Submitted to the Annals of Statistics

NONPARAMETRIC COVARIATE-ADJUSTED REGRESSION

By Aurore Delaigle*, Peter Hall* and Wen-Xin Zhou*

University of Melbourne and Princeton University

We consider nonparametric estimation of a regression curve when the data are observed with multiplicative distortion which depends on an observed confounding variable. We suggest several estimators, ranging from a relatively simple one that relies on restrictive assumptions usually made in the literature, to a sophisticated piecewise approach that involves reconstructing a smooth curve from an estimator of a constant multiple of its absolute value, and which can be applied in much more general scenarios. We show that, although our nonparametric estimators are constructed from predictors of the unobserved undistorted data, they have the same first order asymptotic properties as the standard estimators that could be computed if the undistorted data were available. We illustrate the good numerical performance of our methods on both simulated and real datasets.

1. Introduction. We consider nonparametric estimation of a regression curve $m(x) = E(Y|X = x)$ when X and Y are observed with multiplicative distortion induced by an observed confounder U. Specifically, we observe $\tilde{X}$, $\tilde{Y}$ and U, where $\tilde{Y} = \psi(U)\,Y$, $\tilde{X} = \varphi(U)\,X$, ψ and φ are unknown functions and U is independent of X and Y. This model is known as a covariate-adjusted regression model. It was introduced by Şentürk and Müller (2005a) to generalize an approach commonly employed in medical studies, where the effect of a confounder U, for example body mass index, is often removed by dividing by U. Motivated by the fibrinogen data on haemodialysis patients, where $\tilde{Y}$ was fibrogen level, $\tilde{X}$ was serum transferrin level, and U was body mass index, Şentürk and Müller (2005a) pointed that although it is often reasonable to assume that the effect of U is multiplicative, it does not need to be proportional to U, and a more flexible model is obtained by allowing for distortions represented by the functions φ and ψ. More generally, this model is useful to describe the relationship between variables that are influenced by a confounding variable, and see if this

*Research supported by grants and fellowships from the Australian Research Council. Peter passed away on 9 January 2016. He was a wonderful man and an extraordinary researcher. We will miss him sorely.

Keywords and phrases: discontinuities, local linear estimator, multiplicative distortion, Nadaraya-Watson estimator, nonparametric smoothing, predictors

relationship still exists once the effect of the confounder has been removed.

A number of authors have suggested estimators of the curve m in various parametric settings. Linear regression models were considered by Şentürk and Müller (2005a, 2006) and Şentürk and Nguyen (2006), who generalized them to varying coefficient models (Şentürk, 2006) and generalized linear models (Şentürk and Müller, 2009). A more general nonlinear regression model was suggested by Cui et al. (2009) and Zhang et al. (2012), and in Zhang et al. (2013), the authors considered a partially linear model, where the linear part is observed with multiplicative distortions.

In this work, we propose more flexible nonparametric estimators of the regression function m, which not only relax the parametric assumptions imposed in the existing literature, but also significantly weaken some of the strong assumptions on the curves φ and ψ and on the distribution of the data made by previous authors. In particular, we propose estimators which, unlike in the previous studies, can be applied if EX and EY vanish, and even if the functions ψ and φ are not strictly positive. Our procedures involve estimating the functions φ and ψ, deduce from there predictors of X and Y, and construct nonparametric estimators of m using those predictors. We show that, under the restrictive assumptions made in the existing literature, this is relatively straightforward to do, whereas under the much weaker assumptions we also consider, we need to use a sophisticated approach.

This paper is organized as follows. We describe the covariate-adjusted model and discuss the model restrictions in the existing literature in Section 2. We propose several nonparametric estimators in Section 3, ranging from the most basic estimators which can be applied under similar restrictions as those imposed in the existing literature, to the most sophisticated ones which rely on much milder assumptions. We derive theoretical properties of our estimators in Section 4, where we show that they have the same first order asymptotic properties as the nonparametric estimators that could be computed if X and Y were observed directly. More surprisingly, in some particular cases, our new estimators can even achieve faster convergence rates than the standard estimators based on direct observations from (X, Y). We discuss practical implementation of our methods in Section 5, where we also investigate their performance on simulated data, and apply them to analyze two real datasets studied in Şentürk and Müller (2005b) and Şentürk and Nguyen (2006). We discuss multivariate extensions in Section 6. Our proofs are provided in Section 7 and in a supplementary file.

2. Model and data. We observe independent and identically distributed (i.i.d.) triplets $\{(\tilde{X}_i, \tilde{Y}_i, U_i)\}_{i=1}^n$ generated by the covariate-adjusted model

of Şentürk and Müller (2005a), where

$$Y = m(X) + \sigma(X)\,\varepsilon, \quad \tilde{Y} = \psi(U)\,Y, \quad \tilde{X} = \varphi(U)X, \tag{2.1}$$

with $m(x) = E(Y|X = x)$ an unknown regression curve that we wish to estimate nonparametrically, $\sigma^2(x) = \mathrm{var}(Y|X = x)$ an unknown variance function, and φ and ψ unknown smooth functions. The random variables U, X and ε are mutually independent, Y and U are independent, $E(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = 1$. We use f_X and f_U to denote the densities of X and U, respectively. As in Şentürk and Müller (2005), to make the problem identifiable, we assume that

$$E\{\varphi(U)\} = E\{\psi(U)\} = 1. \tag{2.2}$$

In other words, on average there is no distorting effect, which is similar to the standard condition imposed in the related classical measurement error problems (Carroll and Hall, 1988; Fan and Truong, 1993), where one observes $W = X + U$ with X and U independent, and the measurement error U is assumed to have zero mean.

As mentioned in the introduction, several parametric estimators of m have been suggested in the literature. There, it is commonly assumed that

$$(a)\ \varphi(u), \psi(u) > 0 \ \text{ for all } u \in I_U\,, (b)\ E(X) \neq 0 \ \text{ and } \ E(Y) \neq 0\,, \tag{2.3}$$

where $I_U \equiv [u_L, u_R]$ denotes the compact support of U. Without loss of generality, we assume that $I_U = [0, 1]$ throughout the paper.

An approach used by some authors is based on constructing predictors of the (X_i, Y_i)'s, which can be obtained from the data $(\tilde{X}_i, \tilde{Y}_i, U_i)$, $i = 1, \ldots, n$, on noting that

$$\varphi_0(U_i) \equiv E(\tilde{X}_i|U_i) = \varphi(U_i)\,E(X), \psi_0(U_i) \equiv E(\tilde{Y}_i|U_i) = \psi(U_i)\,E(Y). \tag{2.4}$$

Now, φ and ψ can easily be estimated nonparametrically, say by $\hat{\varphi}$ and $\hat{\psi}$, which motivates Cui et al.'s (2009) predictors $\hat{Y}_i = \{\hat{\psi}(U_i)\}^{-1}\tilde{Y}_i$ and $\hat{X}_i = \{\hat{\varphi}(U_i)\}^{-1}\tilde{X}_i$, and shows that (2.3) is needed by those authors to avoid dividing by zero. In the next section, we shall see that it is possible to construct consistent nonparametric estimators of m, and that this can be done under much less restrictive conditions than (2.3).

3. Methodology.

3.1. *Different methods under different conditions.* The parametric methods developed in the literature crucially rely on assumption (2.3), and the examples considered there are always such that φ, ψ, EX and EY are far from zero. We wish to construct nonparametric estimators of m that are consistent even if those assumptions do not hold. Let $\mathbf{e}_1 = (1,0)^{\mathrm{T}}$, and, for any pairs of random variables (Q,R) and (Q_i,R_i), $i=1,\dots,n$, let $\mathbf{S}_{Q,n}(x;K,h) = n^{-1}\sum_{i=1}^n K_h(Q_i-x)\mathbf{w}\{h^{-1}(Q_i-x)\}\mathbf{w}\{h^{-1}(Q_i-x)\}^{\mathrm{T}} \in \mathbb{R}^{2\times 2}$ and $\mathbf{T}_{Q,R,n}(x;K,h) = n^{-1}\sum_{i=1}^n R_i K_h(Q_i-x)\mathbf{w}\{h^{-1}(Q_i-x)\}$, with $\mathbf{w}(s) = (1,s)^{\mathrm{T}}$ and where K is a kernel function, $h=h_n>0$ is a bandwidth and, for every $t\in\mathbb{R}$, $K_h(t) = h^{-1}K(t/h)$.

If the (X_i,Y_i)'s were available, we could estimate $m(x)$ nonparametrically by a standard local polynomial estimator constructed from the (X_i,Y_i)'s, the two most popular versions of which are the Nadaraya-Watson and the local linear estimators, defined by

$$\tilde{m}_{\mathrm{NW}}(x) = \frac{\sum_{i=1}^n Y_i K_h(x-X_i)}{\sum_{i=1}^n K_h(x-X_i)}\,,\ \tilde{m}_{\mathrm{LL}}(x) = \mathbf{e}_1^{\mathrm{T}}\mathbf{S}_{X,n}^{-1}(x;K,h)\mathbf{T}_{X,Y,n}(x;K,h), \tag{3.1}$$

respectively. In our case, the (X_i,Y_i)'s are not observed and these standard estimators cannot be computed. We develop new nonparametric estimators that can be computed from the $(\tilde{X}_i,\tilde{Y}_i,U_i)$'s, and whose complexity depends on whether (2.3)(a) and (b) are satisfied or not. The simplest situation is the one where (2.3)(a) holds. There, we can estimate m by standard nonparametric estimators based on predictors of the (X_i,Y_i)'s that are similar to, but less restrictive than, those used by Cui et al. (2009); see Section 3.2. The case where we do not assume (2.3)(a) requires more elaborate techniques: in Section 3.3, we suggest a method that can be used when (2.3)(b) is satisfied; we handle the most general case in Section 3.4, where we develop a sophisticated method which is valid regardless of whether (2.3)(a) and (b) hold or not. It involves computing estimators of unknown constant multiples of $|\varphi|$ and $|\psi|$, estimate the zeros of those functions, construct piecewise estimators of unknown constant multiples of φ and ψ, estimate these constants and finally deduce estimators of φ and ψ.

3.2. *Basic method.* We start by deriving simple nonparametric estimators of m that can be computed when (2.3)(a) holds, and which form the basis of the more sophisticated methods we introduce in the subsequent sections. The idea is similar to the one used in the parametric context by Cui at al. (2009): replace the unobserved (X_i,Y_i)'s by predictors $(\hat{X}_i,\hat{Y}_i)$. Under (2.3), motivated by (2.4) and since $EX = E\tilde{X}$ and $EY = E\tilde{Y}$, Cui et al. (2009) take $\hat{Y}_i = \{\hat{\psi}(U_i)\}^{-1}\tilde{Y}_i$ and $\hat{X}_i = \{\hat{\varphi}(U_i)\}^{-1}\tilde{X}_i$, where $\hat{\varphi}$ and $\hat{\psi}$

denote Nadaraya-Watson estimators of φ_0 and ψ_0, divided by, respectively, $\widehat{EX} = n^{-1}\sum_{i=1}^n \tilde{X}_i$ and $\widehat{EY} = n^{-1}\sum_{i=1}^n \tilde{Y}_i$.

It is because of this division that Cui at al. (2009) assume (2.3)(b), but the latter can be avoided and replaced by $E|X|, E|Y| \neq 0$ (which holds for all non-degenerate random variables), by better exploiting (2.3)(a). Specifically, under (2.3)(a), $|\psi| = \psi$, $|\varphi| = \varphi$, and

$$\varphi_0^+(U_i) \equiv E\big(|\tilde{X}_i|\big|U_i\big) = \varphi(U_i)\, E|X|\,, \psi_0^+(U_i) \equiv E\big(|\tilde{Y}_i|\big|U_i\big) = \psi(U_i)\, E|Y|\,. \tag{3.2}$$

Motivated by this, we propose to estimate ψ and φ by

$$\hat{\varphi}_{\mathrm{LL}}(u) = \hat{\varphi}^+_{0,\mathrm{LL}}(u)/\widehat{E|X|} \quad \text{and} \quad \hat{\psi}_{\mathrm{LL}}(u) = \hat{\psi}^+_{0,\mathrm{LL}}(u)/\widehat{E|Y|}\,, \tag{3.3}$$

where $\widehat{E|X|} = n^{-1}\sum_{i=1}^n |\tilde{X}_i|$, $\widehat{E|Y|} = n^{-1}\sum_{i=1}^n |\tilde{Y}_i|$, and where $\hat{\varphi}^+_{0,\mathrm{LL}}(u) = \mathbf{e}_1^{\mathrm{T}}\mathbf{S}^{-1}_{U,n}(u; L, g_1)\mathbf{T}_{U,|\tilde{X}|,n}(u; L, g_1)$ and $\hat{\psi}^+_{0,\mathrm{LL}}(u) = \mathbf{e}_1^{\mathrm{T}}\mathbf{S}^{-1}_{U,n}(u; L, g_2)\mathbf{T}_{U,|\tilde{Y}|,n}(u; L, g_2)$ are local linear estimators of φ_0^+ and ψ_0^+ computed with a kernel function L and bandwidths g_1 and g_2.

Then, we predict Y_i and X_i by taking

$$\hat{Y}_i = \{\hat{\psi}_{\mathrm{LL}}(U_i)\}^{-1}\tilde{Y}_i \quad \text{and} \quad \hat{X}_i = \{\hat{\varphi}_{\mathrm{LL}}(U_i)\}^{-1}\tilde{X}_i\,. \tag{3.4}$$

Finally, replacing (X_i, Y_i) by $(\hat{X}_i, \hat{Y}_i)$ in (3.1), we obtain the following estimators of $m(x)$:

$$\hat{m}_{\mathrm{NW}}(x) = \frac{\sum_{i=1}^n \hat{Y}_i K_h(x - \hat{X}_i)}{\sum_{i=1}^n K_h(x - \hat{X}_i)}, \hat{m}_{\mathrm{LL}}(x) = \mathbf{e}_1^{\mathrm{T}}\mathbf{S}^{-1}_{\hat{X},n}(x; K, h)\mathbf{T}_{\hat{X},\hat{Y},n}(x; K, h)\,. \tag{3.5}$$

REMARK 3.1. Using $E(\tilde{Y}_i|X_i) = E(Y_i|X_i) = m(X_i)$, simpler estimators of m can also be defined by $\hat{m}_{\mathrm{NW},0}(x) = \sum_{i=1}^n \tilde{Y}_i K_h(x - \hat{X}_i)/K_h(x - \hat{X}_i)$ and $\hat{m}_{\mathrm{LL},0}(x) = \mathbf{e}_1^{\mathrm{T}}\mathbf{S}^{-1}_{\hat{X},n}(x; K, h)\mathbf{T}_{\hat{X},\tilde{Y},n}(x; K, h)$. Since they require predicting only the X_i's, these estimators seem more attractive than those in (3.5). However, it can be proved that their asymptotic "variance" is larger than that of the estimators in (3.5). Moreover, they cannot be adapted simply to the case where φ does not satisfy (2.3)(a); see Remark 3.3 in Section 3.3.

3.3. *Refined procedure.* As their parametric counterparts developed in the covariate-adjusted literature, the methods introduced in Section 3.2 can only be computed if (2.3)(a) holds. However, in practice, there is no reason why φ and ψ would always be positive, and even if they are, their estimators may vanish or get close to zero, which can cause numerical problems. In this

section, we suggest a refined approach which can overcome these difficulties when (2.3)(b) holds. The more complex case where (2.3)(b) is violated will be dealt with in Section 3.4.

As in Section 3.2, to estimate m, the first step is to construct predictors $\hat{X}_i$ and $\hat{Y}_i$, and thus estimators of φ and ψ. Recall the notation in (2.4). Since we assume (2.3)(b) but not (2.3)(a), instead of (3.3) we take $\hat{Y}_i = \{\hat{\psi}_{\mathrm{LL}}(U_i)\}^{-1}\tilde{Y}_i$ and $\hat{X}_i = \{\hat{\varphi}_{\mathrm{LL}}(U_i)\}^{-1}\tilde{X}_i$, where $\hat{\varphi}_{\mathrm{LL}}(u) = \hat{\varphi}_{0,\mathrm{LL}}(u)/\widehat{EX}$ and $\hat{\psi}_{\mathrm{LL}}(u) = \hat{\psi}_{0,\mathrm{LL}}(u)/\widehat{EY}$, and the local linear estimators

$$\begin{aligned}\hat{\varphi}_{0,\mathrm{LL}}(u) &= \mathbf{e}_1^{\mathrm{T}}\mathbf{S}_{U,n}^{-1}(u; L, g_1)\mathbf{T}_{U,\tilde{X},n}(u; L, g_1)\,,\\ \hat{\psi}_{0,\mathrm{LL}}(u) &= \mathbf{e}_1^{\mathrm{T}}\mathbf{S}_{U,n}^{-1}(u; L, g_2)\mathbf{T}_{U,\tilde{Y},n}(u; L, g_2)\end{aligned} \tag{3.6}$$

of φ_0 and ψ_0 computed with a kernel function L and bandwidths g_1 and g_2.

To derive consistent estimators of m without imposing (2.3)(a), recall that, for each i, X_i and Y_i are independent of U_i. As a consequence, for any subset $\mathcal{S} \subseteq \mathbb{R}$, we have $E(Y_i|X_i = x, U_i \in \mathcal{S}) = E(Y_i|X_i = x)$. In particular, if X_i, Y_i, φ and ψ were known, then letting $\mathcal{C}_n(\rho_1, \rho_2) = \{1 \le i \le n : |\varphi_0(U_i)| \ge \rho_1, |\psi_0(U_i)| \ge \rho_2\}$, with $\rho_1, \rho_2 > 0$ denoting two small numbers, the following modification of $\tilde{m}_{\mathrm{NW}}(x)$ at (3.1) would be consistent:

$$\tilde{m}_{\mathrm{NW}}(x; \rho_1, \rho_2) = \sum_{i\in\mathcal{C}_n(\rho_1,\rho_2)} Y_i K_h(x - X_i) \Big/ \sum_{i\in\mathcal{C}_n(\rho_1,\rho_2)} K_h(x - X_i)\,,$$

and a similar consistent version $\tilde{m}_{\mathrm{LL}}(x; \rho_1, \rho_2)$ of $\tilde{m}_{\mathrm{LL}}(x)$ at (3.1) could be constructed by replacing, in the definition of $\tilde{m}_{\mathrm{LL}}(x)$, sums over all i by sums over $i \in \mathcal{C}_n(\rho_1, \rho_2)$ as above. The advantage of this approach is that it enables us to exclude the data for which $\psi(U_i)$ or $\varphi(U_i)$ are small, and thus it can be applied even if (2.3)(a) does not hold.

Motivated by this discussion, in the case that interests us, where X_i, Y_i, φ and ψ are unknown, we suggest estimating m as follows. First, let $\hat{\mathcal{C}}_n(\rho_1, \rho_2) = \big\{i = 1, \ldots, n : |\hat{\varphi}_{0,\mathrm{LL}}(U_i)| \ge \rho_1,\ |\hat{\psi}_{0,\mathrm{LL}}(U_i)| \ge \rho_2\big\}$. (The choice of ρ_1 and ρ_2 will be discussed in Section 5.) We define a Nadaraya-Watson estimator of $m(x)$, valid even if (2.3)(a) does not hold, by

$$\hat{m}_{\mathrm{NW}}(x; \rho_1, \rho_2) = \sum_{i\in\hat{\mathcal{C}}_n(\rho_1,\rho_2)} \hat{Y}_i K_h(x - \hat{X}_i) \Big/ \sum_{i\in\hat{\mathcal{C}}_n(\rho_1,\rho_2)} K_h(x - \hat{X}_i)\,. \tag{3.7}$$

Similarly, we define a local linear estimator $\hat{m}_{\mathrm{LL}}(x; \rho_1, \rho_2)$ in the same way as $\hat{m}_{\mathrm{LL}}$ in (3.5), replacing there, and in the definitions of $\mathbf{S}_{\hat{X},n}(x; K, h)$ and $\mathbf{T}_{\hat{X},\hat{Y},n}(x; K, h)$, the indices $i = 1, \ldots, n$ by the indices $i \in \hat{\mathcal{C}}_n(\rho_1, \rho_2)$.

REMARK 3.2. While we shall prove in Section 4 that these estimators are consistent and have the same first order asymptotic properties as their counterparts at (3.1) based on undistorted data, in practice performance can be further improved by excluding a small fraction (say 5%) of the observations corresponding to the U_i's such that a kernel density estimator $\hat{f}_U(U_i)$ of $f_U(U_i)$ is the smallest. (Indeed, we know from standard properties of kernel regression estimators that, at points u where $f_U(u)$ is small, $\hat{\varphi}(u)$ and $\hat{\psi}(u)$ are more variable.) Doing this corresponds to enlarging the set $\mathcal{S}$ slightly, which does not affect consistency and convergence rates, again due to the fact that the U_i's are independent of the (X_i, Y_i)'s,

REMARK 3.3. It is not possible to directly use this approach to modify the estimator discussed in Remark 3.1 for the case where φ has zeros, because $\tilde{Y}_i$ and U_i are dependent. Particularly, we note that in general $E(\tilde{Y}_i|X_i = x, U_i \in \mathcal{S})$ and $E(\tilde{Y}_i|X_i = x)$ are not equal.

3.4. *Elaborate procedure for the most general case.* Finally we construct estimators of m that rely on neither part of (2.3). As before, we start by deriving predictors of the (X_i, Y_i)'s. Constructing predictors $\hat{X}_i$ (resp., $\hat{Y}_i$) without assuming (2.3) requires to derive an estimator of φ (resp., ψ) without this assumption, which, unlike the methods used in the previous sections, turns out to be a challenging task. Our procedure is based on the fact that, from (2.1), $\varphi^*(u) \equiv E(|\tilde{X}|\,|U = u) = |\varphi(u)|\,E|X|$ (resp., $\psi^*(u) \equiv E(|\tilde{Y}|\,|U = u) = |\psi(u)|\,E|Y|$), which implies that we can estimate φ^* (resp., ψ^*) by a standard local linear estimator $\hat{\varphi}^*_{\mathrm{LL}}$ (resp., $\hat{\psi}^*_{\mathrm{LL}}$) with kernel L and bandwidth g_1 (resp., g_2) constructed from the $(U_i, |X_i|)$'s (resp., the $(U_i, |\tilde{Y}_i|)$'s). In what follows, we explain how to deduce an estimator of φ from $\hat{\varphi}^*_{\mathrm{LL}}$. The same procedure can be applied to derive an estimator of ψ from $\hat{\psi}^*_{\mathrm{LL}}$.

Since φ^* is proportional to $|\varphi|$, to extract an estimator of φ from $\hat{\varphi}^*_{\mathrm{LL}}$, we need to estimate the zeros of φ, say $\tau_1, \ldots, \tau_M$ for some finite M, at which φ changes sign. To do this we assume that, for each j, $\varphi''(\tau_j) \neq 0$. Then, it is straightforward to see that the first derivative of φ^* has jump discontinuities at the τ_j's. Moreover, the zeros of φ coincide with those of φ^*, so that, at the τ_j's, φ^* reaches its minimum value, 0. Therefore, the τ_j's can be estimated using procedures for detecting discontinuities in derivatives of a regression curve, such as those in Gijbels et al. (1999) and Gijbels and Goderniaux (2005), combined with the fact that the $\hat{\tau}_j$'s need to correspond to local minima of $\hat{\varphi}^*_{\mathrm{LL}}$; see Section 5.2 for details of implementation. For $j = 1, \ldots, M$, let $\hat{\tau}_j$ denote the resulting estimator of τ_j, and let $I_0 = (-\infty, \hat{\tau}_1)$, $I_M = [\hat{\tau}_M, \infty)$, and, for $j = 1, \ldots, M-1$, $I_j = [\hat{\tau}_j, \hat{\tau}_{j+1})$.

Our next target is to construct an estimator of φ. Recall the notation

$\varphi_0^+ = \varphi \cdot E|X|$ in (3.2). Recalling that φ changes sign at each τ_j, we can obtain a consistent estimator of either φ_0^+ or $-\varphi_0^+$ (we'll see below how to distinguish these two cases) by taking $\hat{\varphi}^+_{\pm,0}(x) = \sum_{j=0}^{M}(-1)^j\, \hat{\varphi}^*_{j,\mathrm{LL}}(x) \cdot I(x \in I_j)$, where, for each j, $\hat{\varphi}^*_{j,\mathrm{LL}}$ denotes the local linear estimator of φ^* constructed using only the $(U_i, |\tilde{X}_i|)$'s for which $U_i \in I_j$. Here we use a different local estimator in each I_j because, under our assumptions, the first derivative of $\varphi^* = |\varphi| \cdot E|X|$ is discontinuous at the τ_j's. It can be shown using standard kernel smoothing arguments that in this case the bias near the τ_j's is reduced by using this piecewise approach.

Our next step is to extract from $\hat{\varphi}^+_{\pm,0}$ an estimator of φ_0^+ (recall that $\hat{\varphi}^+_{\pm,0}$ is an estimator of φ_0^+ or $-\varphi_0^+$, but we can't know of which one). To do this, recall that $E\{\varphi(U)\} = 1$, which implies that $E\{\varphi_0^+(U)\} > 0$. This fact motivates us to estimate $\varphi_0^+(x)$ by $\hat{\varphi}_0^+(x) = \hat{\varphi}^+_{\pm,0}(x) \,/\, \mathrm{sign}\{\sum_{i=1}^n \hat{\varphi}^+_{\pm,0}(U_i)\}$. Since $\varphi_0^+(x) = \varphi(x)\, E|X|$, once we have done this, to estimate φ it remains to construct an estimator of $E|X|$.

Noting that $E\{\varphi_0^+(U)\} = E\{\varphi(U)\}\, E|X| = E|X|$, we can estimate $E|X|$ by $\widehat{E|X|} = n^{-1}\sum_{i=1}^n \hat{\varphi}_0^+(U_i) = \big|n^{-1}\sum_{i=1}^n \hat{\varphi}^+_{\pm,0}(U_i)\big|$. Finally we estimate $\varphi(x)$ by $\hat{\varphi}(x) = \hat{\varphi}_0^+(x)\,/\widehat{E|X|}$. Then, we can predict the X_i's by taking $\hat{X}_i = \{\hat{\varphi}(U_i)\}^{-1}\tilde{X}_i$. We can proceed similarly to construct predictors $\hat{Y}_i$ of the Y_i's. As in Section 3.3, since, to obtain these predictors, we divide by $\hat{\varphi}(U_i)$ and $\hat{\psi}(U_i)$, when constructing our estimator of m we cannot use the $(\hat{X}_i, \hat{Y}_i)$'s for which $|\hat{\varphi}(U_i)|$ or $|\hat{\psi}(U_i)|$ is too small. Therefore, to estimate m we use the estimators $\hat{m}_{\mathrm{NW}}(x; \rho_1, \rho_2)$ and $\hat{m}_{\mathrm{LL}}(x; \rho_1, \rho_2)$ defined in Section 3.3, but with the predictors $\hat{X}_i$ and $\hat{Y}_i$ constructed above.

4. Theoretical properties. We start by establishing theoretical properties of the estimators $\hat{m}_{\mathrm{NW}}$ and $\hat{m}_{\mathrm{LL}}$ from Section 3.2. While these estimators seem intuitively natural, because they are computed using variables obtained through nonparametric prediction, checking whether they are consistent, and deriving detailed asymptotic properties, are quite difficult. Recently, Mammen et al. (2012) gave a deep account of nonparametric estimators computed from nonparametrically generated covariates, but our estimators do not fall into the class of settings they consider, not least because in our case, not only the covariate X, but also the dependent variable Y, are nonparametrically generated, which makes the problem even more complex than theirs. In addition to the basic model assumptions introduced in the first paragraph of Section 2, we make the following regularity assumptions:

(B1) $E|X| \neq 0$, $E|Y| \neq 0$ and $\inf_{u\in I_U} \varphi(u) > 0$, $\inf_{u\in I_U} \psi(u) > 0$.
(B2) $0 < \inf_{u\in I_U} f_U(u) \leq \sup_{u\in I_U} f_U(u) < \infty$; f_U, φ and ψ are twice dif-

ferentiable, and their second derivatives are uniformly continuous and bounded.

(B3) (a) f_X is continuous, $\sup_{x\in\mathbb{R}} f_X(x) < \infty$, and $E\{\exp(c_1|X|)\} < \infty$ for some constant $c_1 > 0$; (b) m and f_X are twice differentiable and their second derivatives are uniformly continuous and bounded; (c) σ is continuous and bounded.

(B4) $E(\varepsilon) = 0$, $E(\varepsilon^2) = 1$ and $E\{\exp(c_2|\varepsilon|)\} < \infty$ for some $c_2 > 0$.

(B5) K and L are twice continuously differentiable, symmetric density functions, and are compactly supported on $[-1, 1]$. Moreover, $\int_0^1 t^2 L(t)\,dt > 2\{\int_0^1 tL(t)\,dt\}^2$.

(B6) The bandwidths $(h, g_1, g_2) = (h_n, g_{1n}, g_{2n})$ are such that $h \asymp n^{-\alpha_0}$ and $g_1 \asymp n^{-\beta_1}$ and $g_2 \asymp n^{-\beta_2}$ for some $0 < \alpha_0, \beta_1, \beta_2 < 1/3$.

Condition (B1) is a relaxed version of assumption (2.3) often assumed in the covariate-adjusted regression literature. See, for example, Şentürk and Müller (2005a, 2006) and Cui et al. (2009). Condition (B2) includes standard regularity and smoothness assumptions for the asymptotic results of kernel-type nonparametric regression estimation. In (B3), we relax the conventional boundedness condition on the covariates used by Şentürk and Müller (2005a, 2006) and Mammen et al. (2012), and assume instead that X has a finite exponential moment (for example this is satisfied if the distribution of X comes from the exponential family or is compactly supported). Condition (B4), which requires exponentially light tails of ε, is similar in spirit to Assumption 1. (iv) in Mammen et al. (2012). Like them, we need this technical assumption to employ an argument based on empirical processes. Condition (B5) is standard in the context of kernel regression, and is easy to satisfy since we can choose the kernels. Condition (B6) states the required range of magnitude of the bandwidths, and is easy to satisfy in practice.

The next two theorems establish uniform consistency and asymptotic normality of our estimators $\hat{m}_{\mathrm{NW}}$ and $\hat{m}_{\mathrm{LL}}$ defined in Section 3.2. Their proof can be found in Section 7 and in Section D in the supplementary file.

THEOREM 4.1. *Assume that* (2.2) *and Conditions* (B1)–(B6) hold and let $[a, b] \subseteq I_X \equiv \{x : f_X(x) > 0\}$.

(i) If $h \asymp g_1 \asymp g_2 \asymp (\log n)^{1/5} n^{-1/5}$, then $\hat{m}_{\mathrm{NW}}$ at (3.5) satisfies $\max_{x\in[a,b]} |\hat{m}_{\mathrm{NW}}(x) - m(x)| = O_P\{(\log n)^{2/5} n^{-2/5}\}$.

(ii) If $\beta_1 \geq 1/5$ and $0 < \alpha_0 < 1/2 - \beta_1$, then for every $x \in [a, b]$,

$$\hat{m}_{\mathrm{NW}}(x) - m(x) = \sqrt{V(x)}\, N(x) + B_0(x) + \tilde{B}(x) + R_0(x), \qquad (4.1)$$

where $N(x) \xrightarrow{\mathscr{D}} N(0,1)$ as $n \to \infty$, $V(x) = \{nhf_X(x)\}^{-1}\sigma^2(x) \int K^2$,

$B_0(x) = \{m''(x) + 2m'(x)f'_X(x)/f_X(x)\}\mu_{K,2}\, h^2/2$, $\tilde{B}(x) = \tilde{B}_\varphi(x) + \tilde{B}_\psi(x)$ with $\tilde{B}_\varphi(x) = xm'(x)E\{\varphi''(U)/\varphi(U)\}\mu_{L,2}\, g_1^2/2$, $\tilde{B}_\psi(x) = -m(x)$ $E\{\psi''(U)/\psi(U)\}\mu_{L,2}\, g_2^2/2$, and the remainder R_0 is such that $|R_0(x)| = o_P\{g_1^2 + g_2^2 + h^2 + (nh)^{-1/2}\}$.

THEOREM 4.2. *Assume that* (2.2) *and Conditions* (B1)–(B6) hold and let $[a, b] \subseteq I_X$.

(i) If $h \asymp g_1 \asymp g_2 \asymp (\log n)^{1/5} n^{-1/5}$, then $\hat{m}_{\mathrm{LL}}$ at (3.5) satisfies $\max_{x\in[a,b]} |\hat{m}_{\mathrm{LL}}(x) - m(x)| = O_P\{(\log n)^{2/5} n^{-2/5}\}$.

(ii) If $\beta_1 \geq 1/5$ and $0 < \alpha_0 < 1/2 - \beta_1$, then for every $x \in [a, b]$,

$$\hat{m}_{\mathrm{LL}}(x) - m(x) = \sqrt{V(x)}\, N(x) + B_1(x) + \tilde{B}(x) + R_1(x), \qquad (4.2)$$

where $N(x) \xrightarrow{\mathscr{D}} N(0, 1)$ as $n \to \infty$, $B_1(x) = m''(x)\mu_{K,2}\, h^2/2$, V and $\tilde{B}$ are as in part (ii) of Theorem 4.1, and R_1 is such that $|R_1(x)| = o_P\{g_1^2 + g_2^2 + h^2 + (nh)^{-1/2}\}$.

We deduce from the theorems that, although they are constructed from distorted data, when computed with appropriate bandwidths, our estimators $\hat{m}_{\mathrm{NW}}$ and $\hat{m}_{\mathrm{LL}}$ defined in Section 3.2 have the same uniform convergence rates as the standard estimators in (3.1) used when the (X_i, Y_i)'s are available. This contrasts with the errors-in-variables models studied by Fan and Truong (1993) and Delaigle et al. (2009), where convergence rates are significantly degraded by the measurement errors. The conclusions arising from the asymptotic distribution of our estimators are also interesting. Abusing terminology, we refer to V and $B_0 + \tilde{B}$ (resp., $B_1 + \tilde{B}$) as the asymptotic variance and bias and of our estimator $\hat{m}_{\mathrm{NW}}$ (resp., $\hat{m}_{\mathrm{LL}}$), and we call asymptotic mean squared error (AMSE) the sum of the asymptotic variance and squared bias. We use similar terminology for the standard estimators of m.

We learn from part (ii) of both theorems that, if we choose g_1 and g_2 of order $o(h)$, the asymptotic bias and variance of our estimators are identical to those of standard estimators, and there, as in the standard case, it is optimal to take $h \asymp n^{-1/5}$, so that AMSE $\asymp n^{-4/5}$. Perhaps more surprisingly, in cases where B_0 (resp., B_1 for $\hat{m}_{\mathrm{LL}}$), B_φ and B_ψ do not all have the same sign, it is possible to choose h and g_1 or g_2 an order of magnitude slightly larger than $n^{-1/5}$ such that the asymptotic bias $B_0 + \tilde{B}$ (resp., $B_1 + \tilde{B}$) vanishes and the AMSE our estimator is of order $o(n^{-4/5})$, thus smaller than the AMSE of the standard estimator (similar results can be established for the integrated AMSE). However, while it is theoretically interesting, we were not able to exploit this result in practice to make our estimator outperform

the standard one, despite several attempts. In part this is because to benefit from this result we need to choose the bandwidths in a very specialized way that requires estimating too many unknowns, and we found that the simpler bandwidths choice suggested in Section 5.2 almost always worked better.

Next, we develop theoretical properties of our estimator defined in Section 3.3. We start by rewriting $\hat{\mathcal{C}}_n(\rho_1,\rho_2)$ as $\hat{\mathcal{C}}_n(\rho_1,\rho_2) = \{1 \le i \le n : U_i \in \hat{\mathcal{L}}_n(\rho_1,\rho_2)\}$, where $\hat{\mathcal{L}}_n(\rho_1,\rho_2) = \{u \in I_U : |\hat{\varphi}_{0,\mathrm{LL}}(u)| \ge \rho_1, |\hat{\psi}_{0,\mathrm{LL}}(u)| \ge \rho_2\}$. We can rewrite the estimator at (3.7) as

$$\hat{m}_{\mathrm{NW}}(x;\rho_1,\rho_2) = \frac{\sum_{i=1}^n \hat{Y}_i K_h(x-\hat{X}_i) I\{U_i \in \hat{\mathcal{L}}_n(\rho_1,\rho_2)\}}{\sum_{i=1}^n K_h(x-\hat{X}_i) I\{U_i \in \hat{\mathcal{L}}_n(\rho_1,\rho_2)\}}.$$

To emphasize the main idea while avoiding repetitive arguments, here we present the theoretical result only for this estimator, assuming that only φ may have zeros, and therefore we take $\rho_2 = 0$ throughout this section. A straightforward adaptation of the arguments used to prove Theorem 4.3 below leads to similar results in the more general case where φ has zeros and $\rho_2 > 0$, and for the local linear estimator $\hat{m}_{\mathrm{LL}}(x;\rho_1,\rho_2)$.

When $\rho_2 = 0$, $\hat{\mathcal{L}}_n(\rho_1,\rho_2)$ depends only on ρ_1; to simplify notation we rewrite it as $\hat{\mathcal{L}}_n(\rho_1) = \{u \in I_U : |\hat{\varphi}_{0,\mathrm{LL}}(u)| \ge \rho_1\}$. Likewise, we rewrite $\hat{m}_{\mathrm{NW}}(x;\rho_1,0)$ as $\hat{m}_{\mathrm{NW}}(x;\rho_1)$. Under certain regularity conditions on φ, the random set $\hat{\mathcal{L}}_n(\rho_1)$ is a consistent estimator of $\mathcal{L}(\rho_1) = \{u \in I_U : |\varphi_0(u)| \ge \rho_1\}$. Recalling that $\varphi_0(u) = E(X)\,\varphi(u)$, this suggests taking ρ_1 to be some value between 0 and $M_0 \equiv |E(X)| \max_{u\in I_U} |\varphi(u)|$. For $0 \le t \le M_0$, let $\partial\mathcal{L}(t) = \{u \in I_U : |\varphi_0(u)| = t\}$. We will need the following assumptions:

(C1) $E(X), E(Y) \neq 0$ and $\inf_{u\in I_U} \psi(u) > 0$.

(C2) φ is such that the set $\Theta = \big\{t \in (0,M_0) : \partial\mathcal{L}(t)$ consists of finitely many points located in the interior of I_U and $\min_{u\in\partial\mathcal{L}(t)} |\varphi'(u)| > 0\big\}$ is non-empty.

The next theorem establishes uniform consistency and asymptotic normality of $\hat{m}_{\mathrm{NW}}(x;\rho)$. See Section E in the supplementary file for its proof.

THEOREM 4.3. *Assume that* (2.2), *Conditions* (B2)–(B5), (C1) and (C2) hold and that $\rho \in (0,M_0)$ in (3.7) is such that $\rho \in \Theta$. Let $[a,b] \subseteq I_X$.

(i) If $g_1 \asymp g_2 \asymp h \asymp (\log n)^{1/5} n^{-1/5}$, then $\hat{m}_{\mathrm{NW}}(x;\rho) \equiv \hat{m}_{\mathrm{NW}}(x;\rho,0)$ at (3.7) satisfies $\max_{x\in[a,b]} |\hat{m}_{\mathrm{NW}}(x;\rho) - m(x)| = O_P\{(\log n)^{2/5} n^{-2/5}\}$.

(ii) If $\beta_1 \ge 1/5$ and $0 < \alpha_0 < 1/2 - \beta_1$, then for every $x \in [a,b]$,

$$\hat{m}_{\mathrm{NW}}(x;\rho) - m(x) = \sqrt{V(x;\rho)}\, N(x) + B_0(x) + \tilde{B}(x;\rho) + R_2(x;\rho), \quad (4.3)$$

where $N(x) \xrightarrow{\mathscr{D}} N(0,1)$ as $n \to \infty$, $V(x;\rho) = V(x)/P\{U \in \mathcal{L}(\rho)\}$, V, B_0 are as in part (ii) of Theorem 4.1, $\tilde{B}(x;\rho) = xm'(x)E[\varphi''(U) I\{U \in \mathcal{L}(\rho)\}/\varphi(U)]\mu_{L,2}\, g_1^2/2 - m(x)E[\psi''(U)I\{U \in \mathcal{L}(\rho)\}/\psi(U)]\mu_{L,2}\, g_2^2/2$, and R_2 is such that $|R_2(x;\rho)| = o_P\{g_1^2 + g_2^2 + h^2 + (nh)^{-1/2}\}$.

We deduce from the theorem that our estimator defined in Section 3.3 has the same uniform convergence rate as the standard Nadaraya-Watson estimator in (3.1), used when the data (X_i, Y_i) are available. Moreover, as long as we choose g_1 and g_2 of order $o(h)$, the asymptotic "bias" and "variance" of our estimator from Section 3.3 are equal to those of the standard Nadaraya-Watson estimator, where $i \in \{1 \le j \le n : U_j \in \mathcal{L}_n(\rho)\}$. As we already indicated below Theorems 4.1 and 4.2, in theory in some cases it is possible to choose the bandwidths in such a way that the AMSE of our estimator tends to zero faster than that of the standard estimator, but it seems very hard to find a way to exploit this in practice. Similar results can be established for the local linear estimator $\hat{m}_{\mathrm{LL}}(x;\rho_1,\rho_2)$.

Establishing theoretical results for the more general procedure described in Section 3.4 is particularly challenging. Recall that this method combines a change point detection algorithm and the ridge-parameter based method introduced in Section 3.3. The complex nature of this approach implies that deriving its theoretical properties rigorously requires long and tedious arguments. Since our paper is already very long, and even the proofs for our simpler methods are fairly tedious, we leave such rigorous derivations for future work. However, our preliminary calculations already indicate that the procedure from Section 3.4 should have asymptotic properties similar to those described in Theorem 4.3. In particular, these calculations indicate that estimating the τ_j's and the sign of φ and/or ψ has no first order asymptotic effect on the properties of our estimators of m.

5. Numerical results.

5.1. *Which method to use.* The approach in Section 3.4 can be applied in essentially all cases, but since the methods from Sections 3.2 and 3.3 are simpler, the user might prefer to use these if all parts of (2.3) hold. While (2.3) can be verified by standard tests of hypothesis applied to the observed data (see Remark 5.1 below), when these conditions are needed, it is because the techniques employed involve dividing by estimators of ψ, φ, EX or EY. Therefore, in practice, to avoid numerical issues, we suggest using the method from Section 3.3, and to use instead the method from Section 3.4 if the absolute values of estimators of EX or EY are small, the extent of which depends on the magnitude of other quantities involved

and the precision of the software employed. This is generally rather easy to determine by examining the data, but if unsure the user can just apply the method of Section 3.4, which is valid in the most general case.

We note too that one does not necessarily need to predict the X_i's and the Y_i's with the same method. For example, if one is confident that EX is far from zero, but is not sure about EY, then the X_i's could be predicted using the approach from Section 3.3, and the predictors of the Y_i's could be obtained from the approach suggested in Section 3.4.

REMARK 5.1. The assumption at (2.3) can be tested in several ways. For example, since $E\tilde{X} = EX$, we can first test the sign of EX by a standard test of hypothesis for the mean applied to the data $\tilde{X}_1, \dots, \tilde{X}_n$, and then test the sign of the function $\varphi_0 = \varphi \cdot EX$ at (2.4), using for example tests such as those in Dümbgen and Spokoiny (2001), Chetverikov (2012) and Lee et al. (2013), applied to the observed data.

5.2. *Details of implementation.* As in the case where the (X_i, Y_i)'s are available, in practice we recommend using the local linear versions of our estimators, and in this section we suggest ways of choosing the parameters required to compute them. Similar ideas can be used for the Nadaraya-Watson estimators. We know from Section 4 that, while we have to choose h with care, we have more flexibility for the bandwidths g_1 and g_2, which can take a large range of values. If we take h to be of the standard size for nonparametric regression, and $g_1 = o(h)$ and $g_2 = o(h)$, then our estimators have the same first order asymptotic properties as the estimators at (3.1).

Motivated by this, for the estimators in Section 3.2, we take $g_1 = n^{-0.1} g_{1,\mathrm{PI}}$, $g_2 = n^{-0.1} g_{2,\mathrm{PI}}$ and $h = h_{\mathrm{PI}}$, where the subscript PI means that we use a standard plug-in bandwidth for local linear estimators (Ruppert et al., 1995) constructed based on, respectively, the data $(U_i, |\tilde{X}_i|)$, $(U_i, |\tilde{Y}_i|)$ and $(\hat{X}_i, \hat{Y}_i)$. For the estimators in Sections 3.3 and 3.4, we take $g_1 = n^{-1/10} g_{1,\mathrm{PI}}$ and $g_2 = n^{-1/10} g_{2,\mathrm{PI}}$, where $g_{1,\mathrm{PI}}$ and $g_{2,\mathrm{PI}}$ denote standard plug-in bandwidths for local linear estimators constructed based on, respectively, the data $(U_i, \tilde{X}_i)$ and $(U_i, \tilde{Y}_i)$. Then, in Section 3.3, we choose $\rho_1 = \max(0.1, \rho_1^*)$ and $\rho_2 = \max(0.1, \rho_2^*)$, where ρ_1^* (resp., ρ_2^*) denotes the square root of an estimator of the asymptotic "mean squared error" of $\hat{\varphi}_{\mathrm{LL}}$ (resp., $\hat{\psi}_{\mathrm{LL}}$), integrated over the set of x-values where $|\hat{\varphi}_{\mathrm{LL}}(x)|$ (resp., $|\hat{\psi}_{\mathrm{LL}}(x)|$) take its smallest values; see Appendix A in the supplementary file for details. We do the same for the method from Section 3.4, except that we use the estimators $\hat{\varphi}$ and $\hat{\psi}$ of φ and ψ derived there. Finally, we take $h = h_{\mathrm{PI}}$, a standard plug-in bandwidth for local linear estimators computed from the data $(\hat{X}_i, \hat{Y}_i)$, $i \in \hat{\mathcal{C}}_n(\rho_1, \rho_2)$.

The estimators from Section 3.4 also require to estimate the zeros $\tau_1, \dots, \tau_M$

at which φ changes sign, and the same is required for ψ if the method in that section is used to compute predictors of the Y_i's. We proceed as follows. First, since the τ_j's all correspond to a local minimum of φ^*, we find all the points at which $\hat{\varphi}^*_{\mathrm{LL}}$ has local minima. Then, among those points we keep only those which are close to the discontinuity points of the derivative φ^* detected by the method of Gijbels and Goderniaux (2005). Here we define "close" by less than $2h$ away, where h is the bandwidth in Section 2.2.1 of Gijbels and Goderniaux (2005). Finally, to slightly improve numerical performance, we implement Remark 3.2 and remove the data corresponding to the 5% smallest $\hat{f}_U(U_i)$'s.

5.3. *Simulations.* We applied our methods to a variety of simulated examples, ranging from the simplest ones in which $\psi > 0$ and $\varphi > 0$, where we can use the method from Section 3.2, to more complex ones in which $EX = 0$ and both ψ and φ oscillate between positive and negative values, where we need to use the sophisticated approach suggested in Section 3.4.

We generated data $(\tilde{X}_i, \tilde{Y}_i, U_i)$, $i = 1, \ldots, n$, from model (2.1) for $n = 100$, 200, 500 and 1000, and considered various combinations of m, φ, ψ and σ, and various distributions of X_i and U_i. We took $\varepsilon_i \sim N(0,1)$, and considered shifted versions of three regression curves m, denoted by m_1, m_2 and m_3 and defined as $m_1(x) = \sin\{\pi(x-1)/2\}/\big[\{1+2(x-1)^2\}\{\mathrm{sign}(x-1)+1\}\big]$, $m_2(x) = x^2\phi_{0,1}(x)$, and $m_3(x) = 2x + \phi_{0.5,0.1}(x)$, where $\phi_{\mu,\theta}$ denotes the density of a $N(\mu, \theta^2)$. In all cases below, the generic constant const. was chosen so that $E\{\varphi(U)\} = E\{\psi(U)\} = 1$.

First, we considered models where the local linear estimators from Sections 3.2 to 3.4 could all be applied: (i.a) $m = m_1$, $X_i \sim N(1, 1.5^2)$, $\sigma(x) = 0.3$; (ii.a) $m = m_2$, $X_i \sim N(1, 1.5^2)$, $\sigma(x) = 0.05$; (iii.a) $m = m_3$, $X_i \sim N(0.5, 0.75^2)$, $\sigma(x) = 0.55$; (i.b) $m(\cdot) = m_1(\cdot - 1) + 2$, $X_i \sim N(2, 1.5^2)$, $\sigma(x) = 0.3$; (ii.b) $m(\cdot) = m_2(\cdot - 1)$, $X_i \sim N(2, 1.5^2)$, $\sigma(x) = 0.05$; (iii.b) $m(\cdot) = m_3(\cdot - 1)$, $X_i \sim N(1.5, 0.75^2)$, $\sigma(x) = 0.55$. Each time we took $U_i \sim \beta(2,5)$, $\psi(u) = \text{const.}\,(u + 0.5)^2$ and $\varphi(u) = \text{const.}\,(u + 0.25)^2$.

Next, we considered models (i.c)–(iii.c) and (i.d)–(iii.d), where we took m, X_i and σ as in models (i.a)–(iii.a) and (i.b)–(iii.b), respectively, but took $U_i \sim \beta(3,5)$ and $\varphi(\cdot) = \psi(\cdot) = \text{const.}\, m_1(5 \cdot -2)$. Here φ and ψ have zeros and change signs, so that the method from Section 3.2 cannot be applied. Finally, in our last models, φ and ψ change signs and have several zeros and $E(X_i) = 0$, so that we can apply only the method from Section 3.4: (iv.a) $m(\cdot) = m_1(\cdot + 1)$, $X_i \sim N(0, 1.5^2)$, $\sigma(x) = 0.3$; (v.a) $m(\cdot) = m_2(\cdot + 1)$, $X_i \sim N(0, 1.5^2)$, $\sigma(x) = 0.05$; (vi.a) $m(\cdot) = m_3(\cdot + 0.5)$, $X_i \sim N(0, 0.75^2)$, $\sigma(x) = 0.55$; (iv.b) $m(\cdot) = m_1(\cdot)$, $X_i \sim \{\chi^2(4) - 4\}/2$, $\sigma(x) = 0.3$; (v.b)

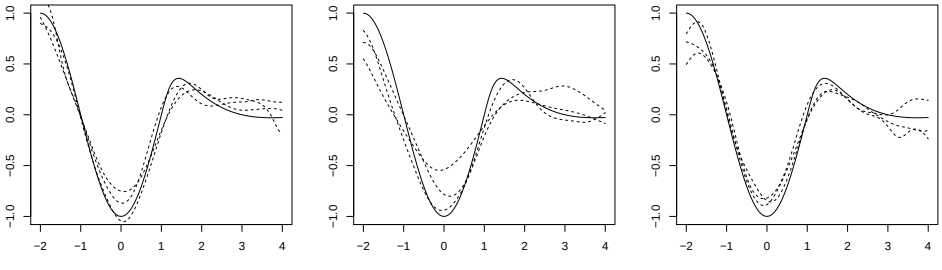

Fig 1: $\hat{m}_{\mathrm{LL}}$ from Section 3.2 (left), $\hat{m}_{\mathrm{LL}}(\cdot;\rho_1,\rho_2)$ from Section 3.3 (center) and estimator $\hat{m}_{\mathrm{LL}}(\cdot;\rho_1,\rho_2)$ from Section 3.4 (right) for three samples coming from model (i.a) with $n = 200$, and corresponding to the 1st, 2nd and 3rd quartiles of the ISEs. The continuous line depicts the true m.

$m(\cdot) = m_2(\cdot)$, $X_i \sim \{\chi^2(4) - 4\}/2$, $\sigma(x) = 0.05$; (vi.b) $m(\cdot) = m_3(\cdot)$, $X_i \sim \{\chi^2(4) - 4\}/3.5$, $\sigma(x) = 0.55$; Each time we took $U_i \sim \beta(3, 5)$ and $\varphi(\cdot) = \psi(\cdot) = \text{const.}\, m_1(5 \cdot -2)$. Heteroscedastic versions of these models gave similar results; see Appendix B in the supplementary file.

We compared each of our estimators with the ideal estimator $\tilde{m}_{\mathrm{LL}}$ at (3.1) computed from the (X_i, Y_i)'s, which are not available in real data applications but are available when we simulate data, and with the inconsistent naive estimator $\hat{m}_{\mathrm{LL,naive}}$, which is the standard local linear estimator computed from the contaminated $(\tilde{X}_i, \tilde{Y}_i)$'s. For each n and each model, we generated 1000 samples and constructed each estimator for each sample. Let $\hat{m}$ denote any one of the estimators considered below. To summarise the performance of $\hat{m}$, we computed, for each sample, the integrated squared error $\mathrm{ISE} = \int_a^b \{\hat{m}(x) - m(x)\}^2\, dx$, where, in each case, a and b were the quantiles 0.025 and 0.975 of the distribution of X.

In Tables 1 to 4 in Appendix B in the supplementary file, for each method we report the first, second and third quartiles of the resulting 1000 ISEs. See Appendix B for a detailed discussion of the simulation results. In summary, we found that, as expected, when φ, ψ, EX and EY were different from zero, but EX and/or EY were relatively close to zero, the estimator that worked best was the one from Section 3.2, but the most complex estimator from Section 3.4 worked well. When EX and EY were far from zero, all three estimators worked well, with the simplest one from Section 3.2 giving the best results and the one from Section 3.4 working the worst. When φ and/or ψ had zeros, the estimator from Section 3.2 could not be applied, and when EX and EY were close to zero, the best results were obtained with the estimator from Section 3.4, whereas when EX and EY were far from zero, the estimator from Section 3.3 worked best. Finally, we found

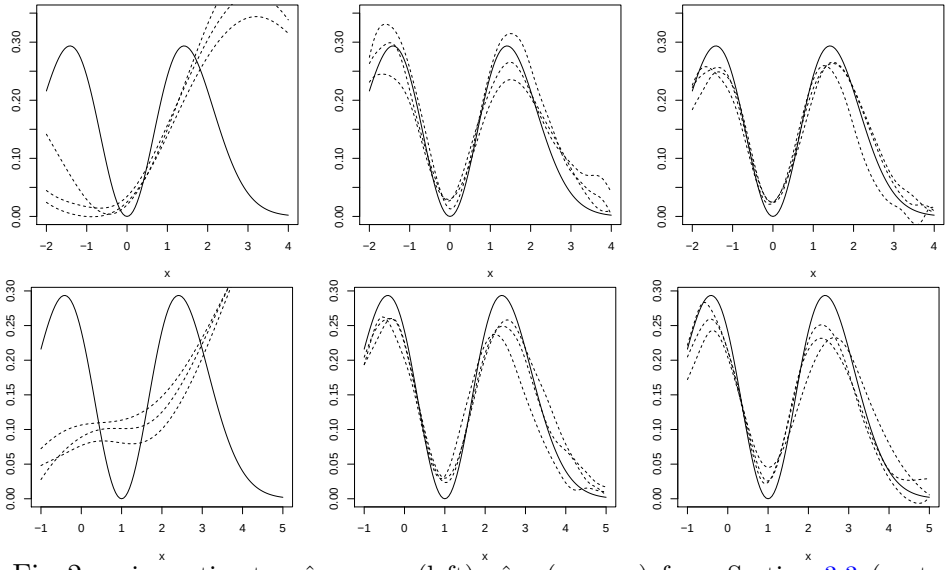

Fig 2: naive estimator $\hat{m}_{\mathrm{LL,naive}}$ (left), $\hat{m}_{\mathrm{LL}}(\cdot;\rho_1,\rho_2)$ from Section 3.3 (center), and estimator $\hat{m}_{\mathrm{LL}}(\cdot;\rho_1,\rho_2)$ from Section 3.4 (right) for three samples coming from model (ii.c) (top) and model (ii.d) (bottom) with $n = 500$, and corresponding to the 1st, 2nd and 3rd quartiles of the ISEs. The continuous line depicts the true m.

that our approach also performed well when the errors were heteroscedastic.

In all cases, our estimators performed considerably better than the naive estimator, but were of course outperformed by the oracle estimator. As expected, the performance of our estimators improved as sample size increased. In all our simulation settings, the estimator from Section 3.4 gave reasonable results. However, if φ and ψ were far from zero, we got better results by using the simplest estimator from Section 3.2, and if EX and EY were far from zero, we got better results using the estimator from Section 3.3.

To illustrate these results graphically, we present a few figures that are representative of the conclusions of our simulations. For each estimator $\hat{m}$ presented in the figures, we show the three estimated curves corresponding to the first three quartiles of the 1000 ISEs defined above. In Figure 1, using example (i.a), we illustrate the fact that, when all three methods can be applied, they often give similar results. Figure 2 shows estimated curves for examples (ii.c) and (ii.d). We can see that, in case (ii.c), where EX is close to zero, the estimator $\hat{m}_{\mathrm{LL}}(\cdot;\rho_1,\rho_2)$ from Section 3.4 worked better than the one from Section 3.3, but that the reverse is true in case (ii.d), where EX and EY are both far from zero. In that figure, we also depict the naive estimator $\hat{m}_{\mathrm{LL,naive}}$, which performed very poorly. Finally, in Figure 3,

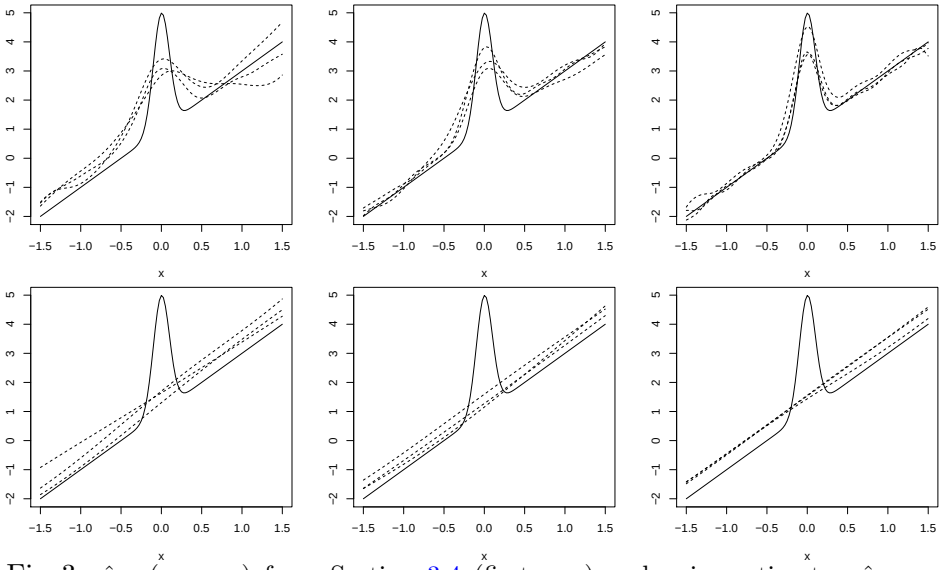

Fig 3: $\hat{m}_{\mathrm{LL}}(\cdot;\rho_1,\rho_2)$ from Section 3.4 (first row) and naive estimator $\hat{m}_{\mathrm{LL,naive}}$ (second row) for three samples coming from model (vi.a) with $n = 100$ (left), $n = 200$ (centre) and $n = 500$ (right), and corresponding to the 1st, 2nd and 3rd quartiles of the ISEs. The continuous line depicts the true m.

we use example (vi.a) to demonstrate the improvement that our estimator $\hat{m}_{\mathrm{LL}}(\cdot;\rho_1,\rho_2)$ from Section 3.4 benefits from as the sample size n increases. Here too, the naive estimator performed very poorly, even for n large.

5.4. *Real data illustrations.* We applied our new method to the Boston house-price dataset described in Harrison and Rubinfeld (1978), available at `http://lib.stat.cmu.edu/datasets`, and which contains information about houses and their owners at 506 locations around Boston. As in Şentürk and Müller (2005b), we are interested in the relationship between the median price (in USD 1000's) of houses, $\tilde{Y}$, and per capita crime rate by town, $\tilde{X}$, with the confounding effect of the proportion of population of lower educational status, U, removed. Şentürk and Müller's (2005b), whose interest was in the correlation between $\tilde{X}$ and $\tilde{Y}$, concluded that this correlation alters dramatically after adjusting for the confounding effect of lower educational status. On the left panel of Figure 4, we depict the covariate-adjusted regression curve obtained using the local linear estimator $\hat{m}_{\mathrm{LL}}(\cdot;\rho_1,\rho_2)$ from Section 3.3, the estimator $\hat{m}_{\mathrm{LL}}(\cdot;\rho_1,\rho_2)$ from Section 3.4, and the naive regression estimator $\hat{m}_{\mathrm{LL,naive}}$ obtained by regressing $\tilde{Y}$ on $\tilde{X}$ after removing a few outliers. In this example, the estimator from Section 3.2 was identical to the one from Section 3.3. We can see that $\hat{m}_{\mathrm{LL,naive}}$ indicates a pronounced

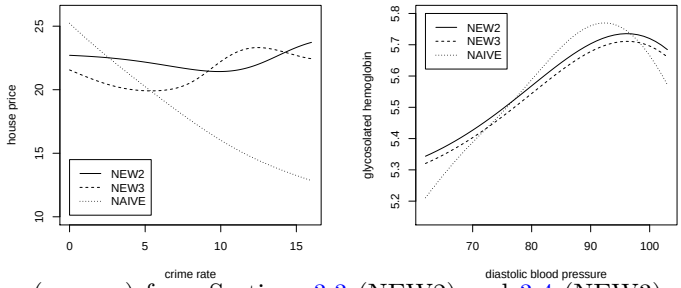

Fig 4: $\hat{m}_{\mathrm{LL}}(\cdot;\rho_1,\rho_2)$ from Sections 3.3 (NEW2) and 3.4 (NEW3), and naive estimator $\hat{m}_{\mathrm{LL,naive}}$ (NAIVE) for the Boston data (left) and the diabetes data (right).

relationship between house price and crime rate (as crime rate increases, house price decreases), but once we adjust for the effect of lower educational status, the regression curve obtained by both versions of our estimator is almost flat, indicating a weak relationship between the adjusted X and Y.

Next, we applied our procedure to the diabetes dataset used by Schorling et al. (1997) and Willems et al. (1997), available at `http://biostat.mc.vanderbilt.edu/DataSets`, which represents a subset of 403 individuals taken from a larger cohort of 1046 subjects who participated in a study for African Americans about obesity, diabetes and related factors in central Virginia. As in Şentürk and Nguyen (2006), our goal was to examine the relationship between glycosolated hemoglobin level $\tilde{Y}$, a biomarker for diabetes, and diastolic blood pressure $\tilde{X}$, adjusting for the effect of body mass index, U, which was found to be a confounder for both variables. As in Şentürk and Nguyen (2006), we removed a few outliers before our analysis. As in the previous example, $\widehat{EX}$ and $\widehat{EY}$ were far from zero, so that we used the estimator $\hat{m}_{\mathrm{LL}}(\cdot;\rho_1,\rho_2)$ from Section 3.3, which we compared with the naive estimator $\hat{m}_{\mathrm{LL,naive}}$. Here too, the estimator from Section 3.2 was identical to the one from Section 3.3. We also computed the estimator from $\hat{m}_{\mathrm{LL}}(\cdot;\rho_1,\rho_2)$ from Section 3.4. These estimators, depicted on the right panel of Figure 4, show that after adjusting for body mass index, the relationship between glycosolated hemoglobin level and diastolic blood pressure is noticeably less pronounced. We should highlight that, in this example, the data were rather sparse for diastolic blood pressure greater than 100, and the few patients for which $\tilde{X}$ was greater than 100 had a rather low value of $\tilde{Y}$, whence the decreasing shape on the right hand side of the graph, which may just be an artifact of the sparseness of the data in that area.

Another interesting application of our method is to the baseline data collected from studies A and B of the Modification of Diet in Renal Disease Study (Levey et al., 1994). The nonlinear relationship between the baseline

unadjusted glomerular filtration rate (GFR) and serum creatinine (SCr) is of particular interest. Taking body surface area (BSA) as the confounder, Cui et al. (2009) used a parametric nonlinear model of the form $m(x) = \beta_1 \exp(-\beta_2 - \beta_3 x^2) + \beta_4$ to study the relationship between GFR and SCr after correcting for the distorting effect of BSA. Because this dataset is not publicly available, we shall not compare the proposed nonparametric method with that of Cui et al. (2009) in this paper.

6. Generalizations to the multivariate case. Our approach can be generalized to the d-variate case, $d \geq 1$, where we observe data distributed like a vector $(U, \tilde{\mathbf{X}}^{\mathrm{T}}, \tilde{Y})$, with $\tilde{\mathbf{X}} \in \mathbb{R}^d$ a distorted version of $\mathbf{X} \in \mathbb{R}^d$. Reflecting the fact that the components of $\tilde{\mathbf{X}}$ may not all be distorted, we write $d = d_1 + d_2$, with $d_1 \geq 0$ and $d_2 \geq 1$, and let $\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, \mathbf{X}_2^{\mathrm{T}})^{\mathrm{T}}$ and $\tilde{\mathbf{X}} = (\mathbf{X}_1^{\mathrm{T}}, \tilde{\mathbf{X}}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{X}_1 = (X_1, \ldots, X_{d_1})^{\mathrm{T}}$ and $\tilde{\mathbf{X}}_2 = (\tilde{X}_{d_1+1}, \ldots, \tilde{X}_d)^{\mathrm{T}}$ is a distorted version of $\mathbf{X}_2 = (X_{d_1+1}, \ldots, X_d)^{\mathrm{T}}$, and where we use the convention that $\mathbf{X} = \mathbf{X}_2$ if $d_1 = 0$. In this notation, the data $\{(U_i, \tilde{Y}_i, \mathbf{X}_{1i}^{\mathrm{T}}, \tilde{\mathbf{X}}_{2i}^{\mathrm{T}})\}_{i=1}^n$ we observe are generated by the model

$$\begin{cases} Y = m(\mathbf{X}) + \varepsilon\, \sigma(\mathbf{X}), \\ \tilde{Y} = \psi(U)\, Y, \quad \tilde{X}_{d_1+r} = \varphi_r(U)\, X_{d_1+r}, \quad r = 1, \ldots, d_2, \end{cases} \tag{6.1}$$

where $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ is a curve we wish to estimate, the random variables $\mathbf{X}$, U and ε are mutually independent, $E(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = 1$. As in (2.2), we assume that $E\{\psi(U)\} = 1$, $E\{\varphi_r(U)\} = 1$, for $r = 1, \ldots, d_2$.

The procedures from Section 3.2 to 3.4 can each be generalized to the multivariate setting, but for space constraint here we show only how to generalize the approach from Section 3.2. The same ideas can be applied for the methods from Sections 3.3 and 3.4. To construct a nonparametric version of the estimator from Section 3.2, we first construct predictors $\hat{Y}_i$ and $\hat{X}_{i,d_1+1}, \ldots, \hat{X}_{id}$ as in equation (3.4), and let

$$\hat{\mathbf{X}}_i = (X_{i1}, \ldots, X_{id_1}, \hat{X}_{i,d_1+1}, \ldots, \hat{X}_{id})^{\mathrm{T}}.$$

Next, we use a standard multivariate local linear regression estimator applied to the data $(\hat{\mathbf{X}}_i^{\mathrm{T}}, \hat{Y}_i)$. That is, we define (see Fan and Gijbels, 1996) $\hat{m}_{\mathrm{LL}}(\mathbf{x}) = \hat{\alpha}_0$, where $(\hat{\alpha}_0, \hat{\alpha}_1) = \arg\min_{\alpha_0 \in \mathbb{R}, \alpha_1 \in \mathbb{R}^d} \sum_{i=1}^n \{\hat{Y}_i - \alpha_0 - \alpha_1^{\mathrm{T}}(\hat{\mathbf{X}}_i - \mathbf{x})\}^2 \mathbf{K}_{\mathbf{h}}(\hat{\mathbf{X}}_i - \mathbf{x})$, with $\mathbf{K}_{\mathbf{h}}(\mathbf{x}) = \prod_{r=1}^d h_r^{-1} K(x_r/h_r)$ a d-dimensional product kernel, K a univariate kernel, and $\mathbf{h} = (h_1, \ldots, h_d)^{\mathrm{T}}$ a vector of bandwidths.

It is well known that fully nonparametric estimators suffer from the curse of dimensionality, which means that as d increases, such estimators can only work reasonably well if the sample size is very large. To overcome

this problem, a common approach is to restrict the regression model so that only univariate curves have to be fitted. A popular example is the additive model (Hastie and Tibshirani, 1990), which assumes that $m(\mathbf{X}) = m_0 + \sum_{j=1}^{d} m_j(X_j)$. In our context, the additive covariate-adjusted regression model can be written as

$$\begin{cases} Y = m_0 + \sum_{j=1}^{d} m_j(X_j) + \varepsilon\, \sigma(\mathbf{X}), \\ \tilde{Y} = \psi(U)\, Y, \quad \tilde{X}_{d_1+r} = \varphi_r(U)\, X_{d_1+r}, \quad r = 1, \ldots, d_2, \end{cases} \tag{6.2}$$

where $m_1, \ldots, m_d$ are unknown univariate functions satisfying $E\{m_j(X_j)\} = 0$ for $j = 1, \ldots, d$ and m_0 is an unknown parameter.

In the standard setting where the $(\mathbf{X}_i^{\mathrm{T}}, Y_i)$'s are directly observed, there are several ways to fit the additive model; see Horowitz (2014) for an overview of estimation and inference for nonparametric additive models. The simplest approach is to adapt to our setting the iterative backfitting algorithm of Buja et al. (1989), as follows. First, let $\hat{m}_0 = n^{-1} \sum_{i=1}^{n} \tilde{Y}_i$ and $\hat{m}_j \equiv 0$ for $j = 1, \ldots, d$. For $j = 1, \ldots, d$, update $\hat{m}_j$ by taking it equal to a local linear regression estimator using the data $\{(\hat{X}_{ij}, \hat{Y}_i - \hat{m}_0 - \sum_{k \neq j} \hat{m}_j(\hat{X}_{ik}))\}_{i=1}^{n}$. Iterate until the estimates $\hat{m}_j$ stabilize. (Here $\hat{X}_{ij} = X_{ij}$ if $j \leq d_1$.)

Alternatively, instead of taking $\hat{m}_j = 0$ as initial estimators, we could start with a linear approximation of the model in (6.2). See Appendix C in the supplementary file for details. We could also apply similar transformations to other existing methods for fitting additive models, such as the approach suggested by Horowitz and Mammen (2004). The main theoretical challenge is a delicate analysis on how the presence of generated response and predictors affects the first order asymptotic properties of the final estimators. However, deriving such results requires much more work than can possibly done in this paper, and so we leave this problem for future research. The method proposed in this section can be applied to creatinine data, which was analyzed by Şentürk and Müller (2006). In this study, serum creatinine level is taken as the response and the two predictors include cholesterol level and serum albumin level. The confounder variable U is taken to be body mass index defined as weight/height^2. The readers can find more details about this dataset in Şentürk and Müller (2006).

7. Proof of Theorem 4.1. We start by introducing basic notations. For a kernel function K, we write $\mu_{K,\ell} = \int u^\ell\, K(u)\, du$ for non-negative integers ℓ. For any set S, we denote its complement by S^{c} and its cardinality by $\#S$. Throughout, we let const. denote a finite positive constant independent of n, which may take different values at each occurrence. We also use the

following notation: $\mu_0 = E(X)$, $m_0 = E(Y)$, $\mu_0^+ = E|X|$, $m_0^+ = E|Y|$ and

$$\varphi_0 = \mu_0\,\varphi, \quad \psi_0 = m_0\,\psi, \quad \varphi_0^+ = \mu_0^+\,\varphi, \quad \psi_0^+ = m_0^+\,\psi. \tag{7.1}$$

We proceed with the proof of Theorem 4.1. For $u \in I_U = [0,1]$, write

$$w_0(u) \equiv 1,\ \hat{w}_X(u) = \hat{\mu}_0^+\,\varphi(u)/\hat{\varphi}_{0,\mathrm{LL}}^+(u)\ ,\ \hat{w}_Y(u) = \hat{m}_0^+\,\psi(u)/\hat{\psi}_{0,\mathrm{LL}}^+(u), \tag{7.2}$$

where $\hat{\mu}_0^+ = \widehat{E|X|} = n^{-1}\sum_{i=1}^n |\tilde{X}_i|$, $\hat{m}_0^+ = \widehat{E|Y|} = n^{-1}\sum_{i=1}^n |\tilde{Y}_i|$ and $\hat{\varphi}_{0,\mathrm{LL}}^+$ and $\hat{\psi}_{0,\mathrm{LL}}^+$ are local linear estimators of φ_0^+ and ψ_0^+ defined below (3.3).

Noting the model at (2.1), and hence by (3.4) and (7.2),

$$\hat{X}_i = X_i\hat{w}_X(U_i), \quad \hat{Y}_i = Y_i\,\hat{w}_Y(U_i). \tag{7.3}$$

Substituting the expressions in (7.3) into (3.5) gives

$$\begin{aligned}\hat{m}_{\mathrm{NW}}(x) - m(x) &= \{n\hat{f}_{\hat{X}}(x)\}^{-1}\sum_{i=1}^n K_h(x-\hat{X}_i)\{\hat{w}_Y(U_i) - w_0(U_i)\}Y_i \\ &+\{n\hat{f}_{\hat{X}}(x)\}^{-1}\sum_{i=1}^n K_h(x-\hat{X}_i)\{m(X_i) - m(x)\} + \{n\hat{f}_{\hat{X}}(x)\}^{-1}\sum_{i=1}^n K_h(x-\hat{X}_i) \\ &\times \sigma(X_i)\varepsilon_i \equiv \hat{\Pi}_{01}(x) + \hat{\Pi}_{02}(x) + \hat{\Pi}_{03}(x),\end{aligned} \tag{7.4}$$

where

$$\hat{f}_{\hat{X}}(x) \equiv n^{-1}\sum_{i=1}^n K_h(x-\hat{X}_i). \tag{7.5}$$

Proof of (i).

We start by establishing uniform bounds for $\hat{w}_X$ and $\hat{w}_Y$ which will be useful throughout the proof. Recalling that the U_i's are supported on $I_U = [0,1]$, for $Z = Y$ or $Z = X$ we use the notation $\|\hat{w}_Z - w_0\|_\infty = \sup_{u\in[0,1]}|\hat{w}_Z(u) - w_0(u)|$. To derive our bounds, note that under Conditions (B1)–(B6), for $\ell = 0, 1, 2$, we have (Masry, 1996; Hansen, 2008)

$$\begin{aligned}\sup_{u\in[0,1]}\left|\hat{\varphi}_{0,LL}^{+\,(\ell)}(u) - \varphi_0^{+\,(\ell)}(u)\right| &= O_P\{\delta_{\ell n}(g_1)\}, \\ \sup_{u\in[0,1]}\left|\hat{\psi}_{0,LL}^{+\,(\ell)}(u) - \psi_0^{+\,(\ell)}(u)\right| &= O_P\{\delta_{\ell n}(g_2)\},\end{aligned} \tag{7.6}$$

where, for all $t > 0$,

$$\delta_{\ell n}(t) \equiv t^2 + (nt^{2\ell+1})^{-1/2}(\log n)^{1/2}. \tag{7.7}$$

In particular, for $g_1 = g_{1n} \asymp n^{-\beta_1}$ and $g_2 = g_{2n} \asymp n^{-\beta_2}$, we have $\delta_{0n}(g_1) = O(n^{-\lambda_1}\sqrt{\log n})$ and $\delta_{0n}(g_2) = O(n^{-\lambda_2}\sqrt{\log n})$, where $\lambda_\nu \equiv \min(2\beta_\nu, 1/2 - \beta_\nu/2) \in (0, 2/5]$, for $\nu = 1, 2$.

Now, using (7.1) and (7.2), we can write

$$\hat{w}_X(u) - w_0(u) = \frac{\hat{\mu}_0^+ \varphi(u) - \hat{\varphi}_{0,\mathrm{LL}}^+(u)}{\hat{\varphi}_{0,\mathrm{LL}}^+(u)} = \frac{(\hat{\mu}_0^+ - \mu_0^+)\varphi(u)}{\hat{\varphi}_{0,\mathrm{LL}}^+(u)} + \frac{\varphi_0^+(u) - \hat{\varphi}_{0,\mathrm{LL}}^+(u)}{\hat{\varphi}_{0,\mathrm{LL}}^+(u)}, \tag{7.8}$$

and a similar equation can be written for $\hat{w}_Y$.

Since, by Condition (B1), $\gamma_1 \equiv \min_{u\in[0,1]} \min\{|\varphi_0^+(u)|, |\psi_0^+(u)|\} > 0$, a direct consequence of (7.6) and Taylor expansion is that $\{\hat{\varphi}_{0,\mathrm{LL}}^+(u)\}^{-1} = \{\varphi_0^+(u) + \hat{\varphi}_{0,\mathrm{LL}}^+(u) - \varphi_0^+(u)\}^{-1} = \{\varphi_0^+(u)\}^{-1} + O_P\{\delta_{0n}(g_1)\}$ uniformly over $u \in [0, 1]$. Moreover, we also have $\hat{\mu}_0^+ = \mu_0^+ + O_P(n^{-1/2})$ and $\hat{m}_0^+ = m_0^+ + O_P(n^{-1/2})$. Substituting the previous two displays into (7.8) gives, for $Z_1 = X$ and $Z_2 = Y$,

$$\|\hat{w}_{Z_\nu} - w_0\|_\infty = O_P\{\delta_{0n}(g_\nu)\} = O_P\{n^{-\lambda_\nu}(\log n)^{1/2}\}. \tag{7.9}$$

Later in our proof, it will also be useful to use the fact that $\delta_{0n}(g_1) = o(h)$ because $\alpha_0 < 2\beta_1$.

Next we study the common denominator $\hat{f}_{\hat{X}}(x)$ of $\hat{\Pi}_{01}(x)$, $\hat{\Pi}_{02}(x)$ and $\hat{\Pi}_{03}(x)$. Let

$$\hat{f}_X(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i) \tag{7.10}$$

denote the standard kernel estimator of $f_X(x)$ that we would use if the X_i's were available. For this estimator, it is well known (see e.g. Theorem 6 in Hansen, 2008) that $\max_{x\in[a,b]} |\hat{f}_X(x) - f_X(x)| = O_P\{\delta_{0n}(h)\}$. Shortly we shall prove that

$$\max_{x\in[a,b]} \left|\hat{f}_{\hat{X}}(x) - \hat{f}_X(x)\right| = O_P\{h^{-1}\delta_{0n}(g_1)\} = o_P(1), \tag{7.11}$$

which further leads to $\max_{x\in[a,b]} |\hat{f}_{\hat{X}}(x) - f_X(x)| = O_P\{h^{-1}\delta_{0n}(g_1) + \delta_{0n}(h)\} = o_P(1)$. In turn, using arguments similar to those we used above to treat the denominator of $\hat{w}_{Z_\nu} - w_0$, and taking into account the fact that $\min_{x\in[a,b]} f_X(x) > 0$, we obtain

$$\begin{aligned}\{\hat{f}_{\hat{X}}(x)\}^{-1} &= \{f_X(x) + \hat{f}_{\hat{X}}(x) - f_X(x)\}^{-1} \\ &= \{f_X(x)\}^{-1} + O_P\{h^{-1}\delta_{0n}(g_1)\} = \{f_X(x)\}^{-1} + o_P(1)\end{aligned} \tag{7.12}$$

uniformly over $x \in [a, b]$, and that $\max_{x\in[a,b]}\{\hat{f}_{\hat{X}}(x)\}^{-1} = O_P(1)$.

Next we prove (7.11). For this, note that for any $C > 0$, we can write

$$P\Big\{\max_{x\in[a,b]}\big|\hat{f}_{\hat{X}}(x)-\hat{f}_X(x)\big|>Ch^{-1}\delta_{0n}(g_1)\Big\}$$
$$\le P\Big\{\max_{x\in[a,b]}\big|\hat{f}_{\hat{X}}(x)-\hat{f}_X(x)\big|>Ch^{-1}\delta_{0n}(g_1),\mathcal{A}_n\Big\}+P(\mathcal{A}_n^{\mathrm{c}}),$$

where $\mathcal{A}_n$ is an event that we shall define below, and which is such that $P(\mathcal{A}_n)\to 1$ as $n\to\infty$. Therefore, to prove (7.11), it suffices to handle the first term on the right side of the inequality above. Towards this end, first, comparing the definitions (7.5) and (7.10) we see that for each $x\in\mathbb{R}$,

$$\big|\hat{f}_{\hat{X}}(x)-\hat{f}_X(x)\big|\le\|K'\|_\infty\frac{\|\hat{w}_X-w_0\|_\infty}{nh^2}\sum_{i=1}^n|X_i|I\big(|X_i-x|\le h\text{ or }|\hat{X}_i-x|\le h\big).\tag{7.13}$$

To further bound the right side of (7.13), we shall show that $\hat{X}_i$ and X_i are uniformly close (see (7.16) below) as long as the estimation error of $\hat{w}_X$ is well-controlled. To see this, for $\lambda\ge 0$, define the event

$$\mathcal{E}_n(\lambda)=\big\{\|\hat{w}_X-w_0\|_\infty\le n^{-\lambda}\big\}.\tag{7.14}$$

By (7.9), we have $P\{\mathcal{E}_n(\lambda)\}\to 1$ as $n\to\infty$ provided that $\lambda<\lambda_1$. Moreover, define events

$$\mathcal{E}_{1n}(\lambda)=\Big\{\max_{1\le i\le n}|X_i|\le\lambda\log n\Big\}\quad\text{and}\quad\mathcal{E}_{2n}(\lambda)=\Big\{\max_{1\le i\le n}|\varepsilon_i|\le\lambda\log n\Big\}.\tag{7.15}$$

In the proof of Lemma F.1 in the supplementary file, we shall show that for every given $c>0$, there exist a constant $C_1>0$ such that $P\{\mathcal{E}_{1n}(C_1)\}\ge 1-\text{const.}\,n^{-c}$.

Let $\alpha\in(\alpha_0,\lambda_1)$ be a constant, such that under Condition (B6), $n^{-\alpha}=o(h)$ and $P\{\mathcal{E}_n(\alpha)^{\mathrm{c}}\}\to 0$ as $n\to\infty$. On the event $\mathcal{E}_n(\alpha)\cap\mathcal{E}_{1n}(C_1)$, we have

$$\max_{1\le i\le n}|\hat{X}_i-X_i|\le\|\hat{w}_X-w_0\|_\infty\max_{1\le i\le n}|X_i|\le C_1n^{-\alpha}\log n,\tag{7.16}$$

such that for every $x\in[a,b]$, $|X_i-x|\le|\hat{X}_i-x|+C_1n^{-\alpha}\log n$. Therefore, on the event $\mathcal{E}_n(\alpha)\cap\mathcal{E}_{1n}(C_1)$ with n sufficiently large,

$$I\big(|\hat{X}_i-x|\le h\big)\le I(|X_i-x|\le 2h).\tag{7.17}$$

It follows from (7.13) and (7.17) that, on $\mathcal{E}_n(\alpha)\cap\mathcal{E}_{1n}(C_1)$ with n large enough,

$$\max_{x\in[a,b]}\big|\hat{f}_{\hat{X}}(x)-\hat{f}_X(x)\big|$$

$$\le \|K'\|_\infty \|\hat{w}_X - w_0\|_\infty (nh^2)^{-1} \max_{x\in[a,b]} \sum_{i=1}^n |X_i| I(|X_i - x| \le 2h)$$

$$\le \|K'\|_\infty \|\hat{w}_X - w_0\|_\infty (nh^2)^{-1} \max_{x\in[a,b]} (|x| + 2h) \sum_{i=1}^n I(|X_i - x| \le 2h)$$

$$\le \text{const.} \|\hat{w}_X - w_0\|_\infty h^{-2} \max_{x\in[a,b]} \{\hat{F}_X(x + 2h) - \hat{F}_X(x - 2h)\}, \tag{7.18}$$

where $\hat{F}_X(x) = n^{-1} \sum_{i=1}^n I(X_i \le x)$ denotes the empirical distribution function. To further bound the right-hand side of (7.18), we let F_X be the distribution function of X and then apply the Dvoretzky-Kiefer-Wolfwitz inequality (Massart, 1990) to obtain that $P(\sqrt{n}\|\hat{F}_X - F_X\|_\infty > y) \le 2\exp(-2y^2)$ for all $y > 0$, where $\|\hat{F}_X - F_X\|_\infty \equiv \sup_{x\in\mathbb{R}} |\hat{F}_X(x) - F_X(x)|$. For $\lambda > 0$, define the event

$$\mathcal{E}_{0n}(\lambda) = \Big\{\sqrt{n}\|\hat{F}_X - F_X\|_\infty \le (\lambda \log n)^{1/2}\Big\}, \tag{7.19}$$

such that $P\{\mathcal{E}_{0n}(1/2)\} \ge 1 - 2n^{-1}$. Under Condition (B3), we deduce that on the $\mathcal{E}_{0n}(1/2)$ with n sufficiently large,

$$\max_{x\in[a,b]} \{\hat{F}_X(x + 2h) - \hat{F}_X(x - 2h)\} \le \max_{x\in[a,b]} \{F_X(x + 2h) - F_X(x - 2h)\}$$
$$+ \{2(\log n)/n\}^{1/2} \le 4\|f_X\|_\infty h + \{2(\log n)/n\}^{1/2} \le \text{const.}\, h. \tag{7.20}$$

Substituting this into (7.18) and taking $\mathcal{A}_n \equiv \mathcal{E}_n(\alpha) \cap \mathcal{E}_{0n}(1/2) \cap \mathcal{E}_{1n}(C_1)$ imply that for all sufficiently large n,

$$P\Big\{\max_{x\in[a,b]} |\hat{f}_{\hat{X}}(x) - \hat{f}_X(x)| > Ch^{-1}\delta_{0n}(g_1), \mathcal{A}_n\Big\}$$
$$\le P\big\{\|\hat{w}_X - w_0\|_\infty > \text{const.}\delta_{0n}(g_1), \mathcal{A}_n\big\}$$
$$\le P\big\{\|\hat{w}_X - w_0\|_\infty > \text{const.}\, \delta_{0n}(g_1)\big\}, \tag{7.21}$$

and that $P(\mathcal{A}_n^c) \to 0$ as $n \to \infty$. Together, (7.9) and (7.21) prove (7.11).

Next we study $\hat{\Pi}_{01}(x)$. For this, we first write $\hat{f}_{\hat{X}}(x)\,\hat{\Pi}_{01}(x)$ as

$$n^{-1} \sum_{i=1}^n K_h(x - X_i)(\hat{w}_Y - w_0)(U_i)\, Y_i + n^{-1} \sum_{i=1}^n \{K_h(x - \hat{X}_i) - K_h(x - X_i)\}$$
$$\times (\hat{w}_Y - w_0)(U_i)\, Y_i \equiv J_1(x) + J_2(x). \tag{7.22}$$

Applying Lemma F.4 with $g_2 \asymp n^{-\beta_2}$ to $J_1(x)$ implies

$$\max_{x\in[a,b]} |J_1(x)| = O_P(g_2^2) = O_P(n^{-2\beta_2}). \tag{7.23}$$

For $J_2(x)$, note that $J_2(x) \le \|K'\|_\infty \|\hat{w}_X - w_0\|_\infty \|\hat{w}_Y - w_0\|_\infty (nh^2)^{-1} \sum_{i=1}^n |X_i|\{|m(X_i)| + \sigma(X_i)|\varepsilon_i|\} I\big(|X_i - x| \le h \text{ or } |\hat{X}_i - x| \le h\big)$. The argument leading to (7.11) can be used to prove that $\max_{x\in[a,b]}(nh)^{-1} \sum_{i=1}^n |X_i\, m(X_i)|$ $I\big(|X_i - x| \le h \text{ or } |\hat{X}_i - x| \le h\big) = O_P(1)$ and the same bound holds if the $m(X_i)$'s are replaced by the $\sigma(X_i)$'s. Moreover, similarly to (F.9) in the proof of Lemma F.1, it can be proved that

$$\max_{1\le i\le n} |\varepsilon_i| = O_P(\log n). \tag{7.24}$$

This, together with (7.9) and the two displays before (7.24) yields

$$\max_{x\in[a,b]} |J_2(x)| = O_P\{h^{-1}\delta_{0n}(g_1)\delta_{0n}(g_2)\log n\} = o_P\{g_2^2 + (ng_2)^{-1/2}\}. \tag{7.25}$$

Here, the last step follows from Condition (B6) and the assumption that $\alpha_0 < 2\beta_1$. Together, (7.12), (7.22), (7.23) and (7.25) imply

$$\max_{x\in[a,b]} \big|\hat{\Pi}_{01}(x)\big| = O_P(g_2^2) + o_P\{(ng_2)^{-1/2}\}. \tag{7.26}$$

For $\hat{\Pi}_{02}(x)$, we write $K_h(x - \hat{X}_i)$ in $\hat{f}_X(x)\,\hat{\Pi}_{02}(x)$ as $K_h(x - \hat{X}_i) - K_h(x - X_i) + K_h(x - X_i)$. A similar argument to what we used to study (7.13) gives

$$\begin{aligned}
&\max_{x\in[a,b]} \Big|n^{-1}\sum_{i=1}^n \{K_h(x - \hat{X}_i) - K_h(x - X_i)\}\{m(X_i) - m(x)\}\Big| \le \|m'\|_\infty \|K'\|_\infty \\
&\times \frac{\|\hat{w}_X - w_0\|_\infty}{nh^2} \max_{x\in[a,b]} \Big|\sum_{i=1}^n |X_i(X_i - x)|\, I\big(|X_i - x| \le h \text{ or } |\hat{X}_i - x| \le h\big)\Big| \\
&= O_P\{\delta_{n,0}(g_1)\}.
\end{aligned} \tag{7.27}$$

Together with (7.7) and (7.12), this implies

$$\max_{x\in[a,b]} \big|\hat{\Pi}_{02}(x) - \Pi_{02}(x)\big| = O_P\{\delta_{n,0}(g_1)\}, \tag{7.28}$$

where $\Pi_{02}(x) \equiv \{n\hat{f}_{\hat{X}}(x)\}^{-1} \sum_{i=1}^n K_h(x - X_i)\{m(X_i) - m(x)\}$.

Next, we write $\hat{f}_{\hat{X}}(x)\,\Pi_{02}(x) = n^{-1}\sum_{i=1}^n K_h(x - X_i)\{m(X_i) - m(x)\}$ as

$$n^{-1}\sum_{i=1}^n \{g_{n,i}(x) - Eg_{n,i}(x)\} + n^{-1}\sum_{i=1}^n Eg_{n,i}(x) \equiv R_n(x) + n^{-1}\sum_{i=1}^n Eg_{n,i}(x), \tag{7.29}$$

where $g_{n,i}(x) = K_h(x - X_i)\{m(X_i) - m(x)\}$. To bound $\max_{x\in[a,b]} |R_n(x)|$, we create a grid using N points of the form $x_j = a + j\epsilon$ with $\epsilon = (b - a)/N$

for some $N \geq 1$ to be determined below (7.32). Since $g'_{n,i}(x) = h^{-1}K'_h(x - X_i)\{m(X_i) - m(x)\} - m'(x)K_h(x - X_i)$, by the mean value theorem we have, for every $x, y \in \mathbb{R}$, $|g_{n,i}(x) - g_{n,i}(y)| \leq (\|K\|_\infty + \|K'\|_\infty)\|m'\|_\infty h^{-1}|x - y|$. Therefore,

$$\max_{x\in[a,b]} |R_n(x)| \leq \max_{1\leq j\leq N} |R_n(x_j)| + 2(\|K\|_\infty + \|K'\|_\infty)\|m'\|_\infty \epsilon h^{-1}. \quad (7.30)$$

For each $x \in \mathbb{R}$ fixed, $g_{n,1}(x), \ldots, g_{n,n}(x)$ are independent random variables satisfying $|g_{n,i}(x)| \leq \|K\|_\infty\|m'\|_\infty$ and $E\{g_{n,i}(x)\}^2 = h^{-1}\int K^2(t)\{m(x - ht) - m(x)\}^2 f_X(x - ht)\,dt \leq \|m'\|_\infty^2\|f_X\|_\infty h\int t^2K^2(t)\,dt$. Hence, by Bernstein's inequality and Boole's inequality, for every $y \geq 0$,

$$\begin{aligned} P\Big\{\max_{1\leq j\leq N}|R_n(x_j)| \geq y\Big\} &\leq \sum_{j=1}^N P\Big[\Big|n^{-1}\sum_{i=1}^n\{g_{n,i}(x_j) - Eg_{n,i}(x_j)\}\Big| \geq y\Big] \\ &\leq 2N\exp\Big\{-\frac{ny^2}{2(c_K^2\|m'\|_\infty^2\|f_X\|_\infty h + \|K\|_\infty\|m'\|_\infty y/3)}\Big\}, \quad (7.31) \end{aligned}$$

where $c_K \equiv \{\int t^2K^2(t)\,dt\}^{1/2}$. For every $\lambda > 0$, define the event

$$\mathcal{C}_n(\lambda) = \Big\{\max_{1\leq j\leq N}|R_n(x_j)| \leq c_K\|m'\|_\infty\|f_X\|_\infty^{1/2}\sqrt{\frac{h\lambda}{n}} + \|K\|_\infty\|m'\|_\infty\frac{\lambda}{n}\Big\}, \quad (7.32)$$

such that in view of (7.31), $P\{\mathcal{C}_n(\lambda)^c\} \leq 2N\exp(-\tau\lambda)$ for some absolute constant $\tau > 0$. By taking $N = n$ and $\lambda = 2\tau^{-1}\log n$, it follows from (7.30) and (7.32) that

$$\max_{x\in[a,b]}|R_n(x)| = O_P\{h^{1/2}(n/\log n)^{-1/2} + n^{-1}\log n + (nh)^{-1}\}. \quad (7.33)$$

For the second term on the right-hand side of (7.29), standard arguments show that, under Conditions (B3) and (B5),

$$Eg_{n,i}(x) = \{m''(x)f_X(x)/2 + m'(x)f'_X(x)\}\mu_{K,2}\,h^2 + o(h^2) \quad (7.34)$$

uniformly in $x \in [a, b]$. Consequently, combining (7.12), (7.28), (7.33) and (7.34), we get

$$\max_{x\in[a,b]}\big|\hat{\Pi}_{02}(x)\big| = O_P(h^2). \quad (7.35)$$

For the last term $\hat{\Pi}_{03}(x)$ in (7.4), we need to control the stochastic error

$$\Delta_{n,\infty} \equiv \max_{x\in[a,b]}\Big|n^{-1}\sum_{i=1}^n K_h(x - \hat{X}_i)\sigma(X_i)\varepsilon_i\Big| \quad (7.36)$$

for $\hat{X}_i = X_i\hat{w}_X(U_i)$ as in (7.3). To this end, we shall use a lattice argument by making a finite approximation of the compact interval $[a, b]$ using a sequence $\{x_j\}_{j=1}^N$ of equidistant points $x_j = a+j\epsilon$ for $\epsilon = (b-a)/N$, and then discretize $\Delta_{n,\infty}$ to define $\Delta_{n,N} \equiv \max_{1\le j\le N} \left|n^{-1}\sum_{i=1}^n K_h(x_j - \hat{X}_i)\sigma(X_i)\varepsilon_i\right|$. Here, N is a positive integer that will be determined after (7.44).

Instead of dealing with $\Delta_{n,\infty}$ directly, we shall prove that $\Delta_{n,N}$ provides a fine approximation to $\Delta_{n,\infty}$, at least with high probability, and then restrict attention to $\Delta_{n,N}$. By definition of $\Delta_{n,N}$, we have $|\Delta_{n,\infty} - \Delta_{n,N}| \le \|\sigma\|_\infty\|K'\|_\infty \epsilon h^{-2}\max_{1\le i\le n}|\varepsilon_i|$. Together with (7.24), this leads to

$$|\Delta_{n,\infty} - \Delta_{n,N}| = O_P(N^{-1}h^{-2}\log n). \tag{7.37}$$

For $\Delta_{n,N}$, shortly we shall prove by taking $N = n$ that

$$\Delta_{n,N} = O_P\{(nh/\log n)^{-1/2}\}, \tag{7.38}$$

which together with (7.37) leads to

$$\Delta_{n,\infty} = O_P\{(nh/\log n)^{-1/2} + (nh^2)^{-1}\log n\} = O_P\{(nh/\log n)^{-1/2}\}, \tag{7.39}$$

where the last step relies on the identity $(nh^2)^{-1}\log n = (nh/\log n)^{-1/2}(nh^3/\log n)^{-1/2}$ and Condition (B6). Combing (7.12) and (7.39) yields

$$\max_{x\in[a,b]}\left|\hat{\Pi}_{03}(x)\right| \le \max_{x\in[a,b]}\{\hat{f}_{\hat{X}}(x)\}^{-1}\Delta_{n,\infty} = O_P\{(nh/\log n)^{-1/2}\}. \tag{7.40}$$

Together, (7.26), (7.35) and (7.40) complete the proof of (4.1).

Next we prove (7.38). For $\lambda > 0$, let $V_{1n}(x) = \{\sum_{i=1}^n K_h^2(x-\hat{X}_i)\sigma^2(X_i)\}^{1/2}$, $V_{2n}(x) = \max_{1\le i\le n} K_h(x - \hat{X}_i)\sigma(X_i)$ and define the event

$$\mathcal{D}_n(N,\lambda) = \left\{|\Delta_{n,N}| \le \max_{1\le j\le N} V_{1n}(x_j)\sqrt{\lambda}/n + \max_{1\le k\le N} V_{2n}(x_j)\lambda/n\right\}. \tag{7.41}$$

To deal with $V_{1n}(x)$, as in the proof of (7.21), put $\mathcal{A}_n = \mathcal{E}_n(\alpha)\cap\mathcal{E}_{1n}(C_1)\cap\mathcal{E}_{0n}(1/2)$ with $\alpha \in (\alpha_0, \lambda_1)$ such that $P(\mathcal{A}_n^c) \to 0$ as $n \to \infty$, where $\mathcal{E}_n(\alpha)$, $\mathcal{E}_{1n}(C_1)$ and $\mathcal{E}_{0n}(1/2)$ are as in (7.14), (7.15) and (7.19), respectively. On the event $\mathcal{A}_n$ with n sufficiently large, it follows from (7.17) and (7.20) that

$$\begin{aligned}\max_{1\le j\le N} V_{1n}(x_j) \le \max_{x\in[a,b]} V_{1n}(x) &\le \frac{\|\sigma\|_\infty\|K\|_\infty}{h}\max_{x\in[a,b]}\sqrt{\sum_{i=1}^n I(|X_i - x| \le 2h)} \\ &\le \text{const.}\|\sigma\|_\infty\|K\|_\infty(n/h)^{1/2}.\end{aligned} \tag{7.42}$$

It is easy to see that $\max_{x\in[a,b]} V_{2n}(x) \le \|\sigma\|_\infty \|K\|_\infty h^{-1}$. This, combined with (7.41) and (7.42) yields, on the event $\mathcal{D}_n(N,\lambda)\cap\mathcal{A}_n$ with n large enough,

$$\Delta_{n,N} \le \text{const.}\|\sigma\|_\infty\|K\|_\infty\big\{\sqrt{\lambda/(nh)} + \lambda/(nh)\big\}. \tag{7.43}$$

Next we show that for properly chosen N and λ, $P\{\mathcal{D}_n(N,\lambda)^{\mathrm{c}}\} \to 0$ as $n\to\infty$. Observe that $\hat{w}_X$ defined in (7.2) is a measurable function of $\{(X_i,U_i)\}_{i=1}^n$ and thus is independent of $\{\varepsilon_i\}_{i=1}^n$. Conditional on $\{(X_i,U_i)\}_{i=1}^n$, taking $\mathbf{a} = (a_1,\dots,a_n)^{\mathrm{T}} = \big(K_h(x-\hat{X}_i)\sigma(X_i),\dots,K_h(x-\hat{X}_n)\sigma(X_n)\big)^{\mathrm{T}}$ in Lemma F.2 and using Boole's inequality, we obtain that for every $\lambda\ge 0$, $P\big[\Delta_{n,N} > \max_{1\le j\le N} V_{1n}(x_j)\sqrt{\lambda}/n + \max_{1\le k\le N} V_{2n}(x_j)\,\lambda/n \,\big|\{(X_i,U_i)\}_{i=1}^n\big] \le 2N\exp(-c\lambda)$ where $c>0$ is a constant independent of n and N. Taking expectations on both sides of the inequality gives that for every $\lambda\ge 0$, $P\{\mathcal{D}_n(N,\lambda)^{\mathrm{c}}\} \le 2N\exp(-c\lambda)$. Taking $N=n$ and $\lambda = 2c^{-1}\log n$ we get

$$P\{\mathcal{D}_n(n,\lambda)^{\mathrm{c}}\} \le 2n^{-1}. \tag{7.44}$$

Combining (7.43) with $N=n, \lambda = 2c^{-1}\log n$, (7.44) and the fact that $P(\mathcal{A}_n^{\mathrm{c}})\to 0$ proves (7.38) as claimed.

Proof of (ii).

To prove the asymptotic normality, we need to use a more refined argument. In what follows, $x\in[a,b]$ is fixed and we deal with the sum in (7.4) over each $\hat{\Pi}_{0j}(x)$ separately.

First, for $\hat{\Pi}_{01}(x)$, recall in (7.22) that $\hat{f}_{\hat{X}}(x)\,\hat{\Pi}_{01}(x) = J_1(x)+J_2(x)$. By (7.25) and Condition (B6), $|J_2(x)| = o_P\{h^{-1}\delta_{0n}(g_1)\delta_{0n}(g_2)\log n\} = o_P(g_1^2+g_2^2)$. For $J_1(x)$, Lemma F.4 with $g_{2n}\asymp n^{-\beta_2}$ implies $J_1(x) = -\frac{1}{2}m(x)f_X(x)\,E\{\psi''(U)/\psi(U)\}\mu_{L,2}\,g_2^2 + o_P(g_2^2)$. The last two displays and (7.12) imply

$$\hat{\Pi}_{01}(x) = -m(x)E\{\psi''(U)/\psi(U)\}\mu_{L,2}\,g_2^2/2 + o_P(g_1^2+g_2^2). \tag{7.45}$$

For $\hat{\Pi}_{02}(x)$, by a first-order Taylor's expansion we obtain

$$\begin{aligned}
\hat{f}_{\hat{X}}(x)\,\hat{\Pi}_{02}(x) &= n^{-1}\sum_{i=1}^n K_h(x-\hat{X}_i)\{m(X_i)-m(x)\}\\
&= n^{-1}\sum_{i=1}^n K_h(x-X_i)\{m(X_i)-m(x)\} + (nh^2)^{-1}\sum_{i=1}^n K'\Big(\frac{x-X_i}{h}\Big)\\
&\quad\times (w_0-\hat{w}_X)(U_i)\,X_i\{m(X_i)-m(x)\} + (2nh^3)^{-1}\sum_{i=1}^n K''(\xi_n)\\
&\quad\times (w_0-\hat{w}_X)^2(U_i)\,X_i^2\{m(X_i)-m(x)\}I\big(|\hat{X}_i-x|\le h\big)\\
&\equiv I_1(x)+I_2(x)+I_3(x),
\end{aligned} \tag{7.46}$$

where ξ_n is a random variable that lies between $(x - X_i)/h$ and $(x - \hat{X}_i)/h$.

A standard argument shows that $I_1(x) = O_P\{h^2 + (nh)^{-1/2}\}$. Together with (7.12), this yields

$$\begin{aligned}\{\hat{f}_{\hat{X}}(x)\}^{-1} I_1(x) &= \{f_X(x)\}^{-1} I_1(x) + O_P[h^{-1}\delta_{n,0}(g_1)\{h^2 + (nh)^{-1/2}\}] \\ &= \{f_X(x)\}^{-1} I_1(x) + o_P\{h^2 + (nh)^{-1/2}\}. \end{aligned} \tag{7.47}$$

For $I_2(x)$, it follows from (F.12) in Lemma F.3 and (7.12) that $\{\hat{f}_{\hat{X}}(x)\}^{-1} I_2(x)$ $= xm'(x)E\{\varphi''(U)/\varphi(U)\}\mu_{L,2}\, g_1^2/2 + o_P(g_1^2)$. For $I_3(x)$, a similar argument to that leading to (7.11) yields $\max_{x\in[a,b]} |I_3(x)|3 = O_P\{h^{-1}\delta_{n,0}^2(g_1)\}$ and hence, $\{\hat{f}_{\hat{X}}(x)\}^{-1} I_3(x) = o_P\{(nh)^{-1/2}\}$. Combining with with (7.46) we get

$$\begin{aligned}\hat{\Pi}_{02}(x) =& \{f_X(x)\}^{-1} I_1(x) + xm'(x)E\{\varphi''(U)/\varphi(U)\}\mu_{L,2}\, g_1^2/2 \\ &+ o_P\{g_1^2 + h^2 + (nh)^{-1/2}\} \end{aligned} \tag{7.48}$$

for $I_1(x)$ as in (7.46). Finally, for the stochastic error term $\hat{\Pi}_{03}(x)$, we shall use an argument similar to that employed in Mammen et al. (2012) based on empirical process theory. Write $\beta_1 = (1 + \xi_0)/5$ for some $\xi_0 \geq 0$. First, we argue that the estimator $\hat{w}_X$ falls within a "nice" function space, the complexity of which can be measured via covering numbers. Let $\mathcal{M}_{0n}$ be the set of functions $[0, 1] \mapsto \mathbb{R}$ whose derivatives up to order two exist and are uniformly bounded in order by $(ng_1^5/\log n)^{-1/2} \asymp n^{\xi_0/2}(\log n)^{1/2}$. Since $\beta_1 \geq 1/5$, we have $\lambda_1 = \min(2\beta_1, 1/2 - \beta_1/2) = 1/2 - \beta_1/2$. For some $\alpha \in (\alpha_0, 1/2 - \beta_1/2)$ to be specified in the paragraph after (7.53), we define the following set of functions:

$$\mathcal{N}_{0n} = \big\{w \in \mathcal{M}_{0n} : \|w - w_0\|_\infty \leq n^{-\alpha}\big\}. \tag{7.49}$$

By (7.6), using the same argument that we used to derive (7.9), we have $P(\hat{w}_X \in \mathcal{N}_{0n}) \to 1$ as $n \to \infty$.

Note that $\hat{f}_{\hat{X}}(x)\,\hat{\Pi}_{03}(x)$ in (7.4) can be written as $n^{-1}\sum_{i=1}^n\{K_h(x - \hat{X}_i) - K_h(x - X_i)\}\sigma(X_i)\varepsilon_i + n^{-1}\sum_{i=1}^n K_h(x - X_i)\sigma(X_i)\varepsilon_i$. For the first term, by Lemma F.1 we have, for any $\kappa_1 \in (0, 1/2 + 3\alpha/4 - 3\alpha_0/2 - \xi_0/8)$, $n^{-1}\sum_{i=1}^n\{K_h(x - \hat{X}_i) - K_h(x - X_i)\}\sigma(X_i)\varepsilon_i = O_P(n^{-\kappa_1})$. On the other hand, it is straightforward to show that $n^{-1}\sum_{i=1}^n K_h(x - X_i)\sigma(X_i)\varepsilon_i = O_P\{(nh)^{-1/2}\} = O_P(n^{-1/2+\alpha_0/2})$. Combining this and (7.12), we get

$$\begin{aligned}\hat{\Pi}_{03}(x) =& \{nf_X(x)\}^{-1}\sum_{i=1}^n K_h(x - X_i)\sigma(X_i)\varepsilon_i \\ &+ O_P\{n^{-\kappa_1} + n^{-1+\beta_1/2+3\alpha_0/2}(\log n)^{1/2}\}\end{aligned}$$

$$=\{nf_X(x)\}^{-1}\sum_{i=1}^{n}K_h(x-X_i)\sigma(X_i)\varepsilon_i+O_P(n^{-\kappa_1}). \tag{7.50}$$

Assembling (7.45), (7.48) and (7.50) we obtain that, for any $\alpha\in(\alpha_0,1/2-\beta_1/2)$ and $\kappa_1\in(0,1/2+3\alpha/4-3\alpha_0/2-\xi_0/8)$,

$$\begin{aligned}\hat m_{\mathrm{NW}}(x)-m(x)=&\tilde B(x)+\{f_X(x)\}^{-1}I_1(x)+\sqrt{V(x)}\,N(x)\\&+o_P(n^{-\kappa_1}+g_1^2+g_2^2+h^2),\end{aligned} \tag{7.51}$$

where $\tilde B(x)$ and $I_1(x)$ are as in part (ii) of Theorem 4.1 and (7.46), respectively, and $N(x)\equiv\{V(x)\}^{-1/2}\{nf_X(x)\}^{-1}\sum_{i=1}^n K_h(x-X_i)\sigma(X_i)\varepsilon_i$ for $V(x)$ is as in part (ii) of Theorem 4.1. Further, for $I_1(x)=n^{-1}\sum_{i=1}^n K_h(x-X_i)\{m(X_i)-m(x)\}$, proceeding as in (7.29) we derive that

$$\{f_X(x)\}^{-1}I_1(x)=B_0(x)+o_P\{h^2+(nh)^{-1/2}\} \tag{7.52}$$

for $B_0(x)$ as in part (ii) of Theorem 4.1. For the third addend on the right-hand side of (7.51), Lyapounov's central limit theorem combined with Slutsky's theorem yield

$$N(x)\xrightarrow{\mathscr{D}}N(0,1),\quad\text{as } n\to\infty. \tag{7.53}$$

In particular, for $h=h_n\asymp n^{-\alpha_0}$ with $\alpha_0\in(0,1/2-\beta_1)$, by taking α and κ_1 in such a way that $\frac{4}{3}\alpha_0<\alpha<\frac12-\frac12\beta_1$ and $\frac12-\frac12\alpha_0<\kappa_1<\frac12+\frac34\alpha-\frac32\alpha_0-\frac18\xi_0$, we have $n^{-\kappa_1}=o\{(nh)^{-1/2}\}$. This, together with (7.51)–(7.53) proves (4.1). □

Acknowledgement. We thank three referees and an Associate Editor for their helpful comments which led to an improved version of the manuscript. This research was supported by the Australian Research Council.

SUPPLEMENTARY MATERIAL

Supplement to "Nonparametric covariate-adjusted regression" (; .pdf). This supplemental material contains more details for the implementation of the proposed estimators, additional simulation results as well as additional proofs omitted in the main text.

References.

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

[2] Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83** 1184–1186.

[3] Chetverikov, D. (2012). Adaptive test of conditional moment inequalities. Available at *arXiv:1201.0167*.
[4] Cui, X., Guo, W., Lin, L. and Zhu, L. (2009). Covariate-adjusted nonlinear regression. *Ann. Statist.* **37** 1839–1870.
[5] Delaigle, A., Fan, J. and Carroll, R. J. (2009). A design-adaptive local polynomial estimator for the errors-in-variables problem. *J. Amer. Statist. Assoc.* **104** 348–359.
[6] Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.* **29** 124–152.
[7] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman & Hall, London.
[8] Fan, J. and Truong, Y. K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **23** 1900–1925.
[9] Gijbels, I., Hall, P. and Kneip, A. (1999). On the estimation of jump points in smooth curves. *Ann. Inst. Statist. Math.* **51** 231–251.
[10] Gijbels, I. and Goderniaux, A.-C. (2005). Data-driven discontinuity detection in derivatives of a regression function. *Comm. Statist. Theory Methods* **33** 851–871.
[11] Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24** 726–748.
[12] Harrison, D. and Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. *J. Environ. Econ. Manag.* **5** 81–102.
[13] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman & Hall, London.
[14] Horowitz, J. L. (2014). Nonparametric Additive Models. In *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics* (J. Racine, L. Su, and A. Ullah, eds.), 129–148. Oxford University Press, Oxford.
[15] Horowitz, J. L. and Mammen, E. (2004). Nonparametric estimation of an additive model with a link function. *Ann. Statist.* **32** 2412–2443.
[16] Lee, S., Song, K. and Whang, Y.-J. (2013). Testing functional inequalities. *J. Econometrics* **172** 14–32.
[17] Levey, A. S., Adler, S., Beck, G. J. et al. (1994). The effects of dietary protein restriction and blood pressure control on the progression of renal disease. *N. Engl. J. Med.* **330** 877–884.
[18] Mammen, E., Rothe, C. and Schienle, M. (2012). Nonparametric regression with nonparametrically generated covariates. *Ann. Statist.* **40** 1132–1170.
[19] Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal.* **17** 571–599.
[20] Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfwitz inequality. *Ann. Probab.* **18** 1269–1283.
[21] Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90** 257–1270.
[22] Şentürk, D. (2006). Covariate-adjusted varying coefficient models. *Biostatistics* **7** 235–251.
[23] Şentürk, D. and Müller, H.-G. (2005a). Covariate-adjusted regression. *Biometrika* **92** 75–89.
[24] Şentürk, D. and Müller, H.-G. (2005b). Covariate adjusted correlation analysis via varying coefficient models. *Scand. J. Stat.* **32** 365–383.
[25] Şentürk, D. and Müller, H.-G. (2006). Inference for covariate adjusted regression via varying coefficient models. *Ann. Statist.* **34** 654–679.
[26] Şentürk, D. and Müller, H.-G. (2009). Covariate-adjusted generalized linear mod-

els. *Biometrika* **96** 357–370.
[27] Şentürk, D. and Nguyen, D. V. (2006). Estimation in covariate-adjusted regression. *Comput. Statist. Data Anal.* **50** 3294–3310.
[28] Schorling, J. B., Roach, J., Siegel, M., Baturka, N., Hunt, D. E., Guterbock, T. M. and Stewart, H. L. (1997). A trial of church-based smoking cessation interventions for rural African Americans. *Preventive Med.* **26** 92–101.
[29] Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.
[30] Willems J. P., Saunders J. T., Hunt, D. E. and Schorling, J. B. (1997). Prevalence of coronary heart disease risk factors among rural blacks: a community-based study. *Southern Med. J.* **90** 814–820.
[31] Zhang, J., Yu, Y., Zhu, L. X. and Liang, H. (2013). Partial linear single index models with distortion measurement errors. *Ann. Inst. Statist. Math.* **65** 237–267.
[32] Zhang, J., Zhu, L. X. and Liang, H. (2012). Nonlinear models with measurement errors subject to single-indexed distortion. *J. Multivariate Anal.* **112** 1–23.

Australian Research Council Centre
of Excellence for Mathematical
and Statistical Frontiers (ACEMS)
and School of Mathematics and Statistics
University of Melbourne
Parkville, Victoria 3010
Australia
E-mail: A.Delaigle@ms.unimelb.edu.au
wenxinz@princeton.edu

Department of Operations Research
and Financial Engineering
Princeton University
Princeton, New Jersey 08544
USA
E-mail: wenxinz@princeton.edu