Statistica Sinica **17**(2007), 797-816

BACKFITTING VERSUS PROFILING IN GENERAL CRITERION FUNCTIONS

Ingrid Van Keilegom and Raymond J. Carroll

Université catholique de Louvain and Texas A&M University

Abstract: We study the backfitting and profile methods for general criterion functions that depend on a parameter of interest β and a nuisance function θ. We show that when different amounts of smoothing are employed for each method to estimate the function θ, the two estimation procedures produce estimators of β with the same limiting distributions, even when the criterion functions are non-smooth in β and/or θ. The results are applied to a partially linear median regression model and a change point model, both examples of non-smooth criterion functions.

Key words and phrases: Backfitting, change points, dioxin, kernel estimation, median regression, nonparametric regression, partially linear model, profile kernel methods, semiparametric estimation, undersmoothing.

1. Introduction

Consider a semiparametric problem that depends on a parameter β_0 and an unknown function $\theta_0(\cdot)$. The purpose of this paper is to compare backfitting and profiling methods in semiparametric regression. Our context is quite general and allows for estimation based on non-smooth criterion functions.

We first introduce the context of the general problem. Assume that the data (X_i, Y_i) $(i = 1, \ldots, n)$ are independent replications of a $(1+d_y)$-dimensional random vector (X, Y). Let β denote a $q \times 1$ vector of parameters of interest, with true value β_0, belonging to a compact subset $\mathcal{B}$ of $I\!R^q$. Let $\theta = \theta(\cdot) : I\!R \to I\!R$ be an infinite dimensional 'nuisance' parameter with true value $\theta_0(\cdot)$, and let $\mathcal{L}\{Y, \beta_0, \theta_0(X)\}$ be a real-valued maximizing function for β_0 and θ_0, in the sense that $E[\mathcal{L}_\beta\{Y, \beta_0, \theta_0(X)\}] = 0$ and $E[\mathcal{L}_\theta\{Y, \beta_0, \theta_0(x)\}|X = x] = 0$ for all x, where $\mathcal{L}_\beta\{y, \beta, \theta(x)\}$ denotes the vector of partial derivatives of $\mathcal{L}\{y, \beta, \theta(x)\}$ with respect to the components of β, and $\mathcal{L}_\theta\{y, \beta, \theta(x)\}$ denotes the partial derivative of $\mathcal{L}(y, \beta, z)$ with respect to z, and evaluated at $z = \theta(x)$. Inference for β_0 is then carried out by maximizing

$$n^{-1} \sum_{i=1}^{n} \mathcal{L}\{Y_i, \beta, \theta(X_i)\} \tag{1}$$

with respect to β for some $\theta(\cdot)$.

An important example, which we treat in detail in this paper, is given by the partially linear regression model

$$\mathcal{Y}_i = Z_i^T\beta_0 + \theta_0(X_i) + \epsilon_i, \tag{2}$$

where Z_i is a possibly vector-valued covariate of dimension q and X_i is a scalar covariate. In our general notation, $Y = (\mathcal{Y}, Z)$. When $\text{med}(\epsilon_i|X_i, Z_i) = 0$, the maximizing function is given by

$$\mathcal{L}\{Y, \beta, \theta(X)\} = -|\mathcal{Y} - Z^T\beta - \theta(X)|. \tag{3}$$

Note that $\mathcal{L}_\beta\{Y, \beta, \theta(X)\}$ and $\mathcal{L}_\theta\{Y, \beta, \theta(X)\}$ do not exist when $\mathcal{Y} - Z^T\beta - \theta(X) = 0$, a point that complicates the theory.

In the semiparametric literature, two approaches have been considered to maximize (1), these approaches differing in the way they treat the unknown function $\theta_0(\cdot)$.

The backfitting procedure has been investigated by many authors in special contexts, including Rice (1986), Speckman (1988), Buja, Hastie and Tibshirani (1989), Hastie and Tibshirani (1990), Opsomer and Ruppert (1997, 1999), Mammen, Linton and Nielsen (1999), Wand (1999) and Opsomer (2000). The basic idea is one of iteration. For any given β, let $\widehat{\theta}(\cdot, \beta)$ be an estimate of $\theta_0(\cdot)$: the estimator we use is defined in the next section. Then define the criterion function

$$m_{BF}\{y, \beta, \theta(x)\} = \mathcal{L}_\beta\{y, \beta, \theta(x)\}.$$

The backfitting estimator $\widehat{\beta}_{BF}$ is now defined by the value of β that minimizes

$$\|n^{-1}\sum_{i=1}^n m_{BF}\{Y_i, \beta, \widehat{\theta}(X_i, \beta)\}\| \tag{4}$$

over $\mathcal{B}$, where $\|\cdot\|$ is the Euclidean norm in $I\!R^q$.

The profile method also has a large literature, see for example Severini and Wong (1992), Severini and Staniswalis (1994), Carroll, Fan, Gijbels and Wand (1997) and Murphy and van der Vaart (2000), among many others. This method again starts with $\widehat{\theta}(x, \beta)$, which may be different from that constructed for backfitting, but it obtains the estimate of β differently. Specifically, it creates a criterion function m_{PR} by differentiating $\mathcal{L}\{y, \beta, \theta(x, \beta)\}$ with respect to β, i.e.,

$$\begin{aligned} m_{PR}\{y, \beta, \theta(x, \beta), \theta_\beta(x, \beta)\} &= \frac{d}{d\beta}\mathcal{L}\{y, \beta, \theta(x, \beta)\} \\ &= \mathcal{L}_\beta\{y, \beta, \theta(x, \beta)\} + \mathcal{L}_\theta\{y, \beta, \theta(x, \beta)\}\frac{\partial}{\partial\beta}\theta(x, \beta), \end{aligned}$$

where $\theta_\beta(x,\beta) = \frac{\partial}{\partial\beta}\theta(x,\beta)$ for any $\theta(x,\beta)$. With these definitions, the profile estimator $\widehat{\beta}_{PR}$ is the value of β in $\mathcal{B}$ for which

$$\|n^{-1}\sum_{i=1}^{n} m_{PR}\{Y_i,\beta,\widehat{\theta}(X_i,\beta),\widehat{\theta}_\beta(X_i,\beta)\}\| \tag{5}$$

is minimal, where $\widehat{\theta}(\cdot,\beta)$ and $\widehat{\theta}_\beta(\cdot,\beta)$ are defined in Section 2.

For the example of a partially linear median regression model, see (2) and (3),

$$\begin{aligned} &m_{BF}\{Y,\beta,\theta(X,\beta)\} = [2I\{\mathcal{Y} - Z^T\beta - \theta(X,\beta) > 0\} - 1]Z,\\ &m_{PR}\{Y,\beta,\theta(X,\beta),\theta_\beta(X,\beta)\}\\ &\quad = [2I\{\mathcal{Y} - Z^T\beta - \theta(X,\beta) > 0\} - 1]\{Z + \theta_\beta(X,\beta)\}. \end{aligned}$$

It is clear that the development of the properties of $\widehat{\beta}_{BF}$ and $\widehat{\beta}_{PR}$ is challenging in this example, given that the nuisance function $\theta(X,\beta)$ is inside an indicator function, whereas this is not the case for a mean regression model. The aim of this paper is to study the asymptotic properties of $\widehat{\beta}_{BF}$ and $\widehat{\beta}_{PR}$ when the functions m_{BF} and m_{PR} are not necessarily smooth, as in the above example.

The comparison of backfitting and profiling has been the subject of some limited research. Consider a Gaussian model with independent data, scalar response $\mathcal{Y}_i$ and predictors (Z_i, X_i), so that in our context $Y_i = (\mathcal{Y}_i, Z_i)$, and suppose that the true mean is $Z_i^{\mathrm{T}}\beta_0 + \theta_0(X_i)$. Opsomer and Ruppert (1999) showed that under certain conditions, backfitting and profiling produce asymptotically equivalent estimators, but only when backfitting an estimated function $\widehat{\theta}(x,\beta)$ undersmoothed compared to that used by profiling. In more global contexts, with correlated data and multiple arguments for the function, backfitting and profiling are no longer necessarily asymptotically equivalent, see Hu, Wang and Carroll (2004) for a counterexample.

In this note we study the two methods when the criterion functions $m_{BF}(y,\beta,z)$ and $m_{PR}(y,\beta,z,z')$ are not necessarily smooth in β, z and/or z', and when θ is estimated by kernel smoothing. We prove that, under certain regularity conditions, the two methods have the same asymptotic distribution, but only when the driving estimation methods $\widehat{\theta}(x,\beta)$ employ different amounts of smoothness.

The paper is organized as follows. In the next section we introduce notation and develop the general conditions under which the estimators $\widehat{\beta}_{BF}$ and $\widehat{\beta}_{PR}$ are asymptotically normal. We show that under certain primitive conditions, the profile estimator and the backfitting estimator have the same asymptotic variance. Section 3 gives two applications. The first, discussed in Section 3.1, deals with the application of the general theory to the partially linear median regression model, defined in (2) and (3). The second, in Section 3.2, is concerned with a varying coefficient change point model, motivated by a problem in toxicology. The proofs and the conditions under which the main results for backfitting and profiling are valid are given in Appendix A and B, respectively.

2. Main Results

Let $K_h(u) = h^{-1}K(u/h)$, K a symmetric kernel density function and h a smoothing parameter. For the backfitting procedure, let $\widehat{\theta}(x,\beta)$ be defined by a value of θ that maximizes

$$n^{-1}\sum_{i=1}^{n} K_h(X_i - x)\mathcal{L}(Y_i, \beta, \theta) \tag{6}$$

for fixed values of β and x. In order to focus on the primary issues, we assume the existence of a well-defined maximizer of (6).

For the profiling estimator, all we need is that $\widehat{\theta}(x,\beta)$ and $\widehat{\theta}_\beta(x,\beta)$ satisfy assumption (PR1) given in Appendix B. It implies that the asymptotic distribution of $\widehat{\beta}$ does not depend on the estimators of θ_0 and $\theta_{0\beta}$ (defined by $\theta_{0\beta}(x,\beta) = \frac{\partial}{\partial\beta}\theta_0(x,\beta)$). While other nonparametric estimators, for example those based on splines or local polynomials, can be used, $\widehat{\theta}$ and $\widehat{\theta}_\beta$ can be estimated in the following way: let $\widehat{\theta}(x,\beta)$ be defined as for the backfitting procedure and let $\widehat{\theta}_\beta(x,\beta)$ be the partial derivative of $\widehat{\theta}(x,\beta)$ with respect to β, or, in case $\widehat{\theta}(x,\beta)$ is not differentiable with respect to β, define $\widehat{\theta}_\beta(x,\beta)$ by

$$\frac{\partial}{\partial\beta}\int \widehat{\theta}(x,b)L_g(\beta - b)\,db,$$

where L is a kernel density function and g is an appropriate bandwidth. Recall that $\widehat{\beta}_{BF}$ and $\widehat{\beta}_{PR}$ are the estimators of β_0 defined in (4) and (5).

Let $\theta_0(x,\beta)$ denote a solution of $E\{\mathcal{L}_\theta(Y,\beta,\theta)|X = x\} = 0$ with respect to θ for fixed β and x, where the expectation is calculated under the distribution induced by $\{\beta_0, \theta_0(\cdot)\}$. We assume that $\theta_0(x,\beta)$ is unique. Clearly, $\theta_0(\cdot,\beta_0) \equiv \theta_0(\cdot)$.

For any function $H \equiv \{H_1(\beta,\theta), \ldots, H_d(\beta,\theta)\}$ of (say) dimension d, we use the notation $\frac{\partial}{\partial\beta}H(\beta,\theta)$ and $\frac{d}{d\beta}H(\beta,\theta)$ to denote the $d \times q$ matrix with (i,j)th-element ($i = 1,\ldots,d, j = 1,\ldots,q$) given by

$$\frac{\partial}{\partial\beta}H(\beta,\theta)_{ij} = \lim_{\tau\to 0}\frac{1}{\tau}[H_i\{\beta + \tau e_j, \theta(\cdot,\beta)\} - H_i\{\beta, \theta(\cdot,\beta)\}] \tag{7}$$

and

$$\frac{d}{d\beta}H(\beta,\theta)_{ij} = \lim_{\tau\to 0}\frac{1}{\tau}[H_i\{\beta+\tau e_j, \theta(\cdot,\beta+\tau e_j)\} - H_i\{\beta,\theta(\cdot,\beta)\}] \tag{8}$$

respectively, where $e_j = (e_{j1}, \ldots, e_{jq})$ and $e_{jk} = \delta_{jk} = I(j = k)$ $(k = 1, \ldots, q)$.

We are now ready to state the main result concerning the asymptotic normality of $\widehat{\beta}_{BF}$ and $\widehat{\beta}_{PR}$. Let $\mathcal{G}(\beta) = \frac{d}{d\beta}E[\mathcal{L}_\beta\{Y, \beta, \theta_0(X, \beta)\}]$ and define

$$\Sigma = \text{cov}\Big[\mathcal{L}_\beta\{Y, \beta_0, \theta_0(X, \beta_0)\} + \mathcal{L}_\theta\{Y, \beta_0, \theta_0(X, \beta_0)\}\frac{\partial}{\partial\beta}\theta_0(X, \beta_0)\Big].$$

Theorem 2.1. *Assume* (BF1)–(BF8) *and, in particular, assume that the bandwidth h satisfies $nh^4 \to 0$ and not $h \propto n^{-1/5}$. Then* $n^{1/2}(\widehat{\beta}_{BF} - \beta_0) \xrightarrow{d}$ Normal $\big\{0, \mathcal{G}^{-1}(\beta_0)\Sigma\mathcal{G}^{-1}(\beta_0)^T\big\}$.

Theorem 2.2. *Assume* (PR1)–(PR4) *and, in particular, allow that the bandwidth h satisfies $h \propto n^{-1/5}$. Then* $n^{1/2}(\widehat{\beta}_{PR} - \beta_0) \xrightarrow{d}$ Normal$\{0, \mathcal{G}^{-1}(\beta_0)\Sigma\mathcal{G}^{-1}(\beta_0)^T\}$.

The proof of these results, as well the assumptions under which they are valid, can be found in Appendix A for the backfitting method, and in Appendix B for the profiling method.

As a consequence, the backfitting and profiling method produce estimators with the same asymptotic distribution even when m_{BF} or m_{PR} are not smooth in β or θ. The backfitting procedure requires however that undersmoothing be used to estimate $\widehat{\theta}(x, \beta)$, whereas the profiling procedure does not.

The assumptions stated in the appendices reveal some of the difficulties related to working with non-smooth criterion functions. For instance, in the statement of conditions (BF6) and (PR2), the expected values are necessary to transform a non-smooth function into a smooth one. If the criterion function were smooth, these expected values would not be needed and the proofs would become considerably easier in that case.

Note that, although the asymptotic variance $\mathcal{G}^{-1}(\beta_0)\Sigma\mathcal{G}^{-1}(\beta_0)^T$ has an explicit formula, its actual computation might be complicated in certain situations. In such cases, a bootstrap approximation can be useful. See Theorem B in Chen, Linton and Van Keilegom (2003) for general conditions under which a naive bootstrap procedure is valid.

3. Applications

In this section we consider two examples in detail : a partially linear median regression model, and a varying coefficient change point model. Both are examples of models with non-smooth criterion functions. The two models can be considered as generic examples, many other examples can be developed along similar lines.

3.1. Partially linear median regression

Consider the model

$$\mathcal{Y}_i = Z_i^{\mathrm{T}}\beta_0 + \theta_0(X_i) + \epsilon_i, \tag{9}$$

where Z_i is a possibly vector-valued covariate of dimension q, X_i is a scalar covariate with compact support R_X, whose density is positive on that support, and $\mathrm{med}(\epsilon_i|X_i, Z_i) = 0$. In our general notation, let $Y = (\mathcal{Y}, Z)$. The maximizing function is

$$\mathcal{L}\{Y, \beta, \theta(X)\} = -|\mathcal{Y} - Z^{\mathrm{T}}\beta - \theta(X)|. \tag{10}$$

The backfitting estimator $\widehat{\beta}_{BF}$ of β_0 has been considered in Chen, Linton and Van Keilegom (2003), see their Example 2. Here, we consider the profiling estimator.

It is readily seen that for fixed β, $\theta_0(x, \beta) = \mathrm{med}(\mathcal{Y} - Z^{\mathrm{T}}\beta|X = x)$. Let $\widehat{\theta}(x, \beta)$ be the kernel estimator of the conditional median of $\mathcal{Y} - Z^T\beta$ given $X = x$. Note that $\widehat{\theta}(x, \beta)$ is not smooth in β, because $\widehat{\theta}(x, \beta)$ is piecewise constant as a function of β. Hence we define $\widehat{\theta}_\beta(x, \beta)$ by

$$\widehat{\theta}_\beta(x, \beta) = \frac{\partial}{\partial\beta}\int \widehat{\theta}(x, b)L_g(\beta - b)\, db.$$

Let $\Theta = \{\theta : \theta(\cdot, \beta) \in C_M^\alpha(R_X) \text{ for all } \beta\}$ for some $\alpha > 1$ and $0 < M < \infty$, see Appendix A for the definition of the class $C_M^\alpha(R_X)$. Assume that θ_0 and the components of $\theta_{0\beta}$ belong to Θ. Then, using kernel theory for median regression (see e.g., Chaudhuri (1991)), it can be seen that assumption (PR1) is valid : for $\widehat{\theta}_\beta - \theta_{0\beta}$ note that

$$\begin{aligned}\|\widehat{\theta}_\beta - \theta_{0\beta}\|_\infty \le &\Big\|\frac{\partial}{\partial\beta}\int \{\widehat{\theta}(x, b) - \theta_0(x, b)\}L_g(\beta - b)\, db\Big\|_\infty \\ &+ \Big\|\frac{\partial}{\partial\beta}\int \{\theta_0(x, b) - \theta_0(x, \beta)\}L_g(\beta - b)\, db\Big\|_\infty\end{aligned}$$

and this is $o_P(n^{-1/4})$ provided L is a symmetric, compactly supported kernel function, $ng^8 \to 0$, $nh^2g^4 \to \infty$, $nh^8g^{-4} \to 0$, $\|\widehat{\theta} - \theta_0\|_\infty = O_P\{(nh)^{-1/2} + h^2\}$, and θ_0 is three times continuously differentiable with respect to the components of β. For example, take $h = C_1n^{-1/5}$ and $g = C_2n^{-1/7}$ for some $C_1, C_2 > 0$. In order to show that $\|\widehat{\theta} - \theta_0\|_\infty = O_P\{(nh)^{-1/2} + h^2\}$, note that Chaudhuri (1991) shows that $\sup_x |\widehat{\theta}(x, \beta) - \theta_0(x)| = O_P\{(nh)^{-1/2} + h^2\}$ for fixed β and for $h \propto n^{-1/5}$. It is possible to extend this result to bandwidths h that satisfy the above constraints, and to prove that the given rate holds uniformly over all β. It suffices for this to replace the supremum over β, in an appropriate way, by a maximum over a set of grid points (of size tending to infinity) and to prove the consistency uniformly over the set of grid points.

Direct calculations show that

$$E[\mathcal{L}_\beta\{Y,\beta,\theta(X,\beta)\}|X] = -E([2F_{\mathcal{Y}|X,Z}\{Z^{\mathrm{T}}\beta + \theta(X,\beta)\} - 1]Z|X),$$
$$E[\mathcal{L}_\theta\{Y,\beta,\theta(X,\beta)\}|X] = -E[2F_{\mathcal{Y}|X,Z}\{Z^{\mathrm{T}}\beta + \theta(X,\beta)\} - 1|X].$$

In addition,

$$\mathcal{G}(\beta_0) = -2E\Big[f_{\mathcal{Y}|X,Z}\{Z^{\mathrm{T}}\beta_0 + \theta_0(X)\}Z\Big\{Z + \frac{\partial}{\partial\beta}\theta_0(X,\beta_0)\Big\}^{\mathrm{T}}\Big].$$

Hence, assumption (PR2) is verified under standard smoothness conditions on $F_{\mathcal{Y}|X,Z}$. Also, (PR4) holds under classical identifiability conditions. It is readily seen that (PR3)(i) is valid for $r_\ell = 2$ and $s_\ell = 1/2$. Finally, for assumption (PR3)(ii) we make use of Theorem 2.7.1 in van der Vaart and Wellner (1996). It is easily checked that

$$\int_0^\infty \sqrt{\log N(\varepsilon^2, \tilde{\Theta}, \|\cdot\|_\infty)}\, d\varepsilon \le C \int_0^{(2M)^{\frac{1}{2}}} \varepsilon^{-\frac{1}{\alpha}} d\varepsilon < \infty,$$

for some $C > 0$. The asymptotic normality of $\widehat{\beta}_{PR}$ now follows. Note that

$$\Sigma = \mathrm{cov}\Big(\{2I(\epsilon \ge 0) - 1\}\Big[Z - \frac{E\{f_{\epsilon|X,Z}(0)Z|X\}}{f_{\epsilon|X}(0)}\Big]\Big),$$

since it is easily seen that $\theta_{0\beta}(X,\beta_0) = -E\{f_{\epsilon|X,Z}(0)Z|X\}/f_{\epsilon|X}(0)$.

3.2. Varying coefficient change point model

Consider the model

$$\mathcal{Y}_i = \theta_{01}(X_i) + \theta_{02}(X_i)|Z_i - \beta_0|_+ + \epsilon_i, \tag{11}$$

where $E(\epsilon_i|X_i, Z_i) = 0$, X_i and Z_i are scalar covariates and $z_+ = zI(z > 0)$. An interesting application can be found in toxicology, where models of the form

$$E(\mathcal{Y}_i|Z_i) = \theta_{01} + \theta_{02}|Z_i - \beta_0|_+^{\lambda_0} \tag{12}$$

are compatible with accepted understanding of the basic structure of dose-response curves for exposure to dioxin. Roberts (1991) states that “new findings suggest that responses to dioxin increase slowly at first but then shoot up after passing a critical concentration”. Indeed, researchers have “agreed that before dioxin can cause any of its myriad toxic effects ... it must first bind to and then activate a receptor. ... If receptor binding is indeed the essential first step

... then that implies there is a safe dose or practical threshold below which no toxic effects occur". Feder (1975) constructs $\sqrt{n}$-consistent and asymptotically normally distributed estimators for $(\theta_{01}, \theta_{02}, \beta_0)$ in (12) when $\lambda_0 = 1$, see his Example 1 on p. 77, setting his $\theta_{12} = 0$. Model (11) goes one step further, in the sense that it allows the average response before and the slope after critical concentration to depend on e.g., age or any other individual characteristic.

Let $Y = (\mathcal{Y}, Z)$ and define the maximizing function

$$\mathcal{L}\{Y, \beta, \theta(X)\} = -\{\mathcal{Y} - \theta_1(X) - \theta_2(X)|Z - \beta|_+\}^2.$$

Note that the theory developed in Section 2 can be extended in an obvious way to bivariate nuisance functions. Straightforward calculations show that for fixed β,

$$\theta_{02}(X, \beta) = \theta_{02}(X)\frac{\mathrm{Cov}\,(|Z - \beta_0|_+, |Z - \beta|_+|X)}{\mathrm{Var}\,(|Z - \beta|_+|X)} = \frac{\mathrm{Cov}\,(\mathcal{Y}, |Z - \beta|_+|X)}{\mathrm{Var}\,(|Z - \beta|_+|X)},$$
$$\theta_{01}(X, \beta) = E(\mathcal{Y}|X) - \theta_{02}(X, \beta)E(|Z - \beta|_+|X).$$

Also, let $\widehat{\theta}_1(X, \beta)$ and $\widehat{\theta}_2(X, \beta)$ be the estimators obtained by replacing the conditional means, variances and covariances in the above expressions by the corresponding kernel estimators, and let $\widehat{\theta}_{1\beta}(X, \beta)$ and $\widehat{\theta}_{2\beta}(X, \beta)$ be obtained by replacing $|Z_i - \beta|_+$ in these kernel estimators by $-I(Z_i \geq \beta)$ $(i = 1, \ldots, n)$.

As for the example on partially linear median regression, the main assumptions to verify are (BF7) for the backfitting procedure and (PR1) and (PR3) for the profiling method. We start with (PR1). Assume (BF1)-(BF4) hold, and let $\Theta = \{\theta : \theta(\cdot, \beta) \in C_M^\alpha(R_X)$ for all $\beta\}$ for some $\alpha > 1/2$, $0 < M < \infty$, and some compact interval R_X, see Appendix A for the definition of $C_M^\alpha(R_X)$. Assume that θ_0 and $\theta_{0\beta}$ belong to Θ. We show that $\|\widehat{\theta}_{j\beta} - \theta_{0j\beta}\|_\infty = o_P(n^{-1/4})$ $(j = 1, 2)$, the other conditions in (PR1) can be proved similarly. Since $\theta_{0j}(\cdot, \beta)$ is composed of variances, covariances and means, it suffices to consider each of these factors separately. For simplicity, we restrict attention to the mean, i.e., we consider

$$\sup_{x,\beta}\left|n^{-1}\sum_{i=1}^n \frac{K_h(X_i - x)}{\sum_{j=1}^n K_h(X_j - x)} I(Z_i \geq \beta) - P(Z \geq \beta|X = x)\right| = o_P(n^{-\frac{1}{4}})$$

provided $nh^2 \to \infty$, $nh^8 \to 0$, and $P(Z \leq \cdot|X = x)$ is twice continuously differentiable with respect to x, see e.g., Proposition 4.1 in Akritas and Van Keilegom (2001).

Part (i) of conditions (BF7) and (PR3) can be easily seen to hold true for $s_\ell = 1$ and $r_\ell = 2$. For part (ii), since $N(\varepsilon, C_M^\alpha(R_X), \|\cdot\|_\infty) = O\{\exp(K\varepsilon^{-1/\alpha})\}$, (see Thm. 2.7.1 in van der Vaart and Wellner (1996)), it follows that the integral in part (ii) is finite. The asymptotic normality of both $\widehat{\beta}_{BF}$ and $\widehat{\beta}_{PR}$ now follows.

The calculation of the asymptotic variance is straightforward but leads to lengthy formulas, and is left to the reader.

Acknowledgement

The first author gratefully acknowledges financial support from the IAP research network nr. P5/24 of the Belgian government (Belgian Science Policy). The research of the second author was supported by a grant from the National Cancer Institute (CA–57030), and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30–ES09106).

Appendix A. Proofs for Backfitting

Make the definitions

$$M_{nBF}(\beta,\theta) = n^{-1}\sum_{i=1}^{n} m_{BF}\{Y_i,\beta,\theta(X_i,\beta)\},$$
$$M_{BF}(\beta,\theta) = E[m_{BF}\{Y,\beta,\theta(X,\beta)\}],$$

and define the $q\times q$ matrix $\Gamma_{BF,\beta}(\beta,\theta) = \frac{d}{d\beta}M_{BF}(\beta,\theta) = \frac{d}{d\beta}E[\mathcal{L}_\beta\{Y,\beta,\theta(X,\beta)\}]$. Also, for a function $\xi(\cdot)=\xi(X,\beta)$, let $\Gamma_{BF,\theta}(\beta,\theta)[\xi]$ denote the Gâteaux-derivative of $M_{BF}(\beta,\theta)$ in the direction ξ, i.e.,

$$\begin{aligned}\Gamma_{BF,\theta}(\beta,\theta)[\xi] &= \lim_{\tau\to 0}\frac{1}{\tau}\{M_{BF}(\beta,\theta+\tau\xi) - M_{BF}(\beta,\theta)\}\\ &= \lim_{\tau\to 0}\frac{1}{\tau}E[\mathcal{L}_\beta\{Y,\beta,(\theta+\tau\xi)(X,\beta)\} - \mathcal{L}_\beta\{Y,\beta,\theta(X,\beta)\}]\\ &= E\left(\frac{\partial}{\partial\theta}E[\mathcal{L}_\beta\{Y,\beta,\theta(X,\beta)\}|X]\xi(X,\beta)\right),\end{aligned}$$

where $\frac{\partial}{\partial\theta}E[\mathcal{L}_\beta\{Y,\beta,\theta(X,\beta)\}|X] = \frac{\partial}{\partial z}E\{\mathcal{L}_\beta(Y,\beta,z)|X\}|_{z=\theta(X,\beta)}$. Note that M_{BF} $(\beta_0,\theta_0) = 0$.

For any function $g = (g_1,\ldots,g_d)$ of (say) dimension d defined on a set $\mathcal{A}$ in $I\!R^a$, for any $y\in\mathcal{A}$ and any k, let $\frac{\partial^k}{\partial y^k}g(y)$ denote the vector of all partial derivatives of order k of the form $\frac{\partial^k}{\partial y_1^{k_1}\ldots\partial y_a^{k_a}}g_j(y)$, where $\sum_{i=1}^a k_i = k$ and $1\le j\le d$. Let $\|g\|_\infty = \max_{1\le j\le d}\sup_{y\in\mathcal{A}}|g_j(y)|$. In particular, for a function $\theta=\theta(x,\beta)$, $\|\theta\|_\infty = \sup_{x,\beta}|\theta(x,\beta)|$ and $\|\frac{\partial\theta}{\partial\beta}\|_\infty = \max_{1\le\ell\le q}\sup_{x,\beta}|\frac{\partial\theta}{\partial\beta_\ell}(x,\beta)|$.

Further, let Θ be some space of functions $\theta=\theta(x,\beta)$ ($x\in I\!R,\beta\in\mathcal{B}$) for which $\|\theta\|_\infty\le M$ for some $M>0$.

The conditions below use the concept of covering number which is defined as follows. For $\epsilon>0$ and any normed space $(\Theta,\|\cdot\|)$ of functions, the covering number $N(\epsilon, \Theta, \|\cdot\|)$ is the minimal number of balls $\{\eta : \|\eta - \theta\| < \epsilon\}$ of radius ϵ needed to cover Θ. The centers of the balls need not belong to Θ, but they should have finite norms.

(BF1) The bandwidth h satisfies $nh^4 \to 0$ as n tends to infinity.

(BF2) The probability density function K has compact support and $\int uK(u)du = 0$.

(BF3) X is absolutely continuous and has compact support R_X, its density f_X is twice continuously differentiable and $\inf_x f_X(x) > 0$.

(BF4) $\theta_0 \in \Theta$, $\frac{\partial^{k+\ell}}{\partial x^k \partial \beta^\ell}\theta_0(x,\beta)$ $(0 \le k + l \le 3)$ exists for almost all x and β, and $\|\frac{\partial^{k+\ell}}{\partial x^k \partial \beta^\ell}\theta_0\|_\infty < \infty$.

(BF5) (i) $P(\widehat{\theta} \in \Theta) \to 1$ as $n \to \infty$ and $\|\widehat{\theta} - \theta_0\|_\infty = o_P(n^{-1/4})$.

(ii) $\sup_x |(\widehat{\theta} - \theta_0)(x, \widehat{\beta}) - (\widehat{\theta} - \theta_0)(x, \beta_0)| = o_P(1)\|\widehat{\beta} - \beta_0\|$.

(iii) $\sup_x |n^{-1}\sum_{i=1}^n K_h(X_i - x)\mathcal{L}_\theta\{Y_i, \beta_0, \widehat{\theta}(x, \beta_0)\}| = o_P(n^{-1/2})$.

(BF6) (i) For all y, $\mathcal{L}(y, \beta, \theta)$ is differentiable with respect to β and θ, for almost all β and θ.

(ii) $\frac{\partial}{\partial\theta}E[\mathcal{L}_\beta\{Y, \beta, \theta_0(X,\beta)\}|X]$ and $\frac{\partial}{\partial\beta}E[\mathcal{L}_\theta\{Y, \beta, \theta_0(X,\beta)\}|X]$ exist for all $\beta \in \mathcal{B}$, and they are equal.

(iii) $E\Big\{\sup_{|\theta|\le M} |\mathcal{L}_\theta(Y, \beta_0, \theta)|^2\Big\} < \infty$.

(iv) $\frac{\partial^{j+k+\ell}}{\partial\theta^j \partial x^k \partial\beta^\ell}E\{\mathcal{L}_\beta(Y, \beta, \theta)|X = x\}$ and $\frac{\partial^{j+k+\ell}}{\partial\theta^j \partial x^k \partial\beta^\ell}E\{\mathcal{L}_\theta(Y, \beta, \theta)|X = x\}$ exist for $0 \le j + k + \ell \le 2$ and for all β, θ and x, and

$$\sup_{\beta\in\mathcal{B}, |\theta|\le M, x\in R_X} \Big|\frac{\partial^{j+k+\ell}}{\partial\theta^j \partial x^k \partial\beta^\ell}E\{\mathcal{L}_\beta(Y, \beta, \theta)|X = x\}\Big| < \infty,$$

$$\sup_{\beta\in\mathcal{B}, |\theta|\le M, x\in R_X} \Big|\frac{\partial^{j+k+\ell}}{\partial\theta^j \partial x^k \partial\beta^\ell}E\{\mathcal{L}_\theta(Y, \beta, \theta)|X = x\}\Big| < \infty.$$

(v) $\mathcal{G}(\beta)$ exists for β in a neighborhood of β_0, is continuous at β_0 and $\mathcal{G}(\beta_0)$ is of full rank.

(BF7) (i)

$$E\left\{\sup_{(\beta',\theta'):\|\beta-\beta'\|\le\delta, \|\theta-\theta'\|_\infty\le\delta} |\mathcal{L}_\theta(Y, \beta, \theta) - \mathcal{L}_\theta(Y, \beta', \theta')|^{r_0}\right\} \le K_0 \delta^{r_0 s_0},$$

$$E\left\{\sup_{(\beta',\theta'):\|\beta-\beta'\|\le\delta, \|\theta-\theta'\|_\infty\le\delta} |\mathcal{L}_{\beta,\ell}(Y, \beta, \theta) - \mathcal{L}_{\beta,\ell}(Y, \beta', \theta')|^{r_\ell}\right\} \le K_\ell \delta^{r_\ell s_\ell},$$

for $r_0 = 2, 2+\eta$, for some $r_\ell \ge 2$ $(\ell = 1, \ldots, q)$, for all $(\beta, \theta) \in \mathcal{B} \times \Theta$, all $\delta > 0$, for some $\eta > 0$, some $0 < s_\ell \le 1$ and some $K_\ell > 0$ $(\ell = 0, \ldots, q)$.

(ii) $\int_0^\infty \sqrt{\log N(\varepsilon^{1/s_\ell}, \widetilde{\Theta}, \|\cdot\|_\infty)}\, d\varepsilon < \infty$, for $\ell = 0, \ldots, q$, where $\widetilde{\Theta} = \{\theta(\cdot, \beta) : \theta \in \Theta, \beta \in \mathcal{B}\}$.

(BF8) (i) For all $\delta > 0$, there exists a $\varepsilon > 0$ such that $\inf_{\|\beta-\beta_0\|>\delta} \|M_{BF}(\beta, \theta_0)\| \geq \varepsilon$.

(ii) Uniformly for all $\beta \in \mathcal{B}$, $M_{BF}(\beta, \theta)$ is continuous in θ at θ_0 (with respect to the $\|\cdot\|_\infty$ norm).

(iii) $\Gamma_{BF,\theta}(\beta, \theta_0)[\theta - \theta_0]$ exists in all directions $\theta - \theta_0 \in \Theta$.

Assumption (BF1) requires that undersmoothing is used to estimate the nuisance function θ_0. This is not required for the profiling method. Conditions (BF2)-(BF4) are standard regularity conditions that are common when using kernel smoothing methods. The properties required in (BF5) are a rate of convergence of $\widehat{\theta}$, a modulus of continuity-type result for $\widehat{\theta}$, and the sample analogue of the equation $E[\mathcal{L}_\theta\{Y, \beta_0, \theta_0(x, \beta_0)\}|X = x] = 0$. In the classical situation where $\|\widehat{\theta} - \theta_0\|_\infty = O_P\{(nh)^{-1/2}(\log n)^{1/2}\}$, condition (BF5)(i) requires that $nh^2(\log n)^{-2} \to \infty$. Note that assumption (BF6) does not impose smoothness conditions on $\mathcal{L}_\beta$ and $\mathcal{L}_\theta$, but instead requires that $E(\mathcal{L}_\beta)$ and $E(\mathcal{L}_\theta)$ are differentiable. Also note that in (BF6)(i), we allow for functions $\mathcal{L}$ that are smooth, except at a finite number of values, as in the example of a partially linear median regression model. The condition on the covering number in (BF7) can be checked by using e.g., the results obtained by van der Vaart and Wellner (1996). A common special case is 'one' in which the class $\tilde{\Theta}$ belongs to $C_M^\alpha(R_X)$, defined by the set of all continuous functions $\theta : R_X \to I\!R$ with $\|\theta\|_\alpha \leq M$, where

$$\|\theta\|_\alpha = \max_{k \leq \underline{\alpha}} \sup_x |\theta^{(k)}(x)| + \sup_{x_1, x_2} \frac{|\theta^{(\underline{\alpha})}(x_1) - \theta^{(\underline{\alpha})}(x_2)|}{|x_1 - x_2|^{\alpha - \underline{\alpha}}},$$

and $\underline{\alpha}$ is the largest integer strictly smaller than α. Theorem 2.7.1 (page 155 in their book) gives a bound on the covering number for this space. For (BF7)(ii) to be valid, it is clear that the space Θ should not be too large. On the other hand, condition (BF5)(i) stipulates that at the same time Θ should not be too small. Finally, (BF8) is a common condition in the context of estimating equations.

For the proofs below, we restrict attention for simplicity to the case $q = 1$. The general case $q \geq 1$ can be obtained in a similar way, but requires more complex notation.

We start with a technical lemma.

Lemma A.1. *Assume* (BF1)−(BF8). *Then,*

$$n^{-1} \sum_{i=1}^n E_X \Big(\frac{K_h(X_i - X)}{f_X(X)} \frac{\partial}{\partial \beta} \theta_0(X, \beta_0) [\mathcal{L}_\theta\{Y_i, \beta_0, \theta_0(X_i)\} - \mathcal{L}_\theta\{Y_i, \beta_0, \widehat{\theta}(X, \beta_0)\}] \Big)$$

$$= E_{X_1,X_2,Y_1}\Big(\frac{K_h(X_1-X_2)}{f_X(X_2)}\frac{\partial}{\partial\beta}\theta_0(X_2,\beta_0)[\mathcal{L}_\theta\{Y_1,\beta_0,\theta_0(X_1)\} \\ -\mathcal{L}_\theta\{Y_1,\beta_0,\widehat{\theta}(X_2,\beta_0)\}]\Big) + o_P(n^{-1/2}),$$

where the expectations are taken conditionally on the data (X_i, Y_i) $(i=1,\ldots,n)$.

Proof. Throughout the proof, C denotes a generic constant, whose value may change from one line to another. The following abbreviated notations will be used : let $\mathcal{H}(Y,\theta) = \mathcal{L}_\theta(Y,\beta_0,\theta)$ and $g(X) = \frac{\partial}{\partial\beta}\theta_0(X,\beta_0)$. To prove this result we make use of modern empirical process theory see e.g., van der Vaart and Wellner (1996). Consider the process $\sum_{i=1}^n Z_{ni}(\theta)$, where

$$Z_{ni}(\theta) = n^{-\frac{1}{2}}\Big\{E_X\Big(\frac{K_h(X_i-X)}{f_X(X)}g(X)[\mathcal{H}\{Y_i,\theta_0(X_i)\} - \mathcal{H}\{Y_i,\theta(X)\}]\Big) \\ - E_{X_1,X_2,Y_1}\Big(\frac{K_h(X_1-X_2)}{f_X(X_2)}g(X_2)[\mathcal{H}\{Y_1,\theta_0(X_1)\} - \mathcal{H}\{Y_1,\theta(X_2)\}]\Big)\Big\},$$

where θ belongs to Θ. For simplicity we suppress the dependence of θ on β_0. Note that by assumption (BF5)(i), $P(\widehat{\theta} \in \Theta) \to 1$. In order to show the weak convergence of this process we verify the conditions of Theorem 2.11.9 in van der Vaart and Wellner (1996):

$$\sum_{i=1}^n E\Big[\sup_{\theta\in\Theta}|Z_{ni}(\theta)|I\Big\{\sup_{\theta\in\Theta}|Z_{ni}(\theta)| > \eta\Big\}\Big] \to 0 \qquad \text{for every } \eta > 0; \tag{13}$$

$$\int_0^{\delta_n}\sqrt{\log N_{[\,]}(\varepsilon,\Theta,L_2^n)}\,d\varepsilon \to 0 \qquad \text{for every } \delta_n \downarrow 0; \tag{14}$$

$$\sum_{i=1}^n Z_{ni}(\theta) \text{ converges marginally for every } \theta \in \Theta, \tag{15}$$

where $N_{[\,]}(\varepsilon,\Theta,L_2^n)$ is the bracketing number, defined as the minimal number of sets N_ε in a partition $\Theta = \cup_{j=1}^{N_\varepsilon}\Theta_{\varepsilon j}$, such that for every $j = 1,\ldots,N_\varepsilon$,

$$\sum_{i=1}^n E\Big\{\sup_{\theta_1,\theta_2\in\Theta_{\varepsilon j}}|Z_{ni}(\theta_1) - Z_{ni}(\theta_2)|^2\Big\} \le \varepsilon^2. \tag{16}$$

The conditions (13) and (14) imply the asymptotic tightness of the process and can be proved separately for the four terms in the definition of $\sum_{i=1}^n Z_{ni}$. We restrict ourselves to showing (13) and (14) for the second term:

$$\sum_{i=1}^n \widetilde{Z}_{ni}(\theta) = n^{-\frac{1}{2}}\sum_{i=1}^n E_X\Big[\frac{K_h(X_i-X)}{f_X(X)}g(X)\mathcal{H}\{Y_i,\theta(X)\}\Big].$$

We start with verifying (14). Fix $\varepsilon > 0$. From assumption (BF7)(ii) it follows that there exist functions $\theta_1, \ldots, \theta_{N_\varepsilon}$ in Θ such that $\int_0^{\delta_n} \sqrt{\log N_\varepsilon}\, d\varepsilon \to 0$, and such that the balls $\{\theta : \|\theta - \theta_j\|_\infty \leq \varepsilon^{1/s_0}\}$ $(j = 1, \ldots, N_\varepsilon)$ cover Θ. We show that for any $1 \leq j \leq N_\varepsilon$,

$$\sum_{i=1}^n E\Bigg\{ \sup_{\|\theta-\theta_j\|_\infty \leq \varepsilon^{\frac{1}{s_0}}} |\widetilde{Z}_{ni}(\theta) - \widetilde{Z}_{ni}(\theta_j)|^2 \Bigg\} \leq \varepsilon^2. \tag{17}$$

The left hand side of (17) equals

$$\begin{aligned}
&E\Big(\sup_{\|\theta-\theta_j\|_\infty \leq \varepsilon^{1/s_0}} |\int K_h(X_1 - x) g(x) [\mathcal{H}\{Y_1, \theta(x)\} - \mathcal{H}\{Y_1, \theta_j(x)\}] dx|^2 \Big) \\
&\quad = E\Big(\sup_{\|\theta-\theta_j\|_\infty \leq \varepsilon^{1/s_0}} |\int K(u) g(X_1 - hu) [\mathcal{H}\{Y_1, \theta(X_1 - hu)\} \\
&\qquad - \mathcal{H}\{Y_1, \theta_j(X_1 - hu)\}] du|^2 \Big) \\
&\quad \leq \sup_x |g(x)|^2 \int K(u) E\Big[\sup_{\|\theta-\theta_j\|_\infty \leq \varepsilon^{\frac{1}{s_0}}} |\mathcal{H}\{Y_1, \theta(X_1 - hu)\} \\
&\qquad - \mathcal{H}\{Y_1, \theta_j(X_1 - hu)\}|^2 \Big] du \\
&\quad \leq C \sup_x |g(x)|^2 \varepsilon^2,
\end{aligned}$$

where the last inequality follows from assumption (BF7)(i). This shows (17), up to a universal constant, and hence, (14) is satisfied for the class $\sum_{i=1}^n \widetilde{Z}_{ni}(\theta)$. We next verify (13). With Z_{ni} replaced by $\widetilde{Z}_{ni}$, the left hand side of (13) is bounded by

$$\begin{aligned}
&n^{\frac{1}{2}} \sup_x |g(x)| E\Big(\sup_\theta |\mathcal{L}_\theta(Y, \beta_0, \theta)| I\Big[\sup_\theta |\mathcal{L}_\theta(Y, \beta_0, \theta)| > \eta n^{\frac{1}{2}} \{\sup_x |g(x)|\}^{-1} \Big] \Big) \\
&\quad = o(1),
\end{aligned}$$

where we have used assumption (BF6)(iii). For the convergence of the marginals of $\sum_{i=1}^n Z_{ni}(\theta)$, we verify Liapunov's condition :

$$\frac{\sum_{i=1}^n E|Z_{ni}(\theta)|^{2+\eta}}{[\sum_{i=1}^n \mathrm{Var}\,\{Z_{ni}(\theta)\}]^{\frac{(2+\eta)}{2}}} \to 0$$

for some $\eta > 0$. First, consider the variance. Using a similar derivation as above, we obtain for any $\theta \in \Theta$,

$$\sum_{i=1}^n \mathrm{Var}\,\{Z_{ni}(\theta)\}$$

$$
\begin{aligned}
&\le \sup_x |g(x)|^2 \int K(u) E|\mathcal{H}\{Y_1, \theta_0(X_1)\} - \mathcal{H}\{Y_1, \theta(X_1 - hu)\}|^2 du \\
&\le C \sup_x |g(x)|^2 \int K(u) \sup_x |\theta_0(x) - \theta(x - hu)|^{2s_0} du \\
&\le Ch^{2s_0} \sup_x |g(x)|^2 \sup_x |\frac{\partial}{\partial x}\theta_0(x)|^{2s_0} \int K(u)|u|^{2s_0} du + 2CM \sup_x |g(x)|^2 \\
&= O(1). \qquad (18)
\end{aligned}
$$

In a similar way one can show that $\sum_{i=1}^n E|Z_{ni}(\theta)|^{2+\eta} = O(n^{-\eta/2})$, since assumption (BF7)(i) assures that $E|\mathcal{H}\{Y_1, \theta_0(X_1)\} - \mathcal{H}\{Y_1, \theta(X_1 - hu)\}|^{2+\eta} \le C \sup_x |\theta_0(x) - \theta(x - hu)|^{2s_0}$. Hence, the Liapunov ratio is $O\{n^{-\eta/2}\} = o(1)$. This shows the weak convergence of the process $\sum_{i=1}^n Z_{ni}(\theta)$ $(\theta \in \Theta)$. It now follows that $\sup_{\theta\in\Theta} |\sum_{i=1}^n Z_{ni}(\theta)| = O_P(1)$. Finally, arguments similar to those in (18) show that $\sum_{i=1}^n \mathrm{Var}\,\{Z_{ni}(\widehat{\theta})\} = o_P(1)$ (where the variance is calculated conditionally on the value of $\widehat{\theta}$), so that $\sum_{i=1}^n Z_{ni}(\widehat{\theta}) = o_P(1)$, from which the result follows.

Lemma A.2. *Assume* (BF1)−(BF8). *Then,*

$$
\Gamma_{BF,\theta}(\beta_0, \theta_0)[\widehat{\theta} - \theta_0] = n^{-1} \sum_{i=1}^n \mathcal{L}_\theta\{Y_i, \beta_0, \theta_0(X_i, \beta_0)\} \frac{\partial}{\partial\beta}\theta_0(X_i, \beta_0) + o_P(n^{-\frac{1}{2}}). \qquad (19)
$$

Proof. Recall the definitions of $\frac{\partial}{\partial\beta}$ and $\frac{d}{d\beta}$ given in (7) and (8). First note that

$$
\begin{aligned}
&\Gamma_{BF,\theta}(\beta_0, \theta_0)[\widehat{\theta} - \theta_0] \\
&\quad = \lim_{\tau\to 0} \frac{1}{\tau} E(\mathcal{L}_\beta[Y, \beta_0, \{\theta_0 + \tau(\widehat{\theta} - \theta_0)\}(X, \beta_0)] - \mathcal{L}_\beta\{Y, \beta_0, \theta_0(X, \beta_0)\}) \\
&\quad = E\Big(\frac{\partial}{\partial\theta} E[\mathcal{L}_\beta\{Y, \beta_0, \theta_0(X, \beta_0)\}|X](\widehat{\theta} - \theta_0)(X, \beta_0)\Big) \\
&\quad = E\Big(\frac{\partial}{\partial\beta} E[\mathcal{L}_\theta\{Y, \beta_0, \theta_0(X, \beta_0)\}|X](\widehat{\theta} - \theta_0)(X, \beta_0)\Big) \\
&\quad = -E\Big(\frac{\partial}{\partial\theta} E[\mathcal{L}_\theta\{Y, \beta_0, \theta_0(X, \beta_0)\}|X](\widehat{\theta} - \theta_0)(X, \beta_0)\frac{\partial}{\partial\beta}\theta_0(X, \beta_0)\Big), \qquad (20)
\end{aligned}
$$

since $E[\mathcal{L}_\theta\{Y, \beta, \theta_0(X, \beta)\}|X] = 0$ for all β. Next, let $g(X) = \frac{\partial}{\partial\beta}\theta_0(X, \beta_0)$ and $\mathcal{H}(Y, \theta) = \mathcal{L}_\theta(Y, \beta_0, \theta)$. The right hand side of (19) equals

$$
\begin{aligned}
&n^{-1} \sum_{i=1}^n E_X \left[\frac{K_h(X_i - X)}{f_X(X)} g(X)\right] \mathcal{H}\{Y_i, \theta_0(X_i, \beta_0)\} + o_P(n^{-\frac{1}{2}}) \\
&\quad = n^{-1} \sum_{i=1}^n E_X \left(\frac{K_h(X_i - X)}{f_X(X)} g(X)[\mathcal{H}\{Y_i, \theta_0(X_i, \beta_0)\} - \mathcal{H}\{Y_i, \widehat{\theta}(X, \beta_0)\}]\right)
\end{aligned}
$$

$$+o_P(n^{-\frac{1}{2}}),$$

since $n^{-1}\sum_{i=1}^n K_h(X_i - x)\mathcal{H}\{Y_i, \widehat{\theta}(x,\beta_0)\} = o_P(n^{-1/2})$ uniformly in x, see assumption (BF5)(iii). Note that throughout this proof all expectations are conditional on the data (X_i, Y_i), which implies that $\widehat{\theta}$ is considered as constant.

Using Lemma A.1 the latter expression can be written as

$$\begin{aligned}
&E_{X_1,X_2,Y_1}\left(\frac{K_h(X_1 - X_2)}{f_X(X_2)} g(X_2)[\mathcal{H}\{Y_1, \theta_0(X_1,\beta_0)\} - \mathcal{H}\{Y_1, \widehat{\theta}(X_2,\beta_0)\}]\right)\\
&\quad + o_P(n^{-\frac{1}{2}})\\
&= E_{X_1,X_2}\left(\frac{K_h(X_1 - X_2)}{f_X(X_2)} g(X_2)[k\{X_1, \theta_0(X_1,\beta_0)\} - k\{X_1, \widehat{\theta}(X_2,\beta_0)\}]\right)\\
&\quad + o_P(n^{-\frac{1}{2}}),
\end{aligned}$$

where $k(X,\theta) = E[\mathcal{H}(Y,\theta)|X]$. Using a Taylor expansion of order two and assumptions (BF1), (BF3), (BF4) and (BF6)(iv) this can be written as

$$\begin{aligned}
&E_{X_2}\left(\frac{E_{X_1}\{K_h(X_1 - X_2)\}}{f_X(X_2)} g(X_2)[k\{X_2, \theta_0(X_2,\beta_0)\} - k\{X_2, \widehat{\theta}(X_2,\beta_0)\}]\right)\\
&\quad + E_{X_2}\Big(\frac{E_{X_1}\{(X_1 - X_2)K_h(X_1 - X_2)\}}{f_X(X_2)} g(X_2)\\
&\qquad\qquad \times \frac{d}{dx}[k\{x, \theta_0(x,\beta_0)\} - k\{x, \widehat{\theta}(X_2,\beta_0)\}]_{x=X_2}\Big) + o_P(n^{-\frac{1}{2}})\\
&= E(g(X)[k\{X, \theta_0(X,\beta_0)\} - k\{X, \widehat{\theta}(X,\beta_0)\}]) + o_P(n^{-\frac{1}{2}})\\
&= -E\Big[g(X)\frac{\partial}{\partial\theta}k\{X, \theta_0(X,\beta_0)\}\{\widehat{\theta}(X,\beta_0) - \theta_0(X,\beta_0)\}\Big] + o_P(n^{-\frac{1}{2}}),
\end{aligned}$$

since $\sup_x |\widehat{\theta}(x,\beta_0) - \theta_0(x,\beta_0)| = o_P(n^{-1/4})$. The latter expression equals $\Gamma_{BF,\theta}(\beta_0,\theta_0)[\widehat{\theta} - \theta_0] + o_P(n^{-1/2})$, by using (20). Hence, the result follows.

Proof of Theorem 2.1. We make use of Theorem 2 in Chen, Linton and Van Keilegom (2003) (CLV hereafter), which states primitive conditions under which $\widehat{\beta}_{BF}$ is asymptotically normal. First of all, we need to show that $\widehat{\beta}_{BF} - \beta_0 = o_P(1)$. For this, we verify the conditions of Theorem 1 in CLV. Condition (1.1) holds by definition of $\widehat{\beta}_{BF}$, while the second, third and fourth conditions are guaranteed by assumptions (BF5)(i) and (BF8). Finally, condition (1.5) is weaker than condition (2.5) of Theorem 2 of CLV, which we verify below. So, the conditions of Theorem 1 are verified, up to condition (1.5) which we postpone to later. Next, we verify conditions (2.1)–(2.6) of Theorem 2 in CLV. Condition (2.1) is, as for condition (1.1), valid by construction of the estimator $\widehat{\beta}_{BF}$, while condition (2.2) follows from assumption (BF6)(v). Since $\Gamma_{BF,\theta}(\beta,\theta_0)[\theta-\theta_0] = E\{\frac{\partial}{\partial\theta}d(X,\theta_0)(\theta-\theta_0)(X,\beta)\}$, where $d(X,\theta) = E[\mathcal{L}_\beta\{Y,\beta,\theta(X,\beta)\}|X]$, we have

$$\begin{aligned} &M_{BF}(\beta,\theta) - M_{BF}(\beta,\theta_0) - \Gamma_{BF,\theta}(\beta,\theta_0)[\theta-\theta_0] \\ &= E\Big\{d(X,\theta) - d(X,\theta_0) - \frac{\partial}{\partial\theta}d(X,\theta_0)(\theta-\theta_0)(X,\beta)\Big\} \\ &= \frac{1}{2}E\Big\{\frac{\partial^2}{\partial\theta^2}d(X,\xi)(\theta-\theta_0)^2(X,\beta)\Big\}, \end{aligned} \tag{21}$$

where $\xi(X)$ is in between $\theta(X,\beta)$ and $\theta_0(X,\beta)$. Hence the norm of (21) is bounded by a constant times $\|\theta-\theta_0\|_\infty^2$. This shows the first part of (2.3). For the second part, it follows from the proof of Theorem 2 in CLV that it suffices to show that

$$\|\Gamma_{BF,\theta}(\widehat{\beta},\theta_0)[\widehat{\theta}-\theta_0] - \Gamma_{BF,\theta}(\beta_0,\theta_0)[\widehat{\theta}-\theta_0]\| = o_P(1)\|\widehat{\beta}-\beta_0\|,$$

and this in turn follows from (BF4), (BF5)(ii) and (BF6)(iv). Next, (2.4) follows from assumption (BF5)(i), while (2.5) is guaranteed by Theorem 3 in CLV together with assumption (BF7). It remains to verify (2.6). Since $\Gamma_{BF,\theta}(\beta_0,\theta_0)[\widehat{\theta}-\theta_0]$ and $M_{nBF}(\beta_0,\theta_0)$ are sums of i.i.d. terms plus negligible terms of lower order (see Lemma A.2.), this follows immediately. The asymptotic normality of $\widehat{\beta}_{BF}$ now follows.

Appendix B. Proofs for Profiling

As for the backfitting estimator define, for any $\theta\in\Theta$ and $\eta\in\Theta^q$,

$$\begin{aligned} M_{nPR}(\beta,\theta,\eta) &= n^{-1}\sum_{i=1}^n m_{PR}\{Y_i,\beta,\theta(X_i,\beta),\eta(X_i,\beta)\}, \\ M_{PR}(\beta,\theta,\eta) &= E[m_{PR}\{Y,\beta,\theta(X,\beta),\eta(X,\beta)\}], \end{aligned}$$

and let $\Gamma_{PR,\beta}(\beta,\theta,\eta) = \frac{d}{d\beta}M_{PR}(\beta,\theta,\eta)$. Note that $M_{PR}(\beta_0,\theta_0,\theta_{0\beta}) = 0$ and that

$$\begin{aligned} \Gamma_{PR,\beta}(\beta,\theta_0,\eta) &= \frac{d}{d\beta}E[\mathcal{L}_\beta\{Y,\beta,\theta_0(X,\beta)\}] + \frac{d}{d\beta}E\Big[\mathcal{L}_\theta\{Y,\beta,\theta_0(X,\beta)\}\eta(X,\beta)\Big] \\ &= \frac{d}{d\beta}E[\mathcal{L}_\beta\{Y,\beta,\theta_0(X,\beta)\}], \end{aligned}$$

since $E[\mathcal{L}_\theta\{Y,\beta,\theta_0(X,\beta)\}|X] = 0$. For functions $\xi(\cdot)$ and $\zeta(\cdot)$, let

$$\Gamma_{PR,\theta,\eta}(\beta,\theta,\eta)[\xi,\zeta] = \lim_{\tau\to 0}\frac{1}{\tau}\{M_{PR}(\beta,\theta+\tau\xi,\eta+\tau\zeta) - M_{PR}(\beta,\theta,\eta)\}.$$

Recall that Θ is some space of functions $\theta = \theta(x, \beta)$ ($x \in \mathbb{R}, \beta \in \mathcal{B}$) for which $\|\theta\|_\infty \leq M$ for some $M > 0$. For any $r \geq 1$ and any $\theta_1, \ldots, \theta_r \in \Theta$, let $\|(\theta_1, \ldots, \theta_r)\|_\infty = \max_{1\leq j\leq r} \|\theta_j\|_\infty$.

The assumptions we need to impose for the main result, are the following:

(PR1) $\theta_0 \in \Theta$, θ_0 is partially differentiable with respect to the components of β, $\frac{\partial\theta_0}{\partial\beta} \in \Theta^q$, $P(\widehat{\theta} \in \Theta) \to 1$ and $P(\widehat{\theta}_\beta \in \Theta^q) \to 1$ as $n \to \infty$, $\|\widehat{\theta} - \theta_0\|_\infty = o_P(n^{-1/4})$, and $\|\widehat{\theta}_\beta - \theta_{0\beta}\|_\infty = o_P(n^{-1/4})$.

(PR2) (i) For all y, $\mathcal{L}(y, \beta, \theta)$ is differentiable with respect to β and θ, for almost all β and θ.

(ii) $\frac{\partial}{\partial\theta} E[\mathcal{L}_\beta\{Y, \beta, \theta_0(X, \beta)\}|X]$ and $\frac{\partial}{\partial\beta} E[\mathcal{L}_\theta\{Y, \beta, \theta_0(X, \beta)\}|X]$ exist for all $\beta \in \mathcal{B}$, and they are equal.

(iii) $\frac{\partial^2}{\partial\theta^2} E\{\mathcal{L}_\beta(Y, \beta, \theta)|X = x\}$ and $\frac{\partial^2}{\partial\theta^2} E\{\mathcal{L}_\theta(Y, \beta, \theta)|X = x\}$ exist for all β, θ and x, and

$$\sup_{\beta\in\mathcal{B}, |\theta|\leq M, x\in R_X} \left|\frac{\partial^2}{\partial\theta^2} E\{\mathcal{L}_\beta(Y, \beta, \theta)|X = x\}\right| < \infty,$$

$$\sup_{\beta\in\mathcal{B}, |\theta|\leq M, x\in R_X} \left|\frac{\partial^2}{\partial\theta^2} E\{\mathcal{L}_\theta(Y, \beta, \theta)|X = x\}\right| < \infty,$$

where R_X is the support of X.

(iv) $\mathcal{G}(\beta)$ exists for β in a neighborhood of β_0, is continuous at β_0 and $\mathcal{G}(\beta_0)$ is of full rank.

(PR3) (i)

$$E\left\{\sup_{(\beta',\theta'):\|\beta-\beta'\|\leq\delta, \|\theta-\theta'\|_\infty\leq\delta, \|\eta-\eta'\|_\infty\leq\delta} |m_{PR,\ell}(Y, \beta, \theta, \eta) - m_{PR,\ell}(Y, \beta', \theta', \eta')|^{r_\ell}\right\} \leq K_\ell \delta^{r_\ell s_\ell}$$

for some $r_\ell \geq 2$, for all $(\beta, \theta, \eta) \in \mathcal{B} \times \Theta^{q+1}$, all $\delta > 0$, for some $0 < s_\ell \leq 1$ and some $K_\ell > 0$ ($\ell = 1, \ldots, q$).

(ii) $\int_0^\infty \sqrt{\log N(\varepsilon^{1/s_\ell}, \widetilde{\Theta}, \|\cdot\|_\infty)}\, d\varepsilon < \infty$ for $\ell = 1, \ldots, q$, where $\widetilde{\Theta} = \{\theta(\cdot, \beta) : \theta \in \Theta, \beta \in \mathcal{B}\}$.

(PR4) (i) For all $\delta > 0$, there exists a $\varepsilon > 0$ such that $\inf_{\|\beta-\beta_0\|>\delta} \|M_{PR}(\beta, \theta_0, \theta_{0\beta})\| \geq \varepsilon$.

(ii) Uniformly for all $\beta \in \mathcal{B}$, $M_{PR}(\beta, \theta, \eta)$ is continuous in (θ, η) at $(\theta_0, \theta_{0\beta})$ (with respect to the $\|\cdot\|_\infty$ norm).

The discussion of these conditions is very similar to that for the backfitting method. We therefore refer to Appendix A for more details.

Lemma B.1. *Assume* (PR1)−(PR4). *Then, for any* $\xi \in \Theta$, $\zeta \in \Theta^q$ *and* $\beta \in \mathcal{B}$, $\Gamma_{PR,\theta,\eta}(\beta, \theta_0, \theta_{0\beta})[\xi, \zeta] = 0$.

Proof. Write

$$\begin{aligned}
&\Gamma_{PR,\theta,\eta}(\beta, \theta_0, \theta_{0\beta})[\xi, \zeta] \\
&= \lim_{\tau\to 0} \frac{1}{\tau} E[\mathcal{L}_\beta\{Y, \beta, (\theta_0 + \tau\xi)(X, \beta)\} - \mathcal{L}_\beta\{Y, \beta, \theta_0(X, \beta)\}] \\
&\quad + \lim_{\tau\to 0} \frac{1}{\tau} E\Big\{[\mathcal{L}_\theta\{Y, \beta, (\theta_0 + \tau\xi)(X, \beta)\} - \mathcal{L}_\theta\{Y, \beta, \theta_0(X, \beta)\}](\theta_{0\beta} + \tau\zeta)(X, \beta)\Big\} \\
&\quad + \lim_{\tau\to 0} \frac{1}{\tau} E\Big[\mathcal{L}_\theta\{Y, \beta, \theta_0(X, \beta)\}\tau\zeta(X, \beta)\Big]. \qquad (22)
\end{aligned}$$

The third term of (22) equals $E(E[\mathcal{L}_\theta\{Y, \beta, \theta_0(X, \beta)\}|X]\zeta(X, \beta)) = 0$, since $E[\mathcal{L}_\theta\{Y, \beta, \theta_0(X, \beta)\}|X] = 0$. The first term of (22) can be written as

$$E\left\{\left(\frac{\partial}{\partial\theta} E[\mathcal{L}_\beta\{Y, \beta, \theta_0(X, \beta)\}|X]\right)\xi(X, \beta)\right\},$$

while the second term equals

$$E\left\{\left(\frac{\partial}{\partial\theta} E[\mathcal{L}_\theta\{Y, \beta, \theta_0(X, \beta)\}|X]\right)\xi(X, \beta)\frac{\partial}{\partial\beta}\theta_0(X, \beta)\right\}. \qquad (23)$$

Since $E[\mathcal{L}_\theta\{Y, \beta, \theta_0(X, \beta)\}|X] = 0$ for all β, it follows that

$$\frac{\partial}{\partial\beta} E\left[\mathcal{L}_\theta\{Y, \beta, \theta_0(X, \beta)\}|X\right] + \frac{\partial}{\partial\theta} E[\mathcal{L}_\theta\{Y, \beta, \theta_0(X, \beta)\}|X]\frac{\partial}{\partial\beta}\theta_0(X, \beta) = 0,$$

and hence, plugging in this expression into (23) gives

$$-E\left\{\frac{\partial}{\partial\beta} E[\mathcal{L}_\theta\{Y, \beta, \theta_0(X, \beta)\}|X]\xi(X, \beta)\right\}.$$

Hence, $\Gamma_{PR,\theta,\eta}(\beta, \theta_0, \theta_{0\beta})[\xi, \zeta] = 0$, since $\frac{\partial}{\partial\beta}E(\mathcal{L}_\theta) = \frac{\partial}{\partial\theta}E(\mathcal{L}_\beta)$.

Proof of Theorem 2.2. In a manner similar to the backfitting procedure, we proceed by checking the primitive conditions of Theorem 2 in Chen, Linton and Van Keilegom (2003) (CLV hereafter). Note that the results in that paper are valid for one-dimensional nuisance functions θ, but it is readily seen how to extend their primitive conditions to the current setup of $(q + 1)$-dimensional nuisance functions.

The verification of the conditions in that theorem is much the same as for the backfitting procedure, except for conditions (2.3) and (2.5). Let us start by verifying (2.3). Since it follows from the proof of Lemma B.1 that $\Gamma_{PR,\theta,\eta}(\beta,\theta_0,\theta_{0\beta})\,[\theta-\theta_0,\eta-\theta_{0\beta}] = E\{\frac{\partial}{\partial\theta}d_1(X,\theta_0)(\theta-\theta_0)(X,\beta)\}+E\{\frac{\partial}{\partial\theta}d_2(X,\theta_0)$ $(\theta-\theta_0)(X,\beta)\frac{\partial}{\partial\beta}\theta_0(X,\beta)\}$, where $d_1(X,\theta) = E[\mathcal{L}_\beta\{Y,\beta,\theta(X,\beta)\}|X]$ and $d_2(X,\theta)$ $= E[\mathcal{L}_\theta\{Y,\beta,\theta(X,\beta)\}|X]$, we have

$$
\begin{aligned}
&M_{PR}(\beta,\theta,\eta) - M_{PR}(\beta,\theta_0,\theta_{0\beta}) - \Gamma_{PR,\theta,\eta}(\beta,\theta_0,\theta_{0\beta})[\theta-\theta_0,\eta-\theta_{0\beta}]\\
&\quad= E\Big\{d_1(X,\theta) - d_1(X,\theta_0) - \frac{\partial}{\partial\theta}d_1(X,\theta_0)(\theta-\theta_0)(X,\beta)\Big\}\\
&\qquad+E\Big[\{d_2(X,\theta) - d_2(X,\theta_0) - \frac{\partial}{\partial\theta}d_2(X,\theta_0)(\theta-\theta_0)(X,\beta)\}\eta(X,\beta)\Big]\\
&\qquad+E\Big\{d_2(X,\theta_0)(\eta-\theta_{0\beta})(X,\beta)\Big\}\\
&\qquad+E\Big\{\frac{\partial}{\partial\theta}d_2(X,\theta_0)(\theta-\theta_0)(X,\beta)(\eta-\theta_{0\beta})(X,\beta)\Big\}\\
&\quad= \frac{1}{2}E\Big\{\frac{\partial^2}{\partial\theta^2}d_1(X,\xi_1)(\theta-\theta_0)^2(X,\beta)\Big\}\\
&\qquad+\frac{1}{2}E\Big\{\frac{\partial^2}{\partial\theta^2}d_2(X,\xi_2)(\theta-\theta_0)^2(X,\beta)\frac{\partial}{\partial\beta}\theta_0(X,\beta)\Big\}\\
&\qquad+E\Big\{\frac{\partial}{\partial\theta}d_2(X,\theta_0)(\theta-\theta_0)(X,\beta)(\eta-\theta_{0\beta})(X,\beta)\Big\},
\end{aligned}
\tag{24}
$$

since $d_2(X,\theta_0)\equiv 0$, where $\xi_1(X)$ and $\xi_2(X)$ are in between $\theta(X,\beta)$ and $\theta_0(X,\beta)$. Hence the norm of (24) is bounded by a constant times $\|(\theta-\theta_0,\eta-\theta_{0\beta})\|_\infty^2$. This shows the first part of (2.3). The second part is obvious by Lemma B.1.

Finally, (2.5) is guaranteed by Theorem 3 in CLV together with assumption (PR3). Note that $N(\varepsilon^{1/s_\ell},\widetilde{\Theta}^q,\|\cdot\|_\infty) \le N(\varepsilon^{1/s_\ell},\widetilde{\Theta},\|\cdot\|_\infty)^q$, and hence the second condition in Theorem 3 in CLV is implied by (PR3)(ii). The result now follows.

References

Akritas, M. G. and Van Keilegom, I. (2001). Nonparametric estimation of the residual distribution. *Scand. J. Statist.* **28**, 549-568.

Buja, A., Hastie, T. J. and Tibshirani, R. J. (1989). Linear smoothers and additive models (with Discussion). *Ann. Statist.* **17**, 453-555.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477-489.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *Ann. Statist.* **19**, 760-777.

Chen, X., Linton, O. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71**, 1591-1608.

Feder, P. I. (1975). On asymptotic distribution theory in segmented regression problems - identified case. *Ann. Statist.* **3**, 49-83.

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

Hu, Z., Wang, N. and Carroll, R. J. (2004). Profile-kernel versus backfitting in the partially linear model for longitudinal/clustered data. *Biometrika* **91**, 251-262.

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.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95**, 449-485.

Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators. *J. Multivariate Anal.* **73**, 166-179.

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

Opsomer, J. D. and Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *J. Comput. Graph. Statist.* **8**, 715-732.

Rice, J. A. (1986). Convergence rates for partially splined models. *Statist. Probab. Lett.* **4**, 204-208.

Roberts, L. (1991). Dioxin risks revisited. *Science* **251**, 624-626.

Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.* **89**, 501-512.

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

Speckman, P. E. (1988). Regression analysis for partially linear models. *J. Roy. Statist. Soc. Ser. B* **50**, 413-436.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Wand, M. P. (1999). A central limit theorem for local polynomial backfitting estimators. *J. Multivariate Anal.* **70**, 57-65.

Institut de Statistique, Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium.

E-mail: vankeilegom@stat.ucl.ac.be

Department of Statistics, Texas A&M University, College Station TX 77843-3143, U.S.A.

E-mail: carroll@stat.tamu.edu

(Received February 2005; accepted September 2005)