsilon\ell(\varepsilon)\big]E\big[\big(h(X) - \nu^\top W^{-1}(U - \mu(X)))\Delta(X)\big] = \int h\Delta\, dG. \end{aligned}$$

Here we used the property $E[\varepsilon\ell(\varepsilon)] = 1$. For the test to detect the local alternative, the shift $\int h\Delta\, dG$ must be non-zero. The shift is always zero if, for example, the covariate $X$ follows a uniform distribution on $\mathscr{C}$. Then $g_{(l)} = g = \mathbf{1}_{\mathscr{C}}$ and therefore $h = 0$. The shift is also zero if the components of $X$ are independent.

**Nonparametric regression.** Our test is easily modified to cover the nonparametric regression model $Y = \varrho(X) + \varepsilon$. In this case we take $\hat{\vartheta} = \vartheta = 0$ and obtain the expansion

$$\sup_{t\in\mathbb{R}} \Big|\hat{\mathbb{F}}(t) - \tilde{\mathbb{F}}(t) + f(t)\frac{1}{n}\sum_{j=1}^{n} \varepsilon_j h(X_j)\Big| = o_p(n^{-1/2}) \tag{3.3}$$

under the assumption that $\varrho$ is additive. Now $\gamma$ simplifies to $\gamma_0 = (E[(h^2(X)])^{1/2}$, and we work with the test

$$\mathbf{1}\Big[T > \|\hat{f}\|\hat{\sigma}\Big(\frac{1}{n}\sum_{j=1}^{n} \hat{h}^2(X_j)\Big)^{1/2} z_{\alpha/2}\Big].$$

Neumeyer and Van Keilegom (2010) consider testing for additivity in the heteroscedastic nonparametric regression model $Y = \varrho(X) + s(X)\eta$ with $E[\eta] = 0$ and $E[\eta^2] = 1$. They study a bootstrap test based on an appropriate version of $n^{1/2}\|\hat{\mathbb{F}} - \tilde{\mathbb{F}}\|$ using standardized residuals. Under the null hypothesis they obtain an expansion which coincides with (3.3) when specialized to the homoscedastic case, i.e. to the case $s(X) = \sigma$ and $\varepsilon = \sigma\eta$.

## 4. Responses missing at random

The results of the previous sections carry over to the situation in which responses are *missing at random*. Then we observe i.i.d. copies $(\delta_1, U_1, X_1, \delta_1 Y_1), \dots, (\delta_n, U_n, X_n, \delta_n Y_n)$ of $(\delta, U, X, \delta Y)$ where $\delta$ is an indicator variable depending on the covariables $(U, X)$, but not on the response $Y$. The simplest approach is the *complete case analysis*, which uses only the $N = \sum_{j=1}^n \delta_j$ completely observed triplets $(U_{i_1}, X_{i_1}, Y_{i_1}), \dots, (U_{i_N}, X_{i_N}, Y_{i_N})$, where $i_1, \dots, i_N$ are the indices $i_j$ for which $\delta_{i_j} = 1$. Koul et al. (2012) show how stochastic expansions carry over from a statistic $T_n = t_n(U_1, X_1, Y_1, \dots, U_n, X_n, Y_n)$ to the corresponding *complete case statistic* $T_c = t_N(U_{i_1}, X_{i_1}, Y_{i_1}, \dots, U_{i_N}, X_{i_N}, Y_{i_N})$. We apply this to our empirical distribution functions $\hat{\mathbb{F}}$ and $\tilde{\mathbb{F}}$. For the conditional probability of $\delta = 1$ given $(U, X, Y)$ we write

$$\pi(U, X) = P(\delta = 1 | U, X) = P(\delta = 1 | U, X, Y).$$

Let $R$ denote the joint law of the covariates $(U, X)$, and $Q(U, X, dy)$ the conditional distribution of the response $Y$ given $(U, X)$. Then the joint law of $(\delta, U, X, \delta Y)$ is

$$P(dz, du, dx, dy) = R(du, dx) B_{\pi(u,x)}(dz) \big( z Q(u, x, dy) + (1 - z) \Delta_0(dy) \big),$$

where $B_p$ denotes the Bernoulli distribution with parameter $p$, and $\Delta_t$ is the Dirac measure at $t$. It follows that the conditional distribution of $(U, X, Y)$ given $\delta = 1$ is

$$P_c(du, dx, dy) = R(du, dx) \frac{\pi(u, x)}{E[\delta]} Q(u, x, dy).$$

Let $\mu_c$, $g_c$, $h_c$ and $\nu_c$ be defined like $\mu$, $g$, $h$ and $\nu$ in Section 2, but now with the distribution of $(U, X)$ replaced by the conditional distribution of $(U, X)$ given $\delta = 1$. Explicitly, $\mu_c(X) = E(U | X, \delta = 1)$, and $g_c$ is the density of $X$ given $\delta = 1$. For $l = 1, \dots, q$ let $g_{c,l}$ denote the conditional density of the $l$-th coordinate of $X$ given $\delta = 1$, let $g_{c,-l}$ denote the conditional density of the vector obtained from $X$ by deleting its $l$-th coordinate, and set

$$g_{c(l)}(x) = g_{c,l}(x_l) g_{c,-l}(x_1, \dots, x_{l-1}, x_{l+1}, \dots, x_q), \quad x = (x_1, \dots, x_q) \in \mathbb{R}^q.$$

Write

$$h_c(x) = \sum_{l=1}^q \frac{g_{c(l)}(x) - g_c(x)}{g_c(x)}, \quad x \in \mathscr{C},$$

and define $\nu_c = \int \mu_c(x) h_c(x) g_c(x)\, dx$.

We must of course assume that $E[\delta] > 0$. The assumptions (G) and (H) on $U$ and $X$ are now required to hold under the conditional distribution of $(U, X)$ given $\delta = 1$. This means that the conditional distribution of $U$ given $\delta = 1$ is quasi-uniform, $E(|U|^2 | \delta = 1) < \infty$, and

$$W_c = E\big( (U - \mu_c(X))(U - \mu_c(X))^\top | \delta = 1 \big)$$

is positive definite. These assumptions are implied by (G) and (H) if $\pi$ is bounded away from zero.

Assume that $\hat\vartheta$ has a stochastic expansion of the form (3.2). Let $\hat\vartheta_c$ denote the version of $\hat\vartheta$ based on the complete observations. From the arguments of Koul et al. (2012) it follows that $\hat\vartheta_c$ has the stochastic expansion

$$\hat\vartheta_c = \vartheta + \frac{1}{n}\sum_{j=1}^n \frac{\delta_j}{E[\delta]} W_c^{-1}(U_j - \mu_c(X_j))\varepsilon_j + o_p(n^{-1/2}).$$

Define the local polynomial smoother $\hat\varrho_c$ and the marginal integration estimator $\tilde\varrho_c$ as in Section 2, now using only the complete observations. Note that minimax properties of such complete case estimators in nonparametric regression are obtained by Efromovich (2011). Define residuals $\hat\varepsilon_{c,j} = Y_j - \hat\vartheta_c^\top U_j - \hat\varrho_c(X_j)$, and $\tilde\varepsilon_{c,j} = Y_j - \hat\vartheta_c^\top U_j - \tilde\varrho_c(X_j)$. The complete case versions of the empirical distribution functions $\mathbb{F}$, $\hat{\mathbb{F}}$ and $\tilde{\mathbb{F}}$ are

$$\mathbb{F}_c(t) = \frac{1}{N}\sum_{j=1}^n \delta_j \mathbf{1}[\varepsilon_j \le t], \quad \hat{\mathbb{F}}_c(t) = \frac{1}{N}\sum_{j=1}^n \delta_j \mathbf{1}[\hat\varepsilon_{c,j} \le t], \quad \tilde{\mathbb{F}}_c(t) = \frac{1}{N}\sum_{j=1}^n \delta_j \mathbf{1}[\tilde\varepsilon_{c,j} \le t].$$

Using Koul et al. (2012) again, from the uniform stochastic expansions for $\hat{\mathbb{F}}$ and $\tilde{\mathbb{F}}$ in Theorems 1 and 2 we obtain uniform stochastic expansions for the complete case versions,

$$\sup_{t\in\mathbb{R}} \Big|\hat{\mathbb{F}}_c(t) - \mathbb{F}_c(t) - f(t)\frac{1}{n}\sum_{j=1}^n \frac{\delta_j}{E[\delta]}\varepsilon_j\Big| = o_p(n^{-1/2}),$$

$$\sup_{t\in\mathbb{R}} \Big|\tilde{\mathbb{F}}_c(t) - \mathbb{F}_c(t) - f(t)\frac{1}{n}\sum_{j=1}^n \frac{\delta_j}{E[\delta]}\varepsilon_j\Big(1 + h_c(X_j) - \nu_c^\top W_c^{-1}(U_j - \mu_c(X_j))\Big)\Big| = o_p(n^{-1/2}).$$

It follows from these two expansions that under the hypothesis of additivity the complete case test statistic

$$T_c = \sup_{t\in\mathbb{R}} N^{1/2}|\hat{\mathbb{F}}_c(t) - \tilde{\mathbb{F}}_c(t)|$$

satisfies the stochastic expansion

$$T_c = \|f\|(E[\delta])^{-1/2}\Big|n^{-1/2}\sum_{j=1}^n \delta_j\varepsilon_j\Big(h(X_j) - \nu_c^\top W_c^{-1}(U_j - \mu_c(X_j))\Big)\Big| + o_p(1).$$

The test is now

$$\mathbf{1}[T_c > \|\hat f_c\|\hat\sigma_c\hat\gamma_c z_{\alpha/2}],$$

where $\hat f_c$, $\hat\sigma_c$ and $\hat\gamma_c$ are complete case versions of the estimators $\hat f$, $\hat\sigma$, and $\hat\gamma$ from the previous section. For example, if $\hat f$ is the kernel estimator

$$\hat f(t) = \frac{1}{nb_n}\sum_{j=1}^n k\Big(\frac{\hat\varepsilon_j - t}{b_n}\Big), \quad t\in\mathbb{R},$$

based on the kernel $k$ and bandwidth $b_n$, then its complete case version is of the form

$$\hat{f}_c(t) = \frac{1}{Nb_N} \sum_{j=1}^{n} \delta_j k\Big(\frac{\hat{\varepsilon}_{c,j} - t}{b_N}\Big), \quad t \in \mathbb{R}.$$

Since the pairs $(U, X)$ are always observed, one might be tempted to use $\hat{\gamma}$ instead of $\hat{\gamma}_c$. However, for missing data we need to estimate $\gamma_c$, which is a functional of the conditional distribution of $(U, X)$ given $\delta = 1$, and not $\gamma$, which is the same functional of the unconditional distribution of $(U, X)$. Thus $\gamma_c$ is typically different from $\gamma$, and it is imperative to use $\hat{\gamma}_c$ in the present case. Note also that the use of $\hat{\gamma}$ instead of $\hat{\gamma}_c$ would not result in a complete case estimator.

## 5. Finite sample performance

A marginal integration estimator of $F$ is used to ensure convergence of the test statistic $T$ to a non-degenerate distribution. Working with the additive series estimator proposed by Müller et al. (2012) gives a *better* estimator of $F$, but yields convergence to a degenerate limiting distribution. Indeed, for the additive series estimator the test statistic $T$ converges to zero in probability.

The convergence to the limiting distribution may be slow. Thus a bootstrap version of our test might be preferable in small to moderate sample sizes. For very large sample sizes, however, the bootstrap test becomes intractable and our proposed test provides a reasonable alternative. The simulations in Neumeyer and Van Keilegom (2010) demonstrate good performance of their bootstrap test in their setting with uniform covariates. We expect this to carry over to our situation.

In the following we will give the results for one particular scenario for which the simulations turned out fairly well, even for the relatively small sample size $n = 100$. We considered the case $q = 2$ and $\vartheta = 0$, i.e. $X$ is two-dimensional and the regression function does not have a linear part. We generated covariates $X = (X_1, X_2)$ from a quasi-uniform distribution on $\mathscr{C} = [0, 1]^2$ specified by the density

$$g(x) = g(x_1, x_2) = 1 + 0.5\,\mathrm{sign}(x_1 - 0.5)\mathrm{sign}(x_2 - 0.5), \quad x = (x_1, x_2) \in \mathscr{C}. \tag{5.1}$$

Note that $g$ alternately takes the values 0.5 and 1.5 on the four quarters of $\mathscr{C}$ and that $g_1 = g_2 = g_{-1} = g_{-2} = 1$ on $[0, 1]$. The conditional probability of $\delta = 1$ given the covariates is $\pi(X) = \pi(X_1) = \cos(X_1)$ so that the data contain on average about 84% complete cases. The errors are generated from a normal distribution with mean zero and standard deviation $\sigma$. The test is

$$\mathbf{1}\Big[T_c > \|\hat{f}_c\|\hat{\sigma}_c\Big(\frac{1}{n}\sum_{j=1}^{n} \hat{h}_c^2(X_j)\Big)^{1/2} z_{\alpha/2}\Big].$$

In order to implement the test statistic $T_c$ we used a locally linear smoother provided by the R routine "loess". We looked at several choices of the smoothing parameter "span"

(between zero and one). Under the null hypothesis of additivity we chose $\varrho(X) = X_1 + X_2$. Since our smoother is locally *linear*, it is not too surprising that the test performs best for large values of "span" (see Table 1; note that the default value in R is 0.75).

TABLE 1. Test performance under the *null* hypothesis, $\varrho(X) = X_1 + X_2$.

| | $n = 100$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $\pi(x) = \cos(x)$ | | $\pi(x) = 1$ | | $\pi(x) = \cos(x)$ | | $\pi(x) = 1$ | |
| Span | $\sigma = 0.1$ | $\sigma = 0.5$ | $\sigma = 0.1$ | $\sigma = 0.5$ | $\sigma = 0.1$ | $\sigma = 0.5$ | $\sigma = 0.1$ | $\sigma = 0.5$ |
| 1 | 0.07 | 0.08 | 0.06 (0.07) | 0.06 (0.07) | 0.03 | 0.03 | 0.01 (0.01) | 0.02 (0.01) |
| 0.95 | 0.13 | 0.14 | 0.10 (0.09) | 0.10 (0.10) | 0.09 | 0.10 | 0.05 (0.04) | 0.08 (0.05) |
| 0.9 | 0.18 | 0.19 | 0.16 (0.14) | 0.17 (0.15) | 0.10 | 0.12 | 0.09 (0.09) | 0.10 (0.09) |
| 0.75 | 0.19 | 0.20 | 0.16 (0.24) | 0.18 (0.26) | 0.13 | 0.14 | 0.12 (0.13) | 0.13 (0.14) |
| 0.5 | 0.28 | 0.29 | 0.27 (0.31) | 0.28 (0.33) | 0.19 | 0.21 | 0.16 (0.14) | 0.17 (0.15) |

The figures are the (rounded) proportions of tests with significance level $\alpha = 0.05$ that reject the null hypothesis (in fact true) of additivity. The entries in parentheses are the proportions if the 'true' quantile is used instead of the estimated quantile.

For estimation of the quantiles we used the routine "density" implemented in R to estimate the error density, with a particularly small bandwidth (the range of the residuals $\hat{\varepsilon}_j$ divided by 20) to avoid oversmoothing, in order to obtain a good estimate of the maximum. Our estimator $\hat{\sigma}_c^2$ is the residual-based empirical estimator. Finally we hand coded $\hat{h}_c$ (which involves estimators of the bivariate covariate density and of the marginal densities) using a uniform kernel function and the bandwidth 1/3.

In Table 1 we consider the situation when the regression function is additive and the null hypothesis should not be rejected. The entries are the proportions of tests that reject the hypothesis of additivity in 1000 trials ($n = 100$) and in 500 trials ($n = 500$). We study both the scenario with missing responses and the scenario where all data are completely observed ($\pi(X) = 1$). In order to keep the significance level of the test ($\alpha = 0.05$ in all simulations), we work with a large "span". For completely observed data and our choice of $g$, we have $Eh^2(X) = 4/3$. The 'true' quantile for the simulation scenario thus computes to 0.903. The rejection rates of the tests that use this quantile are given in Table 1 in parentheses. The results are apparently similar to the case when the quantile is estimated. We also observe that there is no great difference between the two cases missing data /

no missing data, which can perhaps be explained by the fact that only a relatively small percentage of responses (about 16%) is missing.

TABLE 2. Performance under the *alternative* hypothesis with regression function $\varrho(X) + cX_1X_2$, where $\varrho(X) = X_1 + X_2$.

| | | span 1 | | | | span 0.95 | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\pi(x)$ | c | $\sigma = 0.1$ | 0.25 | 0.5 | 0.75 | $\sigma = 0.1$ | 0.25 | 0.5 | 0.75 |
| $\cos(x_1)$ | 0.1 | 0.09 | 0.09 | 0.08 | 0.09 | 0.17 | 0.16 | 0.17 | 0.15 |
| | 0.5 | 0.53 | 0.15 | 0.09 | 0.08 | 0.66 | 0.25 | 0.17 | 0.17 |
| | 1 | 0.94 | 0.41 | 0.18 | 0.13 | 0.98 | 0.52 | 0.28 | 0.19 |
| 1 | 0.1 | 0.10 | 0.09 | 0.08 | 0.07 | 0.18 | 0.15 | 0.14 | 0.11 |
| | | (0.10) | (0.09) | (0.07) | (0.07) | (0.19) | (0.14) | (0.14) | (0.12) |
| | 0.5 | 0.62 | 0.21 | 0.11 | 0.08 | 0.77 | 0.29 | 0.19 | 0.13 |
| | | (0.65) | (0.20) | (0.11) | (0.08) | (0.79) | (0.31) | (0.19) | (0.14) |
| | 1 | 0.98 | 0.47 | 0.20 | 0.12 | 1.00 | 0.61 | 0.29 | 0.21 |
| | | (0.99) | (0.50) | (0.20) | (0.13) | (1.00) | (0.62) | (0.29) | (0.22) |

The figures are proportions of tests that (correctly) reject the null hypothesis as in Table 1. The sample size is $n = 100$.

In Table 2 we work with the regression function $\varrho(X) + cX_1X_2$, $c \in \{0.1, 0.5, 1\}$, which violates the null hypothesis since it contains an additional multiplicative part. Note that for $c = 0.1$ the additional part can be regarded a local alternative since $cX_1X_2 = n^{-1/2}X_1X_2$. We observe that the rejection rates for this particular case are quite low and not greatly affected by the error variance $\sigma^2$. The situation is different for $c = 0.5$ and $c = 1$, which denotes alternatives that are easier to detect. As expected, the rejection rates are high for small $\sigma$ and large $c$, and decrease as $\sigma$ becomes larger. Consider, for example, the rejection rates for span 1 (which yielded reasonable results under the null hypothesis) and $c = 1$: for $\sigma = 0.1$ about 98% of the simulated tests reject the (false) null hypothesis, but only about 12% of the tests reject it when $\sigma = 0.75$.

Finally, we also looked at the case when $g$ is the uniform density on $\mathscr{C} = [0, 1]^2$. Here we expect that our test will fail to detect local alternatives: in Section 3 we have shown that under local alternatives the limiting distribution of the test statistic $T$ is shifted by the value $\int h\Delta\, dG$, which is zero for uniform covariates, i.e. the test has no local asymptotic power. For a quick comparison with table 2 we considered the case $\pi(x) = 1$, $c = n^{-1/2} = 0.1$, $\sigma = 0.1$ and span 0.95. (The results for span 1 were similar.) Only 4% of the tests in the scenario with a uniform covariate density rejected the null hypothesis. This is considerably lower than the 18% listed in Table 2 for the corresponding case with covariate density (5.1). For comparison we also looked at the case $n = 500$ and $c = n^{-1/2} \approx 0.045$. About 12% of the tests with $g$ from (5.1) and 0.4% of the tests with uniform covariates rejected the

null hypothesis. These results are in line with our theoretical findings about uniformly distributed covariates.

Although the simulation results turned out quite well in the example with covariate density (5.1) – at least for small $\sigma$ and sufficiently large $c$ – this was not the case in other scenarios that we considered. The problem seems to be that the distribution of $T$ is not close to the asymptotic distribution. One possible reason is that the marginal integration estimator is based on a multivariate local polynomial smoother, which is affected by the curse of dimensionality. A sample size of several hundreds may not be sufficient for the finite sample approximation to be appropriate. Summing up, the proposed additivity test can be used for large enough samples for which a bootstrap test is no longer practically feasible. Our test will have no local power if the components of the covariate vector are independent which applies, for example, if they follow a multivariate uniform distribution.

## 6. Proof of Theorem 2

The marginal integration estimator $\tilde{\varrho}$ is based on the local polynomial smoother $\hat{\varrho}$. We begin by recalling results on $\hat{\varrho}$ from Müller et al. (2012). Order the multi-indices $i \in I(m)$ lexicographically. Let $\psi$ be the vector with components $\psi_i$, $i \in I(m)$. By definition, $\hat{\varrho}$ is the component $\hat{\beta}_0$ of $\hat{\beta}$, where $\hat{\beta}$ is the solution of the normal equation

$$R(x) = W(x)\hat{\beta}(x)$$

with

$$R(x) = \frac{1}{nc_n^q}\sum_{j=1}^n w\Big(\frac{X_j - x}{c_n}\Big)(Y_j - U_j^\top \hat{\vartheta})\psi\Big(\frac{X_j - x}{c_n}\Big)$$

and

$$W(x) = \frac{1}{nc_n^q}\sum_{j=1}^n w\Big(\frac{X_j - x}{c_n}\Big)\psi\Big(\frac{X_j - x}{c_n}\Big)\psi^\top\Big(\frac{X_j - x}{c_n}\Big).$$

On the Hölder space $\mathcal{H}_q(m,\gamma)$ we introduce the norm

$$\|h\|_{m,\gamma}^{(q)} = \max_{i\in I(m)}\sup_{x\in\mathscr{C}}|D^i h(x)| + \max_{i\in J(m)}\sup_{x,y\in\mathscr{C},x\neq y}\frac{|D^i h(y) - D^i h(x)|}{\|x-y\|^\gamma}$$

with

$$D^i h(x) = \frac{\partial^{i_1+\cdots+i_q}}{\partial x_1^{i_1}\cdots\partial x_q^{i_q}}h(x), \quad x = (x_1,\ldots,x_q) \in \mathscr{C}.$$

Let $\mathcal{B}_q(m,\gamma)$ denote the unit ball of $\mathcal{H}_q(m,\gamma)$ for this norm. In the general partially linear regression model, Müller et al. (2012) have obtained the following uniform stochastic expansion of the regression function estimator,

$$\sup_{u\in\mathbb{R}^p,x\in\mathscr{C}}\big|\hat{\vartheta}^\top u + \hat{\varrho}(x) - \vartheta^\top u - \varrho(x) - (\hat{\vartheta}-\vartheta)^\top(u-\mu(x)) - \hat{c}(x)\big| = o_p(n^{-1/2}) \tag{6.1}$$

with

$$\hat{c}(x) = e^\top (E[W(x)])^{-1} \frac{1}{nc_n^q} \sum_{j=1}^n w\Big(\frac{X_j - x}{c_n}\Big) \varepsilon_j \psi\Big(\frac{X_j - x}{c_n}\Big),$$

where $e$ is the vector $(e_i)_{i \in I(m)}$ with $e_0 = 1$ and $e_i = 0$ for $i \neq 0$. In particular,

$$P\big(\hat{c} \in \mathcal{B}_q(q, \alpha)\big) \to 1, \tag{6.2}$$

$$\int |\hat{c}(x)|^{1+\xi} g(x)\, dx = o_p(n^{-1/2}), \tag{6.3}$$

for some $\alpha > 0$ and $\xi > q/(2s - q)$.

In the *additive* partially linear regression model, let us set

$$\tilde{a}(u, x) = (\hat{\vartheta} - \vartheta)^\top (u - \mu_*(x)) + \tilde{c}(x),$$

where

$$\mu_*(x) = (1 - q)E[U] + \sum_{l=1}^q E[\mu(X_{1,-l}(x_l))],$$

$$\tilde{c}(x) = (1 - q)\frac{1}{n} \sum_{j=1}^n \varepsilon_j + \sum_{l=1}^q \frac{1}{n} \sum_{j=1}^n \hat{c}(X_{j,-l}(x_l)).$$

For the Hölder space $\mathcal{H}_1(q, \alpha)$ the above norm simplifies to

$$\|a\|_{q,\alpha}^{(1)} = \sum_{i=1}^q \sup_{0 \le t \le 1} |a^{(i)}(t)| + \sup_{0 \le s \le t \le 1} \frac{|a^{(q)}(s) - a^{(q)}(t)|}{|s - t|^\alpha}.$$

Let $\mathcal{B}(q, \alpha)$ denote the unit ball of $\mathcal{H}_1(q, \alpha)$ for this norm. We write $\mathcal{D}$ for the set of functions of the form

$$a(u, x) = b^\top (u - \mu_*(x)) + a_1(x_1) + \cdots + a_q(x_q)$$

with $|b| \le 1$ and $a_l \in \mathcal{B}(q, \alpha)$ for $l = 1, \ldots, q$. Here $\alpha$ is so small that (6.2) holds. Let $Q$ denote the joint distribution of $U$ and $X$. Recall that $f$ is bounded. By Theorem (2.2) in Müller et al. (2007), the desired result then follows from the following statements:

$$P(\tilde{a} \in \mathcal{D}) \to 1; \tag{6.4}$$

$$\int |\tilde{a}|^{1+\xi}\, dQ = o_p(n^{-1/2}), \quad \xi > q/(2s - q); \tag{6.5}$$

$$\int \tilde{a}\, dQ = \frac{1}{n} \sum_{j=1}^n \varepsilon_j + \frac{1}{n} \sum_{j=1}^n \varepsilon_j h(X_j) - (\hat{\vartheta} - \vartheta)^\top \nu = o_p(n^{-1/2}); \tag{6.6}$$

$$\sup_{u \in \mathbb{R}^k, x \in \mathscr{C}} \big|\hat{\vartheta}^\top u + \tilde{\varrho}(x) - \vartheta^\top u - \varrho(x) - \tilde{a}(u, x)\big| = o_p(n^{-1/2}). \tag{6.7}$$

Note that requirement (2.1) of Müller et al. (2007) on the bracketing numbers of the class $\mathcal{D}$ is verified as in that paper, but now using the bound (1.5) in Müller et al. (2009) in place of the bound (3.1) in Müller et al. (2007). Statements (6.4) and (6.5) above are simple

consequences of the $\sqrt{n}$-consistency of $\hat\vartheta$ and the properties (6.2) and (6.3) of $\hat c$. Statement (6.6) follows from the identities

$$\int \tilde a\, dQ = (\hat\vartheta-\vartheta)^\top\big(E[U]-E[\mu_*(X)]\big) + (1-q)\frac1n\sum_{j=1}^n \varepsilon_j + \sum_{l=1}^q \frac1n\sum_{j=1}^n \int \hat c(X_{j,-l}(t))g_l(t)\,dt$$

and

$$E[U]-E[\mu_*(X)] = qE[U] - \sum_{l=1}^q \int \mu(x)g_{(l)}(x)\,dx = -\nu$$

and from the expansions

$$\frac1n\sum_{j=1}^n \int \hat c(X_{j,-l}(t))g_l(t)\,dt = \int \hat c(x)g_{(l)}(x)\,dx + o_p(n^{-1/2})$$

and

$$\sum_{l=1}^q \int \hat c(x)g_{(l)}(x)\,dx = \int \hat c(x)(q+h(x))g(x)\,dx = \frac1n\sum_{j=1}^n \varepsilon_j(q+h(X_j)) + o_p(n^{-1/2}).$$

These expansions are proved as in Müller et al. (2009). We omit the details.

We now verify (6.7). Note that (6.1) implies the expansion

$$\sup_{x\in\mathscr{C}} \big|\hat\varrho(x) - \varrho(x) + (\hat\vartheta-\vartheta)^\top\mu(x) - \hat c(x)\big| = o_p(n^{-1/2}).$$

From this we can conclude that

$$\sup_{x\in\mathscr{C}} \Big|\sum_{l=1}^q \frac1n\sum_{j=1}^n \Big(\hat\varrho(X_{j,-l}(x_i)) - \varrho(X_{j,-l}(x_l)) + (\hat\vartheta-\vartheta)^\top\mu(X_{j,-l}(x_l)) - \hat c(X_{j,-l}(x_l))\Big)\Big| = o_p(n^{-1/2}).$$

In view of the additivity of $\varrho$, we have the identity

$$\sum_{l=1}^q \frac1n\sum_{j=1}^n \varrho(X_{j,-l}(x_l)) = \varrho(x) + (q-1)\frac1n\sum_{j=1}^n \varrho(X_j).$$

Using the representation (2.1) for $\bar Y_*$, we have

$$\sup_{x\in\mathscr{C}} \big|\tilde\varrho(x) - \varrho(x) + (\hat\vartheta-\vartheta)^\top\tilde\mu(x) - \tilde c(x)\big| = o_p(n^{-1/2}),$$

where

$$\tilde\mu(x) = (1-q)\frac1n\sum_{j=1}^n U_j + \sum_{l=1}^q \frac1n\sum_{j=1}^n \mu(X_{j,-l}(x_l)).$$

Since $\mu$ is continuous, we obtain

$$\sup_{x\in\mathscr{C}} |\tilde\mu(x) - \mu_*(x)| = o_p(1).$$

Combining the above, we obtain (6.7).

## Acknowledgments

Ursula U. Müller was supported by NSF Grant DMS 0907014. Anton Schick was supported by NSF Grant DMS 0906551. The authors thank two referees for suggestions which improved the manuscript.

## References

[1] Andrews, D.W.K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica* **59**, 307–345.

[2] Breiman, L. and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation. With discussion and with a reply by the authors. *J. Amer. Statist. Assoc.* **80**, 580–619.

[3] Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17**, 453–555.

[4] Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.* **16**, 136–146.

[5] Delecroix, M. and Protopopescu, C. (2000). Are regression series estimators efficient in practice? A computational comparison study. *Comput. Statist.* **15**, 511–529.

[6] Delecroix, M. and Protopopescu, C. (2001). Regression series estimators: the MISE approach. *J. Nonparametr. Stat.* **13**, 453–483.

[7] Dette, H., Von Lieres und Wilkau, C. and Sperlich, S. (2005). A comparison of different nonparametric methods for inference on additive models. *J. Nonparametr. Stat.* **17**, 57–81.

[8] Donald, S.G. and Newey, W.K. (1994) Series estimation of semilinear models. *J. Multivariate Anal.* **50**, 30–40.

[9] Efromovich, S. (2011). Nonparametric regression with responses missing at random. *J. Statist. Plann. Inference* **141**, 3744–3752.

[10] Eubank, R.L. (1999). *Nonparametric Regression and Spline Smoothing.* Second edition. Statistics: Textbooks and Monographs 157. Dekker, New York.

[11] Eubank, R.L., Hart, J.D. and Speckman, P. (1990). Trigonometric series regression estimators with an application to partially linear models. *J. Multivariate Anal.* **32**, 70–83.

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

[13] Khmaladze, E.V. and Koul, H.L. (2004). Martingale transforms goodness-of-fit tests in regression models. *Ann. Statist.* **32**, 995–1034.

[14] Khmaladze, E.V. and Koul, H.L. (2009). Goodness-of-fit problem for errors in non-parametric regression: distribution free approach. *Ann. Statist.* **37**, 3165–3185.

[15] Koul, H.L., Müller, U.U. and Schick, A. (2012). The transfer principle: a tool for complete case analysis. *Ann. Statist.* **40**, 3031–3049.

[16] Li, Q. (2000). Efficient estimation of additive partially linear models. *Internat. Econom. Rev.* **41**, 1073–1092.

[17] Linton, O.B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika* **84**, 469–473.

[18] Linton, O. and Nielsen, J.P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**, 93–100.

[19] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* Second edition. Wiley Series in Probability and Statistics, Wiley, Hoboken.
[20] Mammen, E., Linton, O. and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27**, 1443–1490.
[21] Müller, U.U., Schick, A. and Wefelmeyer, W. (2007). Estimating the error distribution in semiparametric regression. *Statist. Decisions* **25**, 1–18.
[22] Müller, U.U., Schick, A. and Wefelmeyer, W. (2009). Estimating the error distribution function in nonparametric regression with multivariate covariates. *Statist. Probab. Lett.* **79**, 957–964.
[23] Müller, U.U., Schick, A. and Wefelmeyer, W. (2012). Estimating the error distribution function in semiparametric additive regression models. *J. Statist. Plann. Inference*, **142**, 552–566.
[24] Neumeyer, N. and Van Keilegom, I. (2010). Estimating the error distribution in nonparametric multiple regression with applications to model testing. *J. Multivariate Anal.* **101**, 1067–1078.
[25] Newey, W.K. (1994). Series estimation of regression functionals. *Econometric Theory* **10**, 1–28.
[26] Ruppert, D. and Wand, M.P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346–1370.
[27] Schick, A. (1996a). Root-n consistent estimation in partly linear regression models. *Statist. Probab. Lett.* **28**, 353–358.
[28] Schick, A. (1996b). Root-n consistent and efficient estimation in semiparametric additive regression models. *Statist. Probab. Lett.* **30**, 45–51.
[29] 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.
[30] Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8**, 1348–1360.
[31] Stone, C.J. (1982). Optimal global rates of convergence for nonparametric estimators. *Ann. Statist.* **10**, 1040–1053.
[32] Stone, C.J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689-705.
[33] Stute, W. (1997). Nonparametric model checks for regression. *Ann. Statist.* **25**, 613–641.
[34] Stute, W., Xu, W.L. and Zhu, L.X. (2008). Model diagnosis for parametric regression in high-dimensional spaces. *Biometrika* **95**, 451–467.
[35] Tjøstheim, D. and Auestad, B.H. (1994). Nonparametric identification of nonlinear time series: projections. *J. Amer. Statist. Assoc.* **89**, 1398–1409.
[36] Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data.* Springer Series in Statistics. Springer, New York.
[37] Van Keilegom, I, González Manteiga, W. and Sánchez Sellero, C. (2008). Goodness-of-fit tests in parametric regression based on the estimation of the error distribution. *TEST* **17**, 401–415.

Ursula U. Müller, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA, uschi@stat.tamu.edu

Anton Schick, Department of Mathematical Sciences, Binghamton University, Binghamton, NY 13902-6000, USA, anton@math.binghamton.edu

Wolfgang Wefelmeyer, Mathematical Institute, University of Cologne, Weyertal 86–90, 50931 Cologne, Germany, wefelm@math.uni-koeln.de