Bernoulli **17**(2), 2011, 736–748
DOI: 10.3150/10-BEJ296

Semi-parametric regression: Efficiency gains from modeling the nonparametric part

KYUSANG YU[1], ENNO MAMMEN[2] and BYEONG U. PARK[3]

[1]*Konkuk University, Seoul, Korea. E-mail: kyusangu@konkuk.ac.kr*
[2]*University of Mannheim, Mannheimun, Germany. E-mail: emammen@rumms.uni-mannheim.de*
[3]*Seoul National University, Seoul, Korea. E-mail: bupark@stats.snu.ac.kr*

It is widely admitted that structured nonparametric modeling that circumvents the curse of dimensionality is important in nonparametric estimation. In this paper we show that the same holds for semi-parametric estimation. We argue that estimation of the parametric component of a semi-parametric model can be improved essentially when more structure is put into the nonparametric part of the model. We illustrate this for the partially linear model, and investigate efficiency gains when the nonparametric part of the model has an additive structure. We present the semi-parametric Fisher information bound for estimating the parametric part of the partially linear additive model and provide semi-parametric efficient estimators for which we use a smooth backfitting technique to deal with the additive nonparametric part. We also present the finite sample performances of the proposed estimators and analyze Boston housing data as an illustration.

Keywords: partially linear additive models; profile estimator; semi-parametric efficiency; smooth backfitting

1. Introduction

Structured nonparametric models such as additive models are known to circumvent the curse of dimensionality and allow reliable estimation when a full nonparametric model does not work. In the present paper we show that a similar assertion applies for semi-parametric models: structural modeling of the nonparametric part can lead to accurate estimation of the parametric part even in situations where otherwise only very poor, unreliable or unstable estimates would be available. We show this by comparing the partially linear and the partially linear additive model. In particular, we demonstrate that using an additive model for the nonparametric part in the partially linear model can lead to drastic gains of efficiency in the estimation of the parametric components. This holds if the dimension of the nonparametric covariates is high, or the parametric covariates can be approximated by non-additive transformations of the nonparametric covariates. In the extreme of the latter case, if the approximation is exact, then estimation of the parametric part in the partially linear model breaks down. If the approximation is very crude, one sees large efficiency gains by using additive models for the nonparametric part.

Suppose we observe the i.i.d. copies $(Y^1, \mathbf{X}^1, \mathbf{Z}^1), \ldots, (Y^n, \mathbf{X}^n, \mathbf{Z}^n)$ of a random vector $(Y, \mathbf{X}, \mathbf{Z})$, where $\mathbf{X} = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ and $\mathbf{Z} = (Z_1, \ldots, Z_d)^\top \in \mathbb{R}^d$. The partially linear model assumes

$$Y = m_0 + \mathbf{X}^\top \boldsymbol{\beta} + m(Z_1, \ldots, Z_d) + \epsilon, \tag{1}$$

1350-7265

where $\boldsymbol{\beta}$ is an unknown p-vector and m is an unknown d-variate function. The partially linear additive model puts an additive structure to the nonparametric function m:

$$Y = m_0 + \mathbf{X}^\top \boldsymbol{\beta} + m_1(Z_1) + \cdots + m_d(Z_d) + \epsilon. \tag{2}$$

These models exclude the interesting case where $\mathbf{X}$ or $\mathbf{Z}$ includes some endogeneous variables of Y, but they simplify our discussion on semi-parametric efficiency. We believe that our results can be extended to the corresponding semi-parametric models with time series data by following, for example, the arguments in [7].

For identifiability of the additive component functions m_j, we put the constraints $Em_j(Z_j) = 0$, $1 \le j \le d$. We assume that $(\mathbf{X}, \mathbf{Z})$ has a joint density q with respect to $\nu = \nu_1 \times \nu_2$, where ν_1 is a σ-finite measure and ν_2 is the Lebesgue measure on each support of $\mathbf{X}$ and $\mathbf{Z}$, and that the marginal density of $\mathbf{Z}$ (with respect to ν_1), denoted by $q_{\mathbf{Z}}$, has compact support, say $[0, 1]^d$. The model (2) enjoys the advantages of both the partially linear model (1) and the nonparametric additive model to the fully nonparametric model. It accommodates discrete covariates since we only require that ν_1 is a σ-finite measure, and also interaction effects between covariates by putting them into the parametric part. By the additive structure in the nonparametric part it avoids the curse of dimensionality, but retains the flexibility of the model. It also renders easy interpretation of the individual role of each covariate.

We discuss semi-parametric efficient estimation of the parameter $\boldsymbol{\beta}$ in the model (2). We present the semi-parametric Fisher information bound and provide an estimator that achieves the efficiency bound. Semi-parametric efficient estimation when $d = 1$ has been studied by Bhattacharya and Zhao [1], Cuzick [5] and Schick [17]. Their works can be easily extended to the model (1) for $d > 1$. Comparing the Fisher information bounds for the models (1) and (2), we find that the information bound under the model (2) is smaller than the bound under the model (1). In our semi-parametric model (2), we do not specify the distribution of the error term ϵ or the distribution q of the covariates. We show that one can do as well without knowing those distributions.

There have been a few works on the model (2). Opsomer and Ruppert [13] obtained a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$ by a backfitting method with undersmoothing. Recently Liang *et al.* [8] and Carroll *et al.* [4] studied the model with measurement error and repeated measurements, respectively. But they did not discuss semiparametric efficiency. The model (1) has been studied more often; see [19], among others. Most studies, however, are rather focused on the cases where there is only a single-dimensional (or at most low-dimensional) nonparametric function m. This is because high-dimension costs higher-order smoothness in theory and poor small sample performances in practice.

2. Semi-parametric efficiency

To avoid unnecessary complexity, we assume $m_0 = 0$. We also assume that ϵ is independent with $(\mathbf{X}, \mathbf{Z})$, and that g, the density of ϵ, is symmetric and is absolutely continuous with respect to the Lebesgue measure, having a derivative g' and finite Fisher information $\int (g')^2/g < \infty$. Below, we give a heuristic argument for deriving the semi-parametric efficiency and present a rigorous statement in a theorem.

Suppose that g is known and $p = 1$. We write $m(\mathbf{z}) = m_1(z_1) + \cdots + m_d(z_d)$ and adopt the convention $m_j(\mathbf{z}) = m_j(z_j)$. The logarithm of the joint density of $(Y, \mathbf{X}, \mathbf{Z})$ as a function of the parameters is given by $\ell(\beta, m; (y, x, \mathbf{z})) = \log g(y - x\beta - m(\mathbf{z}))$, neglecting those terms that do not depend on (β, m), and the log-likelihood of (β, m) by $\sum_{i=1}^n \ell(\beta, m; (Y^i, X^i, \mathbf{Z}^i))$. Let $\mathcal{H}$ denote the space of all additive functions m such that $m(\mathbf{z}) = m_1(z_1) + \cdots + m_d(z_d)$, $Em_j(Z_j) = 0$ and $Em(\mathbf{Z})^2 < \infty$.

Calculation of the Fisher information in a semi-parametric model is made locally: fix a value (β^0, m^0) of the parameter (β, m) and think of all 'regular' parametric submodels $\{(\beta, m_\beta) : \beta \in \mathbb{R}\}$ passing through (β^0, m^0), where $m_{\beta^0} = m^0$ and the mapping $\beta \mapsto m_\beta$ is Fréchet differentiable as a function from $\mathbb{R}$ to $\mathcal{H}$. Define $\varphi = g'/g$. Then, each finite-dimensional submodel $\{(\beta, m_\beta) : \beta \in \mathbb{R}\}$ has the score function

$$\begin{aligned} \mathrm{d}\ell(\beta, m_\beta)/\mathrm{d}\beta|_{\beta=\beta^0} &= \partial\ell(\beta, m^0)/\partial\beta|_{\beta=\beta^0} + \partial\ell(\beta^0, m)/\partial m|_{m=m^0}(\delta) \\ &= \varphi(\epsilon)X + \varphi(\epsilon)\delta(\mathbf{Z}), \end{aligned}$$

where $\delta = \partial m_\beta/\partial\beta|_{\beta=\beta^0} \in \mathcal{H}$ is the tangent of the mapping $\beta \mapsto m_\beta$ at β^0, and $\partial\ell/\partial m$ denotes the Fréchet derivative of ℓ with respect to m. This gives the Fisher information for estimating β in each submodel as $\mathcal{I}(\delta) \equiv E[\varphi(\epsilon)X + \varphi(\epsilon)\delta(\mathbf{Z})]^2$.

The Fisher information at $(\beta^0, m^0) \in \mathbb{R} \times \mathcal{H}$ in the *full* semi-parametric model typically equals to the Fisher information at $(\beta^0, m^0) \in \mathbb{R} \times \mathcal{H}$ in the most difficult parametric submodel that gives minimal $\mathcal{I}(\delta)$. Theorem 1 below demonstrates that this is the case with our problem. The least favorable direction δ^* that minimizes $\mathcal{I}(\delta)$ over $\delta \in \mathcal{H}$ is the solution of the following integral equation: for all $\delta \in \mathcal{H}$,

$$\begin{aligned} 0 &= E[\varphi(\epsilon)X + \varphi(\epsilon)\delta^*(\mathbf{Z})]\varphi(\epsilon)\delta(\mathbf{Z}) \\ &= I_g \cdot E\big[\big(E(X|\mathbf{Z}) + \delta^*(\mathbf{Z})\big)\delta(\mathbf{Z})\big], \end{aligned}$$

where $I_g = \int (g')^2/g$. This shows that $\delta^* = -\Pi(E(X|\mathbf{Z} = \cdot)|\mathcal{H})$, where $\Pi(\cdot|\mathcal{H})$ denotes the projection operator onto $\mathcal{H}$, and that the 'curve' m_β^* corresponding to the least favorable submodel equals $m_\beta^* = (\beta^0 - \beta)\Pi(E(X|\mathbf{Z} = \cdot)|\mathcal{H}) + m^0$. The Fisher information for the least favorable submodel is thus given by $\mathcal{I}(\delta^*) = I_g \cdot E[X - \Pi(E(X|\mathbf{Z})|\mathcal{H})]^2$, where, with a slight abuse of notation, we write $\Pi(E(X|\mathbf{Z} = \cdot)|\mathcal{H})(\mathbf{Z}) = \Pi(E(X|\mathbf{Z})|\mathcal{H})$.

The above arguments can be generalized to the case where $p > 1$. Writing $\eta_j = \Pi(E(X_j|\mathbf{Z} = \cdot)|\mathcal{H})$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^\top$, the least favorable direction equals $\boldsymbol{\delta}^* = -\boldsymbol{\eta}$ so that the Fisher information matrix for the least favorable submodel equals $\mathcal{I}(\boldsymbol{\delta}^*) = I_g \cdot E[\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z})][\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z})]^\top$. In the following theorem we show that the Fisher information $\mathcal{I}(\boldsymbol{\delta}^*)$ given above is indeed the semi-parametric information bound, as defined in [3], in our original semi-parametric model where the error density g and the density q of the covariate $(\mathbf{X}, \mathbf{Z})$ are not specified. To state the theorem, let $\mathcal{G}$ denote the set of all symmetric and absolutely continuous (with respect to the Lebesgue measure) functions g such that $I_g < \infty$. Let $\mathcal{Q}$ be an arbitrary class of density functions q. For the spaces of m, we consider Hilbert spaces defined by

$$\mathcal{H}(q) = \left\{ m \in L_2(q) : m(\mathbf{z}) = \sum_{j=1}^d m_j(z_j) \text{ and } Em_j(Z_j) = 0 \text{ for all } 1 \le j \le d \right\},$$

where $L_2(q)$ denotes the space of functions $m:\mathbb{R}^d \to \mathbb{R}$ such that $E_q m(\mathbf{Z})^2 < \infty$ and E_q means the expectation under the density q. The semi-parametric model (2) under study is then expressed as $\mathcal{P} = \{p(\cdot;\boldsymbol{\beta}, m, g, q) : \boldsymbol{\beta} \in \mathbb{R}^p, m \in \mathcal{H}(q), g \in \mathcal{G}, q \in \mathcal{Q}\}$. Let $(\boldsymbol{\beta}^0, m^0, g_0, q_0)$ be a fixed point where we are calculating the semi-parametric Fisher information. Denote by P_0 the distribution corresponding to $(\boldsymbol{\beta}^0, m^0, g_0, q_0)$, and by $I(P_0|\boldsymbol{\beta}, \mathcal{P})$ the semi-parametric Fisher information at P_0 for estimating $\boldsymbol{\beta}$ under the model $\mathcal{P}$. In the theorem below, the 'efficient score' ℓ^* for estimating $\boldsymbol{\beta}$ is the score for $\boldsymbol{\beta}$ at $\boldsymbol{\beta}^0$ in the least favorable parametric submodel that is indexed only by $\boldsymbol{\beta}$ and passes through P_0. Let E_0 denote the expectation under P_0.

Theorem 1. *The efficient score at P_0 for estimating $\boldsymbol{\beta}$ is given by*

$$\begin{aligned} &\ell^*(\mathbf{x}, \mathbf{z}, y; P_0|\boldsymbol{\beta}, \mathcal{P}) \\ &\quad = -[\mathbf{x} - \boldsymbol{\eta}(\mathbf{z})]\frac{g_0{}'}{g_0}\bigl(y - \mathbf{x}^\top\boldsymbol{\beta}^0 - m^0(\mathbf{z})\bigr), \end{aligned}$$

where $\boldsymbol{\eta} = (\Pi[E_0(X_j|\mathbf{Z} = \cdot)|\mathcal{H}(q_0)])_{j=1}^p$. The information bound at P_0 for estimating $\boldsymbol{\beta}$ equals $I(P_0|\boldsymbol{\beta}, \mathcal{P}) = I_{g_0} \cdot E_0[\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z})][\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z})]^\top$.

A proof of Theorem 1 can be found in an extended version of this paper that can be downloaded from http://stat.snu.ac.kr/theostat/papers/BEJ296_ExtendedVersion.pdf.

Let $\mathcal{P}_{\mathrm{PL}} \supset \mathcal{P}$ denote the semi-parametric model (1). One can show $I(P_0|\boldsymbol{\beta}, \mathcal{P}_{\mathrm{PL}}) = I_{g_0} \cdot E_0[\mathbf{X} - E_0(\mathbf{X}|\mathbf{Z})][\mathbf{X} - E_0(\mathbf{X}|\mathbf{Z})]^\top$ using the arguments to derive $I(P_0|\boldsymbol{\beta}, \mathcal{P})$. Note that $I(P_0|\boldsymbol{\beta}, \mathcal{P}) \geq I(P_0|\boldsymbol{\beta}, \mathcal{P}_{\mathrm{PL}})$ by the property of conditional expectation, and that the equality $I(P_0|\boldsymbol{\beta}, \mathcal{P}) = I(P_0|\boldsymbol{\beta}, \mathcal{P}_{\mathrm{PL}})$ holds if $E_0(X_j|\mathbf{Z} = \mathbf{z})$ are additive for all $1 \leq j \leq d$. According to the theory of semi-parametric efficiency, the minimal asymptotic variance that any regular estimator of $\boldsymbol{\beta}$ can achieve equals the inverse of the Fisher information matrix. The inequality $I(P_0|\boldsymbol{\beta}, \mathcal{P}) \geq I(P_0|\boldsymbol{\beta}, \mathcal{P}_{\mathrm{PL}})$ implies $I(P_0|\boldsymbol{\beta}, \mathcal{P})^{-1} \leq I(P_0|\boldsymbol{\beta}, \mathcal{P}_{\mathrm{PL}})^{-1}$, with equality holding if $E_0(X_j|\mathbf{Z} = \mathbf{z})$ are all additive.

Theorem 2. *Suppose $I(P_0|\boldsymbol{\beta}, \mathcal{P}_{\mathrm{PL}})$ is positive definite. Then, $I(P_0|\boldsymbol{\beta}, \mathcal{P})^{-1} < I(P_0|\boldsymbol{\beta}, \mathcal{P}_{\mathrm{PL}})^{-1}$ unless $E_0[\boldsymbol{\eta}(\mathbf{Z}) - E_0(\mathbf{X}|\mathbf{Z})][\boldsymbol{\eta}(\mathbf{Z}) - E_0(\mathbf{X}|\mathbf{Z})]^\top = \mathbf{O}$, where $\mathbf{O}$ is the $p \times p$ matrix with all entries being zero, and $A < B$ means that $B - A$ is non-negative definite and $A \neq B$.*

Theorem 2 tells that using an additive model for the nonparametric part can lead to drastic gains of efficiency in the estimation of the parametric components. The efficiency gains occur if the parametric covariates $\mathbf{X}$ are approximated by non-additive transformations of the nonparametric covariates $\mathbf{Z}$. If the approximation is exact, then estimation of the parametric part in the partially linear model (1) breaks down since $I(P_0|\boldsymbol{\beta}, \mathcal{P}_{\mathrm{PL}}) = \mathbf{O}$, while it does not with the partially linear additive model (2). If the approximation is very crude, one has large efficiency gains by using additive models for the nonparametric part.

3. Semi-parametric efficient estimation

Let $\boldsymbol{\beta}^0$ and m^0 denote the true parameter values. In this section we present the semi-parametric efficient estimator of $\boldsymbol{\beta}^0$ that achieves the minimal asymptotic variance $I(P_0|\boldsymbol{\beta},\mathcal{P})^{-1}$. The construction is based on a smooth backfitting technique and a profiling method. The latter is basically for estimating the least favorable curve, and is applied to the Gaussian error model to produce an initial estimator of $\boldsymbol{\beta}^0$ to be used in the construction of the semi-parametric efficient estimator.

3.1. Smooth backfitting methods

The smooth backfitting method, introduced by Mammen, Linton and Nielsen [10], is known to be a powerful technique for estimating additive regression functions. Since our profiling method involves smooth backfitting for non-additive functions, we discuss some properties of the method when the target function is not additive.

Let W be a random variable and $\{W^i\}$ be a random sample distributed as W. The smooth backfitting estimator, $\hat{m}_W^{\mathrm{add}}(\mathbf{z}) \equiv \hat{m}_{W,0}^{\mathrm{add}} + \hat{m}_{W,1}^{\mathrm{add}}(z_1) + \cdots + \hat{m}_{W,d}^{\mathrm{add}}(z_d)$, with responses W^i and regressors $\mathbf{Z}^i$, are defined as the solution of following integral equations:

$$\hat{m}_{W,j}^{\mathrm{add}} = \tilde{m}_{W,j} - \sum_{l=1,\neq j}^{d} \hat{\Pi}_j(\hat{m}_{W,l}^{\mathrm{add}}) - \hat{m}_{W,0}^{\mathrm{add}}, \qquad 1 \le j \le d, \tag{3}$$

with the constraints $\langle \hat{m}_{W,j}^{\mathrm{add}}, \mathbf{1}\rangle = 0$ for $1 \le j \le d$. Here, $\hat{m}_{W,0}^{\mathrm{add}} = n^{-1}\sum_{i=1}^n W^i$ and $\tilde{m}_{W,j}(z_j)$ denotes the marginal regression kernel estimator obtained by regressing W^i on Z_j^i only. The operator $\hat{\Pi}_j$ stands for a projection onto a Hilbert space equipped with a scalar product $\langle\cdot,\cdot\rangle$; see [23] for details. For example, in the case where $\tilde{m}_{W,j}(z_j)$ are the local constant marginal estimators, $\langle g,h\rangle = \int g(\mathbf{z})h(\mathbf{z})\hat{q}_{\mathbf{Z}}(\mathbf{z})\,\mathrm{d}\mathbf{z}$, with $\hat{q}_{\mathbf{Z}}(\cdot)$ being the kernel estimator of the design density $q_{\mathbf{Z}}$. Smoothing to the direction of Z_j is done by the boundary corrected kernel $K_{h_j}(u,v) = c_j(v)h_j^{-1}K^0((u-v)/h_j)$, where K^0 is a base kernel function, h_j is the bandwidth, and $c_j(v)$ is a factor that gives $\int K_{h_j}(u,v)\,\mathrm{d}u = 1$.

Let $m_W(\mathbf{z}) = E(W|\mathbf{Z}=\mathbf{z})$. We do not assume that m_W is an additive function. Define $m_W^{\mathrm{add}} = m_{W,1}^{\mathrm{add}} + \cdots + m_{W,d}^{\mathrm{add}}$ to be the projection of m_W onto the space of additive functions $\mathcal{H}(q_{\mathbf{Z}})$. Then, $E[m_W(\mathbf{Z}) - E(W) - m_W^{\mathrm{add}}(\mathbf{Z})]\delta(\mathbf{Z}) = 0$ for any $\delta \in \mathcal{H}(q_{\mathbf{Z}})$. The additive function $m_W^{\mathrm{add}}(\mathbf{z})$ plays the role of the target function that the smooth backfitting estimator $\hat{m}_W^{\mathrm{add}}(\mathbf{z})$ aims at. Lu *et al.* [9] discussed the property of the smooth backfitting estimators under non-additive regression models in the context of spatial data analysis. However, they treated only the case where the bandwidth is asymptotic to $n^{-1/5}$. Below, we give a uniform expansion of the smooth backfitting estimator for a wider range of the bandwidths, after tedious asymptotic calculation following the lines of the arguments in [10]. To state the theorem, let $\varepsilon = W - E(W) - m_W^{\mathrm{add}}(\mathbf{Z})$ and define ε^i accordingly. Let $\tilde{m}_{\varepsilon,j}(z_j)$ and $\tilde{m}_{\varepsilon,j}^{LL}(z_j)$ denote, respectively, the local constant and linear estimators with responses ε^i and the scalar regressors Z_j^i. Let h_j be the bandwidth associated with Z_j. The theorem relies on the following assumptions.

Assumptions A.

A1. *For* $1 \le j \ne k \le d$, q_{Z_j,Z_k} *are bounded away from zero and infinity on its support*, $[0,1]^2$, *and have continuous partial derivatives*.

A2. *The base kernel function* K^0 *is symmetric*, *supported on a compact support and has bounded derivative*.

A3. *The functions* $m^{\mathrm{add}}_{W,j}$*'s are twice continuously differentiable.*

A4. $E|W - m_W(\mathbf{Z})|^{r_0} < \infty$ *for some* $r_0 > 5/2$.

Theorem 3. *Assume that the conditions* A1–A4 *hold, and that* h_j *are asymptotic to* $n^{-\alpha}$ *for* $1/5 \le \alpha < 1/2$. *Then, for* $1 \le j \le d$, *it holds that*

$$\sup_{z_j \in [0,1]} |\hat{m}^{\mathrm{add}}_{W,j}(z_j) - m^{\mathrm{add}}_{W,j}(z_j) - h_j a_{1,j,n}(z_j) - h_j^2 a_{2,j}(z_j) - \tilde{m}_{\varepsilon,j}(z_j)| = \mathrm{o}_p((nh_j)^{-1/2})$$

in the local constant case, *and that*

$$\sup_{z_j \in [0,1]} |\hat{m}^{\mathrm{add}}_{W,j}(z_j) - m^{\mathrm{add}}_{W,j}(z_j) - h_j^2 a_{3,j}(z_j) - \tilde{m}^{LL}_{\varepsilon,j}(z_j)| = \mathrm{o}_p((nh_j)^{-1/2})$$

in the local linear case, *for some functions* $a_{1,j,n}$ *that are uniformly bounded and non-zero only for* $z_j \in [0, ch_j) \cup (1 - ch_j, 1]$ *for some constant* $0 < c < \infty$, *and for some functions* $a_{2,j}$ *and* $a_{3,j}$ *that are continuous.*

A proof of Theorem 3 can be found in an extended version of this paper that can be downloaded from http://stat.snu.ac.kr/theostat/papers/BEJ296_ExtendedVersion.pdf.

3.2. Profiling with Gaussian error models

We apply a profiling technique to remove the infinite-dimensional parameter m in the estimation of $\boldsymbol{\beta}^0$. For a general framework of profiling approaches to semi-parametric models, we refer to [18]. See also [12] for a more recent work on profile likelihood.

Define $\hat{m}^{\mathrm{add}}_{\mathbf{X}} = (\hat{m}^{\mathrm{add}}_{X_1}, \ldots, \hat{m}^{\mathrm{add}}_{X_p})^\top$. We note that $\hat{m}^{\mathrm{add}}_{\mathbf{X}}$ is an estimator of $\boldsymbol{\eta}$ and $\hat{m}^{\mathrm{add}}_{Y}$ is an estimator of $\boldsymbol{\beta}^{0\top}\boldsymbol{\eta} + m^0$. For each given $\boldsymbol{\beta}$, let $\hat{m}^{\mathrm{add}}(\mathbf{z}; \boldsymbol{\beta}) = \sum_{j=1}^d \hat{m}^{\mathrm{add}}_j(z_j; \boldsymbol{\beta})$ be the smooth backfitting estimator obtained by taking $Y^i - \mathbf{X}^{i\top}\boldsymbol{\beta} = \mathbf{X}^{i\top}(\boldsymbol{\beta}^0 - \boldsymbol{\beta}) + m^0(\mathbf{Z}^i) + \epsilon^i$ as responses and $\mathbf{Z}^i$ as covariates. Recall that the least favorable curve is given by $m^*(\cdot, \boldsymbol{\beta}) \equiv \boldsymbol{\eta}^\top(\boldsymbol{\beta}^0 - \boldsymbol{\beta}) + m^0$. Thus, we may regard $\hat{m}^{\mathrm{add}}(\cdot; \boldsymbol{\beta})$ as an estimator of the least favorable curve $m^*(\cdot, \boldsymbol{\beta})$. Since $\hat{m}^{\mathrm{add}}(\mathbf{z}; \boldsymbol{\beta}) = \hat{m}^{\mathrm{add}}_Y(\mathbf{z}) - \hat{m}^{\mathrm{add}}_{\mathbf{X}}(\mathbf{z})^\top\boldsymbol{\beta}$ by the fact that the smooth backfitting operation is linear in response vectors, the estimated profile likelihood based on the Gaussian error model is given by

$$-\sum_{i=1}^n [Y^i - \mathbf{X}^{i\top}\boldsymbol{\beta} - \hat{m}^{\mathrm{add}}(\mathbf{Z}^i; \boldsymbol{\beta})]^2 = -\sum_{i=1}^n \bigl[Y^i - \hat{m}^{\mathrm{add}}_Y(\mathbf{Z}^i) - \bigl(\mathbf{X}^i - \hat{m}^{\mathrm{add}}_{\mathbf{X}}(\mathbf{Z}^i)\bigr)^\top \boldsymbol{\beta}\bigr]^2.$$

The estimator that maximizes the above Gaussian profile likelihood is then given by

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \tilde{\mathbf{X}}^i \tilde{\mathbf{X}}^{i\top}\right)^{-1} \left(\sum_{i=1}^{n} \tilde{\mathbf{X}}^i \tilde{Y}^i\right),$$

where $\tilde{\mathbf{X}}^i = \mathbf{X}^i - \hat{m}_{\mathbf{X}}^{\mathrm{add}}(\mathbf{Z}^i)$ and $\tilde{Y}^i = Y^i - \hat{m}_Y^{\mathrm{add}}(\mathbf{Z}^i)$.

Theorem 4. *Suppose that the assumptions* A1–A4 *hold with $W = Y$ and X_j, $1 \le j \le p$. Also, assume that $E[\exp(|X_j - E(X_j|\mathbf{Z})|)|\mathbf{Z}] < C$ a.s. for some $C > 0$, $1 \le j \le p$. If the bandwidths h_j are asymptotic to $n^{-\alpha}$ for $1/5 \le \alpha < 1/2$, then it holds that*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \overset{d}{\Rightarrow} N\left(\mathbf{0}, \mathrm{var}(\epsilon)\left[E\left(\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z})\right)\left(\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z})\right)^\top\right]^{-1}\right).$$

A proof of Theorem 4 is given in the Appendix. We note that the asymptotic variance of the estimator $\hat{\boldsymbol{\beta}}$ is larger than $I(P_0|\boldsymbol{\beta}, \mathcal{P})^{-1}$. This can be seen directly from a projection property. In fact, $\mathrm{var}(\epsilon) \ge I_g^{-1}$ and the equality hold if g is Gaussian. This means that the estimator $\hat{\boldsymbol{\beta}}$ achieves the semi-parametric efficiency in the reduced model where g is specified as a Gaussian density. It is also interesting to see what happens if $\eta_0(\mathbf{X}, \mathbf{Z}) \equiv E_0(Y|\mathbf{X}, \mathbf{Z})$ does not belong to the partially linear additive model of the form (2). In this case, our estimator of η_0 converges to η^*, which is the $L_2(q)$-projection of η_0 onto the space

$$\mathcal{F} = \{f \in L_2(q) \mid f(\mathbf{x}, \mathbf{z}) = \boldsymbol{\beta}^\top \mathbf{x} + m(\mathbf{z}),\ \boldsymbol{\beta} \in \mathbb{R}^p,\ m \in \mathcal{H}\}. \tag{4}$$

3.3. Adapting to unknown error density

In this subsection, we construct the semi-parametric efficient estimator that achieves the minimal asymptotic variance discussed in Section 2. We follow the approach adopted by Bickel [2], Schick [16,17], Park [14], Cuzick [5] and Bhattacharya and Zhao [1]. Write $I = I(P_0|\boldsymbol{\beta}, \mathcal{P})$ and define $\boldsymbol{\beta}_n^* = \boldsymbol{\beta}^0 - I^{-1} n^{-1} \sum_{i=1}^n [\mathbf{X}^i - \eta(\mathbf{Z}^i)]\varphi(\epsilon)$. Then, the random sequence $\boldsymbol{\beta}_n^*$ achieves the efficiency bound. We plug some estimators of the unknown quantities into $\boldsymbol{\beta}_n^*$. We estimate the error density g by using the 'pseudo' errors $\hat{\epsilon}^i \equiv \tilde{Y}^i - \tilde{\mathbf{X}}^{i\top}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the Gaussian profile estimator constructed in Section 3.2. In particular, we take $\hat{g}(t) = b + (na)^{-1} \sum_{i=1}^n L((t - \hat{\epsilon}^i)/a)$ and $\hat{g}'(t) = \mathrm{d}\hat{g}(t)/\mathrm{d}t$, where a and b are positive constants that depend on the sample size n, and L is a symmetric differentiable density function. Define

$$\hat{I} = \left(n^{-1} \sum_{i=1}^{n} \tilde{\mathbf{X}}^i \tilde{\mathbf{X}}^{i\top}\right)\left(n^{-1} \sum_{i=1}^{n} \hat{\varphi}(\hat{\epsilon}^i)^2\right),$$

where $\hat{\varphi}$ is the 'symmetrized' estimator of φ defined by $\hat{\varphi}(e) = [(\hat{g}'/\hat{g})(e) - (\hat{g}'/\hat{g})(-e)]/2$. Our semi-parametric efficient estimator is then given by

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \hat{I}^{-1} \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{X}}^i \hat{\varphi}(\hat{\epsilon}^i).$$

Assumptions B.

B1. *The error ϵ has an absolutely continuous and symmetric density g with respect to the Lebesgue measure, μ, and $I_g = \int (g'^2/g)\,\mathrm{d}\mu < \infty$.*

B2. *The kernel L is a symmetric density function with three bounded and Lipschitz continuous derivatives.*

B3. *The sequences a and b converge to zero, as $n \to \infty$, and satisfy $n^{1/2} h_j b(a^2 \wedge b^2) \to \infty$ and $a^2/\{h_j (\log n)^2\} \to \infty$ for all $1 \le j \le d$.*

Theorem 5. *Assume that the conditions of Theorem 4 and the assumptions* B1–B3 *hold. Then,* $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \overset{d}{\Rightarrow} N(\mathbf{0}, I(P_0|\boldsymbol{\beta}, \mathcal{P})^{-1})$.

A proof of Theorem 5 is given in the Appendix. For a choice of the bandwidth a in $\hat{g}$, one can devise a data-driven choice along the lines of Park [15]. For h, one can follow the approach of Mammen and Park [11]. In this adaptation step, misspecification of the model may result in a meaningless estimator. This is in contrast to the estimation in the initial step where the procedure estimates the projection of the mean function onto the model space $\mathcal{F}$ at (4). The reason is that the residuals from the initial step include not only the pure errors but also the deviation of the true regression function from its projection onto $\mathcal{F}$. These residuals mislead estimation of the score function.

4. Numerical properties

We generated 500 random samples of the size $n = 400$. We used Epanechnikov kernel for the regression and the Gaussian density kernel for the estimation of the score function. We applied a local constant version of smooth backfitting. We took $m_1(z_1) = \sin\{2\pi(z_1 - 0.5)\}$ and $m_2(z_2) = z_2 - 0.5 + \sin\{2\pi(z_2 - 0.5)\}$. We set $m_0 = 3$, $\beta_1 = 1.5$ and $\beta_2 = 0.8$. We drew (Z_1, Z_2) from $N_2((0.5, 0.5)^\top, \Sigma)$ truncated to $[0, 1]^2$, where $\Sigma = \{(1 - \rho)I + \rho\mathbf{1}\mathbf{1}^\top\}/4$. We generated $X_1 = CZ_1(1 - 2Z_2) + U$ for some constant C, where $U \sim N(0, 0.5)$, and X_2 from Bernoulli$(p(X_1, Z_1, Z_2))$, where $p(X_1, Z_1, Z_2) = g(\exp((Z_1 + Z_2)/2) + \sin(2\pi Z_1) - X_1^2)$ and $g(t) = \exp(t)/(1 + \exp(t))$. Note that $E(X_1|\mathbf{Z} = \cdot)$ is orthogonal to the space of additive functions.

We compared the Gaussian profile estimator (SAM), given in Section 3.2, and the profile kernel estimator (PL), given in [19], which is for the partial linear model without the additive structure. For this, we generated ϵ from $N(0, 1)$ and set $\rho = 0$. In the case where $p = 1$, that is, X_2 does not enter the model, the theoretical value of the ratio of the asymptotic variance of SAM to that of PL equals $1/(1 + 0.1707C^2)$. The empirical values from our simulation study for the bandwidth pair (h_1, h_2) that gave the best mean square error (MSE) were 0.7818, 0.5868 and 0.4082 for $C = 1, 2$ and 3, respectively, which nearly coincided with the theoretical values. We tried other values of ρ, but the lesson was the same. In the case where $p = 2$ and $d = 5$ with $(Z_1, \ldots, Z_5)$ from $N_5((0.5, \ldots, 0.5)^\top, \Sigma)$ truncated to $[0, 1]^5$ and $m_j(z_j) = z_j^2$ for $3 \le j \le 5$, we took $C = 1$ and found that SAM beat PL for all bandwidth choices that we tried. The Gaussian profile estimator was stable while PL broke down for small bandwidths. The best MSE of SAM

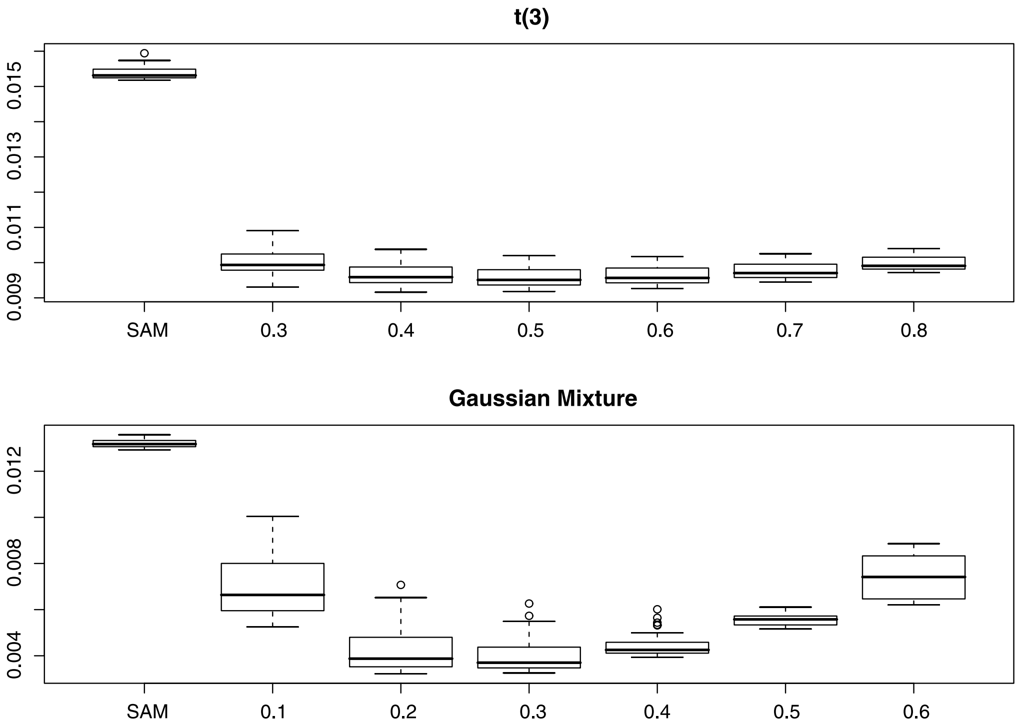

Figure 1. Mean square errors of SAM and ASAM.

and that of PL, respectively, for various choices of the bandwidth pair (h_1, h_2) were 0.0032 and 0.0051 for β_1 and 0.0186 and 0.0269 for β_2.

Next, we compared SAM with the semi-parametric efficient estimator (ASAM). For this, we considered the case where $p = d = 2, C = 1$ and $\rho = 0.8$, and generated ϵ from $N(0, 1)$, t-distribution with degree of freedom 3, and $\frac{1}{2}N(-1.5, 0.6^2) + \frac{1}{2}N(1.5, 0.6^2)$. For ASAM, we took $b = 0.01$, and six different choices of a: $a_i = 0.3 + 0.1i, 0 \le i \le 5$, for $N(0, 1)$ and $t(3)$ errors and $a_i = 0.1 + 0.1i, 0 \le i \le 5$, for the Gaussian mixture error. We used 36 different choices for the bandwidth pair $(h_1, h_2) \in \{0.05, 0.10, , \dots, 0.30\}^2$. Figure 1 is for the estimators of β_1. Each box-plot was obtained from the 36 values of MSE that corresponded to the 36 bandwidth pairs (h_1, h_2). For ASAM, the value of a is indicated on the horizontal scale. The figure suggests that the values of the MSE of ASAM are far smaller than those of SAM for the entire range of the bandwidth a, under $t(3)$ and the Gaussian mixture error models. The box-plots for the Gaussian error model are not given here since SAM and ASAM gave similar performance. The results for β_2 are not reported either since they give a similar lesson.

5. Boston housing data

We applied the semi-parametric efficient estimators to Boston housing data as an illustration. As in [6,22], we took the median price in 1,000 USD (MEDV) as the response Y. Also, we chose as covariates X_1, X_2 and $Z_1, \dots, Z_6$, respectively, the eight variables LSTAT (percentage values of lower status population), CHAS (a dummy variable that takes the value 1 if the tract borders Charles River; 0 otherwise), CRIM (per capita crime rate), RM (average numbers of rooms per dwelling), NOX (nitric oxides concentration), PTRATIO (pupil–teacher ratios), DIS (weighted distances to five Boston employment centers) and TAX (full-value property tax rate per 10,000 USD). The logarithms of LSTAT, DIS and TAX were taken to reduce sparse areas, as in [22]. We chose the model $Y = m_0 + \beta_1 X_1 + \beta_2 X_2 + \sum_{j=1}^{6} m_j(Z_j) + \epsilon$. In the data set, there were 16 cases for which Y took the maximal value 50. These may be censored responses that one may remove from analysis. Indeed, an initial analysis showed a strong asymmetry in the distribution of the residuals, which led us to exclude the 16 cases for further analysis. For additive regression, we applied local constant smooth backfitting with the Epanechnikov kernel and bandwidths h_j chosen by a rule of thumb.

With SAM, we obtained $\hat{\beta}_1 = -6.203$ and $\hat{\beta}_2 = 0.985$. Their estimated standard errors were 0.420 and 0.597, respectively. This suggests that $\hat{\beta}_2$ is not strongly significant while $\hat{\beta}_1$ is. The generalized R^2 was 0.862. For ASAM, in the estimation of the score function, we used a bandwidth a that was obtained by R function `bw.SJ()`. With ASAM, we got $\tilde{\beta}_1 = -6.172$ and $\tilde{\beta}_2 = 1.366$, and their estimated standard errors were 0.399 and 0.567, respectively. Thus, with ASAM, both the estimated coefficients are strongly significant. This may be an indication that a Gaussian error model is not appropriate for the data set. The generalized R^2 was almost the same as in the analysis with SAM.

Appendix

Proof of Theorem 4. We only treat the case with local constant smooth backfitting. The case with local linear smooth backfitting can be dealt with similarly. We prove

$$n^{-1/2} \sum_{i=1}^{n} \tilde{\mathbf{X}}^i (\tilde{Y}^i - \tilde{\mathbf{X}}^{i\top} \boldsymbol{\beta}^0) - n^{-1/2} \sum_{i=1}^{n} \big(\mathbf{X}^i - \boldsymbol{\eta}(\mathbf{Z}^i)\big)\epsilon^i = \mathrm{o}_p(1). \tag{5}$$

Write $\Delta(\mathbf{z}) = m^0(\mathbf{z}) - \hat{m}^{\mathrm{add}}(\mathbf{z}; \boldsymbol{\beta}^0)$. The left-hand side of equation (5) equals $C_1 + C_2 + C_3$, where $C_1 = n^{-1/2} \sum_{i=1}^{n} (\mathbf{X}^i - \boldsymbol{\eta}(\mathbf{Z}^i))\Delta(\mathbf{Z}^i)$, $C_2 = n^{-1/2} \sum_{i=1}^{n} (\boldsymbol{\eta}(\mathbf{Z}^i) - \hat{m}_{\mathbf{X}}^{\mathrm{add}}(\mathbf{Z}^i))\epsilon^i$, and $C_3 = n^{-1/2} \sum_{i=1}^{n} (\boldsymbol{\eta}(\mathbf{Z}^i) - \hat{m}_{\mathbf{X}}^{\mathrm{add}}(\mathbf{Z}^i))\Delta(\mathbf{Z}^i)$. Write $\Delta(\mathbf{z}) = \Delta_0 + \sum_{j=1}^{d} \Delta_j(z_j)$. By Theorem 3, standard techniques of kernel smoothing, integration by part and the representation of m^0 and $\hat{m}^{\mathrm{add}}(\mathbf{z}; \boldsymbol{\beta}^0)$ as a solution of an integral equation with differentiable kernel (see equation (3)), we have

$$\sup_{\mathbf{z} \in [0,1]^d} |\Delta(\mathbf{z})| = \mathrm{o}_p(\delta_n), \qquad \sup_{z_j \in [0,1]} \left| \frac{\mathrm{d}}{\mathrm{d}z_j} \Delta_j(z_j) - h_j b_{n,j}(z_j) \right| = \mathrm{o}_p(\delta_n)$$

for some uniformly bounded non-random functions $b_{n,j}$, where $\delta_n = n^{-a}$ for some $a \in (0, 1/2 - \alpha)$. These imply that $\delta_n^{-1} \Delta \in B(\mathbf{0}, 1)$ with probability tending to one, where $B(\mathbf{0}, 1)$ denotes a class of additive functions $\sum_{j=1}^{d} g_j(z_j)$ such that each g_j is a real function defined on [0, 1] and satisfies $\sup_{t,t'\in[0,1]} |g_j(t) - g_j(t')| \le |t - t'|$. The covering number with bracketing of $B(\mathbf{0}, 1)$ with respect to sup-norm, $N_{[\cdot]}(\eta) \equiv N_{[\cdot]}(\eta, B(\mathbf{0}, 1), \|\cdot\|_\infty)$, is bounded by $(2\eta^{-1})^d 3^{d\eta^{-1}}$. Define random functionals $F(X_j^i, \mathbf{Z}^i) : B(\mathbf{0}, 1) \to \mathbb{R}$ by $[F(X_j^i, \mathbf{Z}^i)](g) = (X_j^i - \eta_j(\mathbf{Z}^i))g(\mathbf{Z}^i)$, and $F_j : B(\mathbf{0}, 1) \to \mathbb{R}$ by $F_j = n^{-1/2} \sum_{i=1}^n F(X_j^i, \mathbf{Z}^i)$. Then, using Corollary 8.8 of van de Geer [20] and the tail condition assumed in the theorem, one can show $\sup_{g\in B(\mathbf{0},1)} |F_j g| = \mathrm{O}_p(1)$. Let $C_{1,j}$ denote the jth element of C_1. Since $P(|\delta_n^{-1} C_{1,j}| > M) \le P(\sup_{g\in B(\mathbf{0},1)} |F_j g| > M) + P(\delta_n^{-1}\Delta \notin B(\mathbf{0}, 1))$, we obtain $C_{1,j} = \mathrm{O}_p(\delta_n) = \mathrm{o}_p(1)$. One can prove $C_2 = \mathrm{o}_p(1)$ using a truncation argument with Theorem 3 and applying the Chebyshev inequality conditioning on $(\mathbf{X}^i, \mathbf{Z}^i)$. The fact that $C_3 = \mathrm{o}_p(1)$ follows from $P(Z_j^i$ lies in $[0, ch_j) \cup (1 - ch_j, 1]) = \mathrm{O}(h_j)$ for some constant $0 < c < \infty$ and Theorem 3. □

Proof of Theorem 5. We will show that $\tilde{\boldsymbol\beta} - \boldsymbol\beta_n^* = \mathrm{o}_p(n^{-1/2})$. It suffices to show

$$\hat{I}^{-1} n^{-1} \sum_{i=1}^n \tilde{\mathbf{X}}^i \hat\varphi(\hat\epsilon^i) = \hat{\boldsymbol\beta} - \boldsymbol\beta^0 + I^{-1} n^{-1} \sum_{i=1}^n [\mathbf{X}^i - \eta(\mathbf{Z}^i)]\varphi(\epsilon^i) + \mathrm{o}_p(n^{-1/2}). \tag{6}$$

By Theorem 3 and standard techniques of kernel smoothing along with assumption B3, it holds that, uniformly over i,

$$\hat\varphi(\hat\epsilon^i) = \hat\varphi(\epsilon^i) - \tilde{\mathbf{X}}^{i\top}(\hat{\boldsymbol\beta} - \boldsymbol\beta^0)\hat\varphi'(\epsilon^i) - \{\hat m^{\mathrm{add}}(\mathbf{Z}^i; \hat{\boldsymbol\beta}) - m^0(\mathbf{Z}^i)\}\hat\varphi'(\epsilon^i) + \mathrm{o}_p(n^{-1/2}). \tag{7}$$

Also, using the proof of Lemma 4.1 in [2] and standard calculus, one can show $\hat{I} = I + o_p(1)$ and $n^{-1} \sum_{i=1}^n \tilde{\mathbf{X}}^i \tilde{\mathbf{X}}^{i\top} \hat\varphi'(\epsilon^i) = -I + \mathrm{o}_p(1)$. Thus, the proof of the theorem is completed if we verify

$$n^{-1} \sum_{i=1}^n \tilde{\mathbf{X}}^i \{\hat m^{\mathrm{add}}(\mathbf{Z}^i; \hat{\boldsymbol\beta}) - m^0(\mathbf{Z}^i)\}\hat\varphi'(\epsilon^i) = \mathrm{o}_p(n^{-1/2}); \tag{8}$$

$$n^{-1} \sum_{i=1}^n \tilde{\mathbf{X}}^i \hat\varphi(\epsilon^i) - n^{-1} \sum_{i=1}^n \{\mathbf{X}^i - \boldsymbol\eta(\mathbf{Z}^i)\}\varphi(\epsilon^i) = \mathrm{o}_p(n^{-1/2}). \tag{9}$$

Proofs of (8) and (9) can be based on the following lemma, which follows from Corollary 2.7.4 in [21] and assumption B2 on L. Note that the moment condition on ϵ ensures the entropy bound. To state the lemma, define

$$\mathcal{C}_M^\alpha(\mathcal{X}) = \left\{ f : \mathcal{X} \to \mathbb{R} : \sup_x |f(x)| + \sup_{x,y} \frac{|f(x) - f(y)|^\alpha}{|x - y|} \le M \right\}$$

for a set $\mathcal{X} \subset \mathbb{R}$ and a real number $\alpha \in (0, 1]$. Let $\|\cdot\|_g$ denote the L_2 norm with respect to the density g.

Lemma 1. *Assume the conditions of Theorem 5. Then there exists a constant M such that, with probability tending to one, $b(a \wedge b)\hat\varphi \in \mathcal{C}_M^1(\mathbb{R})$, $[nh_{\max} a^6 b/(\log n)^2]^{1/2}(\hat\varphi - \varphi_n) \in \mathcal{C}_M^1(\mathbb{R})$*

and $b(a^2 \wedge b^2)\hat{\varphi}' \in \mathcal{C}_M^1(\mathbb{R})$. *Moreover, there exist constants* $\delta > 0$ *and* $C_1 > 0$ *such that* $\log N_{[\cdot]}(\eta, \mathcal{C}_M^1(\mathbb{R}), \|\cdot\|_g) \leq C_1\eta^{-(2-\delta)}$. $\square$

Acknowledgement

Research of Kyusang Yu was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0023488). Research of Byeong U. Park was supported by the Mid-career Researcher Program through NRF grant funded by the MEST (No. 2010-0017437).

References

[1] Bhattacharya, P. and Zhao, P. (1997). Semiparametric inference in a partial linear model. *Ann. Statist.* **25** 244–262. MR1429924

[2] Bickel, P. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671. MR0663424

[3] Bickel, P., Klaassen, A., Ritov, Y. and Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore, MD: Johns Hopkins Univ. Press. MR1245941

[4] Carroll, R., Maity, A., Mammen, E. and Yu, K. (2009). Efficient semiparametric marginal estimation for the partially linear additive model for longitudinal/clustered data. *Statist. Biosci.* **1** 10–31.

[5] Cuzick, J. (1992). Efficient estimates in semiparametric additive regression models with unknown error distribution. *Ann. Statist.* **20** 1129–1136. MR1165611

[6] Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11** 1031–1057. MR2189080

[7] Koul, H.L. and Schick, A. (1997). Efficient estimation in nonlinear autoregressive time-series models. *Bernoulli* **3** 247–277. MR1468305

[8] Liang, H., Thurston, S., Ruppert, D., Apanasovich, T. and Hauser, R. (2008). Additive partial linear models with measurement errors. *Biometrika* **95** 667–678.

[9] Lu, Z., Lundervold, L., Tjøstheim, D. and Yao, Q. (2007). Exploring spatial nonlinearity using additive approximation. *Bernoulli* **13** 447–472. MR2331259

[10] 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. MR1742496

[11] Mammen, E. and Park, B.U. (2005). Bandwidth selection for smooth backfitting in additive models. *Ann. Statist.* **33** 1260–1294. MR2195635

[12] Murphy, S. and van der Vaart, A. (2000). On profile likelihood (with comments). *J. Amer. Statist. Assoc.* **95** 449–485. MR1803168

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

[14] Park, B.U. (1990). Efficient estimation in the two sample semiparametric location-scale model. *Probab. Theory Related Fields* **86** 21–39. MR1061946

[15] Park, B.U. (1993). A cross-validatory choice of smoothing parameter in adaptive location estimation. *J. Amer. Statist. Assoc.* **88** 848–854. MR1242935

[16] Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14** 1139–1151. MR0856811

[17] Schick, A. (1993). On efficient estimation in regression models. *Ann. Statist.* **21** 1486–1521. MR1241276

[18] Severini, T. and Wong, W. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20** 1768–1802. MR1193312

[19] Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* **50** 413–436. MR0970977

[20] van de Geer, S. (2000) *Empirical Processes in M-Estimation.* Cambridge: Cambridge Univ. Press.

[21] van der Vaart, A. and Wellner, J. (1996) *Weak Convergence and Empirical Processes. With Applications to Statistics.* New York: Springer. MR1385671

[22] Wang, J. and Yang, L. (2009). Efficient and fast spline-backfitted kernel smoothing of additive models. *Ann. Inst. Statist. Math.* **61** 663–690. MR2529970

[23] Yu, K., Mammen, E. and Park, B.U. (2008). Smooth backfitting in generalized additive models. *Ann. Statist.* **36** 228–260. MR2387970

Received April 2010