Maximum likelihood curve estimation

Abstract

We consider a maximum likelihood based local linear regression method for the classical nonparametric regression model. To estimate the nonparametric regression function, we maximize an estimated likelihood via a density estimation of the random error. An algorithm is provided to achieve the maximization. The proposed estimator is consistent and asymptotically normal, and is optimal in the sense that the estimator of the linear functional form of the nonparametric regression function is semiparametrically efficient. Simulation studies provide supportive evidence. The proposed method is extended to estimate the partial linear regression model. Analysis of a real example is presented.

Key words and phrases: Maximum likelihood function; the simple nonparametric regression model; local linear method; semiparametric efficiency.

1 Introduction

Let (X, Y) be a random sample from a bivariate population. It is of common interest to explore the association between X and Y. One of the most fundamental statistical models to describe such an association is via the nonparametric regression model

$$Y = g(X) + \varepsilon, \tag{1.1}$$

where $g(\cdot)$ is the unknown regression function and ε is the random error with unknown distribution f.

There have been many studies focused on the estimation of the unknown regression function $g(\cdot)$. There are, typically, two classes of estimators: spline smoothing methods (Wahba, 1978; Carter and Eagleson, 1992; Carter et al., 1992; Eubank, 1988; Green and Silverman 1994) and kernel-based estimators (Nadaraya, 1964; Watson, 1964; Gasser and Muller, 1979; Muller and Stadtmuller, 1987; Hall and Carroll, 1989; Hall and Marron, 1990; Neumann, 1994 and Fan and Gijbel, 1996). These estimators are primarily based on the penalized least squares and the local least squares methods, respectively. The least squares (LS) criterion is appropriate when the error ε is normally distributed but may be unstable and inefficient if the error ε deviate significantly from the normal distribution. Hence, some adjustments are made, for example, the kernel M-smoother (Härdle, 1989), median smoothing (Tukey, 1977), locally weighted regression (Stone, 1977 and Cleveland, 1979), and the local least absolute method (Wang and Scott, 1994). However, all of these methods choose the criterion functions based on mathematical convenience or robust considerations, rather than from the perspective of the data itself. As a result, none of these existing methods is proven optimal. In other words, efficient curve estimation of $g(\cdot)$ in the model (1.1) is unclear or has not been explored. One of the major difficulties is the computational complexity for more sophisticated and possibly more accurate estimation. With the availability of ever increasing computational capacity, many previously considered implausible methodologies are now computationally feasible. The purpose of this paper is to introduce an estimation method using an optimal criterion function to produce more accurate and semiparametrically efficient curve estimation (as defined by Bickel et al., 1993) for the nonparametric regression curve $g(\cdot)$.

This paper is organized as follows. In Section 2, we give the criterion function and the estimator of $g(\cdot)$. We derive the asymptotic properties including consistency, asymptotic normality and semiparametric efficiency of the proposed estimator in Section 3. The proposed method is extended to estimate the partial linear regression model in Section 4. We present simulation results of the robustness and efficiency of the estimator in Section 5. Finally, we illustrate our methods using a real example in Section 6.

2 Estimation

If the unknown function $g(\cdot)$ is parameterized, the parameters can be estimated by minimizing

$$l\{g(\cdot)\} = \sum_{i=1}^{n} \rho\{Y_i - g(X_i)\}, \tag{2.1}$$

where ρ is a given loss function. The least squares and least absolute criteria correspond to $\rho(t) = t^2$ and $\rho(t) = |t|$. When the form of the unknown function $g(\cdot)$ is not available, we can only rely on its qualitative trait. Assume that $g(\cdot)$ is smooth so that Taylor's expansions is applicable: for each given x and w around x,

$$g(w) \quad \approx \quad g(x) + \dot{g}(x)(w - x) \equiv \alpha + \beta(w - x). \tag{2.2}$$

Substituting this into (2.1), we obtain the following local ρ-objective function:

$$\sum_{i=1}^{n} \rho\{Y_i - \alpha - \beta(X_i - x)\} K_i(x), \tag{2.3}$$

where $K_i(x) = K_h(X_i - x)$, $K_h(u) = K(u/h)/h$, K is a nonnegative symmetric kernel function with support on $[-1, 1]$ and h is the bandwidth. The minimized values of α and β are the estimate of $g(x)$ and its derivative. The minimization with respect to (α, β) is not affected by the observation i that $|X_i - x| > h$ because any non-local observations are eliminated in the above local objective function by the kernel function $K_i(x)$. The reasonability of the local ρ-objective function is that the non-local observations do not carry sufficient information about $g(x)$ under this given criterion.

The existing kernel-based method uses the local objective function (2.3) with various different functions ρ. If the density function of the error, f, is known, an ideal choice of the loss function $\rho(\cdot)$ is $-\log f(\cdot)$. It is appealing to consider an objective function

$$\sum_{i=1}^{n} \log f\{Y_i - \alpha - \beta(X_i - x)\} K_i(x). \tag{2.4}$$

The maximizer of α is then the estimator of $g(x)$ and the local maximum likelihood estimator (MLE) of $g(x)$ then has the desired optimal properties. However, the objective function (2.4) is useless since f is unknown. This paper proposes an iterative algorithm to achieve the maximization of (2.4) with respect to α and β when f is unknown. Specifically, we replace the unknown density function $f(\cdot)$ in (2.4) with an estimated one derived in the previous step and we estimate the density function $f(\cdot)$ by the Nadaraya-Watson kernel, given by $f_n(w) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_{\hbar}\left(Y_i - g(X_i) - w\right)$, where $\mathcal{K}$ is a nonnegative symmetric kernel function with support on $[-1, 1]$ and $\hbar$ is the bandwidth. It is known that under regularity conditions, we have $\sup_w \|f_n(w) - f(w)\| \to 0$. Hence, the estimator of $g(x)$ based on the following local maximum likelihood,

$$\sum_{i=1}^{n} \log f_n\{Y_i - \alpha - \beta(X_i - x)\} K_h(X_i - x). \tag{2.5}$$

is closed to the one based on (2.4). Since f is unknown, g is identifiable up to a location shift, so we set $g^{(k)}(X_n) = 0$ for all $k \geq 0$ for notational and computational convenience. The iterative algorithm is formally presented as follows.

- **Step 0.** Choose an initial estimator of $g(\cdot)$, denoted as $g^{(0)}(\cdot)$. For example, this can be the usual local linear estimator of $g(\cdot)$.

- **Iteration Step from $k-1$ to k.**

  (a) For every given $w = Y_1 - g^{(k-1)}(X_1), \cdots, Y_n - g^{(k-1)}(X_n)$, we estimate $f(w)$ by

$$f^{(k)}(w) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_{\hbar}\left(Y_i - g^{(k-1)}(X_i) - w\right) \tag{2.6}$$

  (b) For every given $x = X_1, \cdots, X_{n-1}$, we estimate $\alpha = g(x)$ and $\beta = \dot{g}(x)$ by the estimated local maximum likelihood,

$$\sum_{i=1}^{n} \log[f^{(k)}\{Y_i - \alpha - \beta(X_i - x)\}]K_h(X_i - x) \tag{2.7}$$

  with respect to α and β. Let $\widehat{\alpha}(x)$ and $\widehat{\beta}(x)$ be the solutions of α and β. Then $g^{(k)}(X_i) = \widehat{\alpha}(X_i)$ for $i = 1, \cdots, n-1$.

- Repeat this iteration procedure untill convergence.

- For every x in the range of X, the estimates of $g(x)$ and $\dot{g}(x)$, denoted by $\widehat{g}(x)$ and $\widehat{\dot{g}}(x)$, respectively, are obtained by maximizing the (2.7) for α and β by replacing $f^{(k)}$ with their estimators defined as the limits of the above iteration.

Remark 1. In the actual implementation of the algorithm, at each step, only $g^{(k)}(x)$ for x values of $X_1, \ldots, X_{n-1}$ need to be computed, since only these quantities are used in the next step.

Remark 2. In each step of the above iterations, we use all observations to estimate the density function and then use the estimated density to obtain the estimate of $g(x)$. As a result, the estimate of $g(x)$ as the limit of the iteration, also involves all observations, inside and outside the neighborhood of x. In this sense, it is a genuine global estimation, different from the standard local estimation in (2.3) using a known loss function, where only the local data inside the neighborhood of x is used. In Section 3, the proposed estimator is shown to be semiparametrically efficient (as defined by Bickel et al., 1993).

Remark 3. When f is known, our estimate is reduced to the local maximum likelihood estimator. Substituting (2.6) into (2.7), our estimator can be regarded as a nonparametric version of the maximum profile likelihood estimator. The difference between our estimator and the traditional maximum profile likelihood estimator is that the dimension of our parameter space is infinity.

Remark 4. The derivative of f is required to estimate g. Non-parametric estimators for derivative functions may converge relatively slowly, so a higher-order kernel $\mathcal{K}$ that can be taken from Muller (1984) is needed to ensure sufficiently rapid convergence. In the simulation studies and the real example, we use the fourth-order kernel for $\mathcal{K}$ to satisfy the theoretical conditions for consistency and asymptotical normality of the estimators. The consistency and asymptotical normality of the estimator are studied in Section 3.

3 Large sample properties

We now establish the uniform consistency and asymptotic normality of the proposed likelihood estimator. Without loss of generality, we assume the support of X is $[0,1]$. Other regularity conditions are stated in the Appendix A. For the following discussion, we fix $\widehat{g}(0) = g(0) = 0$ in theory. The uniform consistency of $\widehat{g}(\cdot)$ is presented in Theorem 1.

Theorem 1 *Under Conditions 1-6 stated in Appendix A, we have*

$$\sup_{0<x<1} |\widehat{g}(x) - g(x)| \to 0 \ \textit{in probability.}$$

To express explicitly the asymptotic expression of the estimator $\widehat{g}(x) - g(x)$, we introduce the following notation. Denote $\mu_i = \int x^i K(x)dx$, $\nu_i = \int x^i K^2(x)dx$, $\tau = E\left(\frac{\dot{f}^2(\varepsilon_i)}{f^2(\varepsilon_i)}\right)$, $f(x)$ is the density function of ε and $p(x)$ is the density function of X.

Theorem 2 *Under Conditions 1-6 stated in Appendix A, for* $0 < x < 1$, $\widehat{g}(x) - g(x)$ *satisfies the following Fredholm integral equation,*

$$\begin{aligned}\widehat{g}(x) - g(x) = & \int_0^1 [\widehat{g}(u) - g(u)]p(u)du \\ & + \frac{1}{\sqrt{nh}}\sigma(x)\varphi + \frac{h^2}{2}\ddot{g}(x)\mu_2 + o_p\{h^2 + 1/\sqrt{nh}\}, \end{aligned} \tag{3.1}$$

where $\sigma^2(x) = \frac{\nu_0}{\tau p(x)}$, φ *is a standard normal random variable and* $\ddot{g}(x) = d^2 g(x)/dx^2$.

Let $\mathcal{A}$ be the linear operator satisfying $\mathcal{A}(\phi) = \int_0^1 p(u)\phi(u)du$ for any function ϕ, and I be the identity operator. Then (3.1) can be written as

$$(I - \mathcal{A})\left(\widehat{g} - g\right)(x) = \frac{1}{\sqrt{nh}}\sigma(x)\varphi + \frac{h^2}{2}\ddot{g}(x)\mu_2 + o_p\{h^2 + 1/\sqrt{nh}\}. \tag{3.2}$$

Since $(I-\mathcal{A})^{-1}$ is linear, (3.2) can be expressed as

$$\widehat{g}(x) - g(x) = (nh)^{-1/2}(I-\mathcal{A})^{-1}(\sigma)(x)\varphi \\ +\tfrac{1}{2}h^2\mu_2(I-\mathcal{A})^{-1}(\ddot{g})(x) + o_p\{h^2+(nh)^{-1/2}\}.$$

The order of the asymptotic bias of $\widehat{g}(x)-g(x)$ is h^2 and the order of the asymptotic covariance is $(nh)^{-1}$. As a consequence, the theoretical optimal bandwidth is of the order $n^{-1/5}$. It follows from (3.2) in straightforward fashion that

Corollary 1 *Under Conditions 1-6 in Appendix A, for* $0<x<1$,

$$(nh)^{1/2}[\widehat{g}(x) - g(x) - \frac{1}{2}h^2\mu_2(I-\mathcal{A})^{-1}(\ddot{g})(x)] \xrightarrow{\mathcal{D}} N(0,\Sigma(x)), \tag{3.3}$$

where $\Sigma(x) = [(I-\mathcal{A})^{-1}(\sigma)(x)]^2$.

To evaluate the optimality of the proposed approach, we provide a justification via semiparametric efficiency in the following theorem. Define $\mathcal{D}_0 = \{\phi(\cdot)$: ϕ has a continuous second derivative on $[0,1]$, $\int_0^1 \phi(x)dx = 0\}$. Then,

Theorem 3 *Assume Conditions 1-6 in Appendix A hold. If* $nh^4 \to 0$, $nh^2/(\log n)^2 \to \infty$ *and* $n\hbar^5 h/(\log n)^2 \to \infty$, *then for any* $\phi \in \mathcal{D}_0$,

$$\int_0^1 \phi(x)\widehat{g}(x)dx \;\; \textit{is an efficient estimator of} \;\; \int_0^1 \phi(x)g(x)dx.$$

Estimation of the parameter $\int_0^1 \phi(x)g(x)dx$ at the rate $n^{-1/2}$ requires undersmoothing with $h = o(n^{-1/4})$ to ensure that the bias is of order $o(n^{-1/2})$. The necessity of undersmoothing to obtain root-n-consistent estimation is standard in nonparametric regression; see, for example, Carroll et al. (1997) and Hastie & Tibshirani (1990).

4 Applications to partial linear models

Partial linear models have attracted much attention due to their flexibility to combine traditional linear models with nonparametric regression models. The partial linear model is written as:

$$Y = g(X) + Z'\gamma + \varepsilon, \tag{4.1}$$

where X is a $1-$dimensional random variable, $g(\cdot)$ is an unknown measurable function, Z is a $d-$dimensional random vector, γ is an unknown parameter vector of $d-$dimension, and ε is the error with unknown distribution f. For identification, the covariate Z excludes the intercept term. There exist many methods to estimate γ and g when f is unknown. Two primary approaches are the penalized least squares (Engle et al., 1986; Green et al., 1985; Shiau et al., 1986; Wahba, 1984a, 1984b) and the local least squares (Speckman,1988; Härdle, 2000) methods. Several efficient estimators for γ also have been considered in the literature. Particularly, Heckman (1986, 1988), Chen (1988), Speckman (1988) and Cuzick (1992) have shown that γ can be efficiently estimated in the model (4.1) if g is smooth but unknown when the error distribution is normal. Cuzick (1992) discusses a general theory suggesting that efficient estimation of γ should be possible even when the error distribution is unknown. Ma et al. (2006) derive a semiparametrically efficient estimator of γ using a constant weight function. However, the efficiency is obtained only for γ, not for the regression function g. Here, we extend the idea described in Section 2 to estimate γ and g when the error distribution is unknown. Our estimator is semiparametrically efficient not only for γ, but also for the regression function $g(\cdot)$.

We set $g^{(k)}(X_n) = 0$ for all $k \geq 0$ for identification purposes. An algorithm to estimate γ and g, by taking advantage of the method described in Section 2 is:

- Choose an initial estimator of $g(\cdot)$ and γ, denoted as $g^{(0)}(\cdot)$ and $\gamma^{(0)}$. For example, these can be the usual local least squares estimators for $g(\cdot)$ and the least squares estimators for γ.

- **Iteration Step from $k-1$ to k.**

  (a) For every given $w = Y_1 - g^{(k-1)}(X_1) - Z_1'\gamma^{(k-1)}, \cdots, Y_n - g^{(k-1)}(X_n) - Z_n'\gamma^{(k-1)}$, we estimate $f(w)$ by

$$f^{(k)}(w) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_{\hbar}\{Y_i - g^{(k-1)}(X_i) - Z_i'\gamma^{(k-1)} - w\}.$$

  (b) For every given $x = X_1, \cdots, X_{n-1}$, we estimate $\alpha = g(x)$ and $\beta = \dot{g}(x)$ by the estimated local maximum likelihood,

$$\sum_{i=1}^{n} \log[f^{(k)}\{Y_i - Z_i'\gamma^{(k-1)} - \alpha - \beta(X_i - x)\}] K_h(X_i - x) \tag{4.2}$$

  with respect to α and β. Let $\widehat{\alpha}(x)$ and $\widehat{\beta}(x)$ be the solutions of α and β. Then $g^{(k)}(X_i) = \widehat{\alpha}(X_i)$ for $i = 1, \cdots, n-1$.

(c) We estimate γ by maximizing

$$\sum_{i=1}^{n} \log f^{(k)}\{Y_i - g^{(k)}(X_i) - Z_i'\gamma\}$$

with respect to γ.

- Repeat this iteration procedure till convergence.
- For every x in the range of X, the estimates of $g(x)$ and $\dot{g}(x)$, denoted by $\widehat{g}(x)$ and $\widehat{\dot{g}}(x)$, respectively, are obtained by maximizing the (4.2) for α and β and replacing $f^{(k)}$ and $\gamma^{(k-1)}$ with their estimators defined as the limits of the above iteration.

Following similar arguments to those for Theorems 1–3, we have the following Theorems 4–7. The proof of Theorem 4 is analogous to that of Theorem 1. The proofs of Theorems 5 to 7 are provided in Supplementary Material.

Theorem 4 *Under Conditions 1–6 stated in Appendix B, we have*

$$\|\widehat{\gamma} - \gamma\| \to 0 \text{ and } \sup_{0<x<1} |\widehat{g}(x) - g(x)| \to 0, \text{ in probability.}$$

Theorem 5 *Under Conditions 1–6 stated in Appendix B, if* $nh^4 \to 0$, $nh^2/(\log n)^2 \to \infty$ *and* $n\hbar^5 h/(\log n)^2 \to \infty$ *as* $n \to \infty$, *we have*

$$n^{1/2}(\widehat{\gamma} - \gamma) \to N\left(0, [Cov(Z) - Cov\{b_1(X)\}]^{-1}\tau^{-1}\right),$$

in distribution, where $b_1(x) = E(Z \mid X = x) - E(Z)$ *and* $\tau = E\left(\frac{\dot{f}^2(\varepsilon_i)}{f^2(\varepsilon_i)}\right)$.

Remark 5. Estimation of the parameter γ at the rate $n^{-1/2}$ again requires undersmoothing with $h = o(n^{-1/4})$ to ensure that the bias of $\widehat{\gamma}$ is of order $o_p(n^{-1/2})$.

Remark 6. If X and Z are independent, we have $b_1(x) = E(Z \mid X = x) - E(Z) \equiv 0$. Then, under the conditions in Theorem 5,

$$n^{1/2}(\widehat{\gamma} - \gamma_0) \to N\left(0, \{\mathrm{Cov}(Z)\}^{-1}\tau^{-1}\right).$$

The asymptotic distribution of $\widehat{\gamma}$ is the same as if $g(\cdot)$ were known. This implies that the estimations of γ and $g(\cdot)$ are independent if X and Z are independent. The underlying reason for the phenomenon lies in the additive structure for g and γ so that the model (4.1) can be written as $Y = Z'\gamma + \tilde{\varepsilon}$, where $\tilde{\varepsilon} = g(X) + \varepsilon$, and thus $\tilde{\varepsilon}$ and Z are independent if X and Z are independent.

Theorem 6 *Under Conditions 1–6 stated in Appendix B, for* $0 < x < 1$, *we have the Fredholm integral equation,*

$$\widehat{g}(x) - g(x) = \int \{\widehat{g}(u) - g(u)\}p(u)du + \frac{h^2\ddot{g}(x)\mu_2}{2}$$
$$- \frac{1}{n\tau p(x)} \sum_{i=1}^{n} \frac{\dot{f}(\varepsilon_i)}{f(\varepsilon_i)} K_h(X_i - x) + o_p\left(h^2 + (nh)^{-1/2}\right). \qquad (4.3)$$

Then, using the same notations as those in Corollary 1, we give the asymptotical normality for $\widehat{g}(x)$ in the following Corollary,

Corollary 2 *Under Conditions 1–6 in Appendix B, for* $0 < x < 1$,

$$(nh)^{1/2}[\widehat{g}(x) - g(x) - \frac{1}{2}h^2\mu_2(I - \mathcal{A})^{-1}(\ddot{g})(x)] \xrightarrow{\mathcal{D}} N(0, \Sigma(x)), \qquad (4.4)$$

where $\Sigma(x) = [(I - \mathcal{A})^{-1}(\sigma)(x)]^2$.

This indicates that the order of the asymptotic biases of $\widehat{g}(x)$ is h^2 and the order of the asymptotic variances is $(nh)^{-1}$. As a consequence, the theoretical optimal bandwidth $O(n^{-1/5})$ in nonparametric problem can be achieved. In addition, from Corollary 1 and Corollary 2, we observe that the asymptotic bias and the variance of $\widehat{g}(\cdot)$ are the same as if γ were known. This occurs because $\widehat{\gamma}$ converges to the true value at the rate $O_p(n^{-1/2})$ (see Theorem 5), while $\widehat{g}(x)$ converges to the true value at the rate $O_p(n^{-2/5})$, so that the uncertainty of $\widehat{g}(\cdot)$ due to estimating γ is ignorable.

Theorem 5 shows that $\widehat{\gamma}$ is a $n^{1/2}$-consistent and asymptotically normal estimator of γ. Theorem 7 below shows that $\widehat{\gamma}$ is also an efficient estimator of γ. For any function $\phi(x) = \{\phi_1', \phi_2(x)\}'$, where $\phi_2(\cdot)$ has a continuous second derivative on $[0, 1]$ and satisfies $\int_0^1 \phi_2(x)dx = 0$. Let $\phi_1'\widehat{\gamma} + \int_0^1 \phi_2(x)\widehat{g}(x)dx$ be the estimator of $\phi_1'\gamma + \int_0^1 \phi_2(x)g(x)dx$, then

Theorem 7 *Under the conditions of Theorem 5,*

$$\phi_1'\widehat{\gamma} + \int_0^1 \phi_2(x)\widehat{g}(x)dx \ \textit{ is an efficient estimator of } \ \phi_1'\gamma + \int_0^1 \phi_2(x)g(x)dx.$$

Theorem 7 implies that by taking $\phi_2(x) \equiv 0$, $\widehat{\gamma}$ is an efficient estimator of γ; and that by taking $\phi_1 = 0$, $\int_0^1 \phi_2(x)\widehat{g}(x)dx$ is an efficient estimator of $\int_0^1 \phi_2(x)g(x)dx$.

5 Numerical studies

In this section we investigate the performance of the proposed maximum likelihood curve estimator by comparing our method with the local least squares (LLS) estimator. The performance of the estimator $\widehat{g}(\cdot)$ is assessed via $bias = \left[\frac{1}{n_{grid}} \sum_{i=1}^{n_{grid}} \{E\widehat{g}(x_i) - g(x_i)\}^2\right]^{1/2}$, $sd = \left[\frac{1}{n_{grid}} \sum_{i=1}^{n_{grid}} E\{\widehat{g}(x_i) - E\widehat{g}(x_i)\}^2\right]^{1/2}$ and RMSE $= [bias^2 + sd^2]^{1/2}$, where x_i $(i = 1, \ldots, n_{grid})$ are the grid points where the function $g(\cdot)$ is estimated and $E\widehat{g}(x_i)$ is approximated by its sample mean based on N simulated data sets. In the following examples, the Epanechnikov kernel will be used, and $n_{grid} = 200$. The results presented below are based on $N = 300$ replications and $n = 300$ sample size.

We first consider the simple nonparametric regression model $Y = g(X) + \varepsilon$, where $X \sim U[-1, 1]$, $g(x) = x^3$, and Model 1: $\varepsilon \sim U[-0.4, 0.4]$; or Model 2: $\varepsilon \sim 0.5N(-0.4, 0.1^2) + 0.5N(0.4, 0.1^2)$.

Within each of the two sampling conditions, we apply the proposed estimator and the LLS estimator to the generated data. Table 1 shows the performance results of the two methods under Models 1 and 2. From Table 1, we see that for each model, given a bandwidth h, the biases of the proposed estimator and the LLS estimator are comparable, but the proposed estimator has a much smaller standard deviation than the LLS estimator. The empirical efficiency of the LLS estimator relative to our method is 46.85% on average for Model 1 and 8.99% on average for Model 2. This implies that our method only requires about 46.85% of the sample size that is required by the LLS estimator for Model 1, and about 8.99% of the sample size that is required by the LLS estimator for Model 2. As a result of the smaller standard deviation, the proposed estimator has better performance than the LLS estimator in each case according to the RMSE. The improvement of the proposed method, compared to the LLS estimator, is much more significant for Model 2 than that for Model 1. This is likely because of the fact that the error term in Model 2 deviates further from the normal random variable distribution than that in Model 1.

Table 1 : The simulation results for Models 1 and 2

h	Proposed			LLS		
	bias	sd	RMSE	bias	sd	RMSE
			Model 1			
0.1	0.0055	0.0358	0.0363	0.0089	0.0628	0.0634
0.2	0.0131	0.0256	0.0287*	0.0140	0.0378	0.0403*
0.3	0.0258	0.0217	0.0337	0.0252	0.0325	0.0412
0.4	0.0410	0.0219	0.0464	0.0390	0.0297	0.0490
0.5	0.0562	0.0212	0.0601	0.0536	0.0281	0.0605
			Model 2			
0.1	0.0050	0.0294	0.0298	0.0080	0.0897	0.0901
0.2	0.0127	0.0209	0.0244*	0.0112	0.0679	0.0688
0.3	0.0238	0.0170	0.0293	0.0220	0.0586	0.0626*
0.4	0.0381	0.0151	0.0409	0.0359	0.0535	0.0644
0.5	0.0528	0.0145	0.0547	0.0508	0.0501	0.0714

Note: * stands for the minimum RMSE

Now we consider the partial linear nonparametric regression model $Y = g(X) + Z'\boldsymbol{\beta} + \varepsilon$, where $X \sim 0.5N(-0.5, 0.5^2) + 0.5N(0.5; 0.5^2)$, $g(x) = \sin(x)$, $Z = (Z_1, Z_2)'$, $Z_1 \sim N(0, 0.5^2)$, $Z_2 \sim U(0, 1)$, $\boldsymbol{\beta} = (\beta_1, \beta_2)' = (0.5, 1)'$, and Model 3: $\varepsilon \sim U[-0.4, 0.4]$; or Model 4: $\varepsilon \sim 0.5N(-0.4, 0.1^2) + 0.5N(0.4, 0.1^2)$.

Table 2 provides the bias, empirical standard deviations and root mean square errors (RMSE) of the nonparametric estimators of the function $g(\cdot)$ using the proposed method and the LLS method for Models 3 and 4, respectively. We obtain similar conclusions from Table 2 as we did from the results in Table 1. Furthermore, the amount of data used by the proposed estimator is more than that used by the LLS estimator, so the optimal bandwidth of the proposed estimator is smaller than that for the LLS estimator.

Table 2 : The simulation results of Models 3 and 4 for the nonparameter

h	Proposed			LLS		
	bias	sd	RMSE	bias	sd	RMSE
	Model 3					
0.2	0.0028	0.0264	0.0266	0.0022	0.0444	0.0444
0.3	0.0050	0.0252	0.0257	0.0037	0.0387	0.0389
0.4	0.0077	0.0211	0.0225	0.0067	0.0357	0.0363
0.5	0.0115	0.0186	0.0219*	0.0105	0.0340	0.0356*
0.6	0.0158	0.0165	0.0228	0.0146	0.0330	0.0361
	Model 4					
0.2	0.0028	0.0208	0.0210	0.0034	0.0779	0.0780
0.3	0.0045	0.0165	0.0171	0.0037	0.0678	0.0679
0.4	0.0077	0.0141	0.0161*	0.0066	0.0623	0.0626
0.5	0.0114	0.0123	0.0168	0.0106	0.0590	0.0599
0.6	0.0156	0.0111	0.0191	0.0148	0.0570	0.0589*
0.7	0.0199	0.0103	0.0224	0.0191	0.0560	0.0592

Table 3 : The simulation results of Models 3 and 4 for the regression coefficients

		Proposed			LLS		
	h	bias	sd	RMSE	bias	sd	RMSE
				Model 3			
β_1	0.2	-0.0002	0.0159	0.0159	-0.0004	0.0286	0.0286
	0.3	0.0005	0.0198	0.0198	-0.0003	0.0282	0.0282
	0.4	0.0006	0.0192	0.0192	-0.0002	0.0280	0.0280
	0.5	0.0010	0.0191	0.0191	-0.0002	0.0279	0.0279
	0.6	0.0010	0.0191	0.0191	-0.0001	0.0278	0.0278
β_2	0.2	0.0017	0.0227	0.0228	0.0021	0.0462	0.0462
	0.3	0.0040	0.0358	0.0361	0.0026	0.0462	0.0462
	0.4	0.0024	0.0357	0.0358	0.0028	0.0464	0.0464
	0.5	0.0024	0.0350	0.0350	0.0029	0.0467	0.0468
	0.6	0.0019	0.0350	0.0351	0.0029	0.0471	0.0472
				Model 4			
β_1	0.2	0.0001	0.0115	0.0115	-0.0003	0.0487	0.0487
	0.3	-0.0001	0.0112	0.0112	-0.0002	0.0482	0.0482
	0.4	-0.0001	0.0112	0.0112	-0.0001	0.0478	0.0478
	0.5	-0.0003	0.0114	0.0114	0.0001	0.0476	0.0476
	0.6	-0.0003	0.0115	0.0115	0.0001	0.0475	0.0475
	0.7	-0.0003	0.0116	0.0116	0.0001	0.0473	0.0473
β_2	0.2	0.0001	0.0193	0.0193	-0.0023	0.0780	0.0781
	0.3	0.0001	0.0208	0.0208	-0.0031	0.0779	0.0779
	0.4	-0.0000	0.0201	0.0201	-0.0030	0.0779	0.0779
	0.5	-0.0001	0.0193	0.0193	-0.0029	0.0781	0.0782
	0.6	-0.0003	0.0192	0.0192	-0.0026	0.0784	0.0785
	0.7	-0.0001	0.0193	0.0193	-0.0022	0.0791	0.0792

Table 3 provides the bias, sd and RMSEs of the regression coefficients estimators $\widehat{\boldsymbol{\beta}}$ using the proposed method and the LLS method for Models 3 and 4, respectively. With a given bandwidth, both methods are unbiased but the proposed estimator has a much smaller standard deviation than the LLS estimator. Therefore, the proposed estimator has better performance than the local estimator in terms of the RMSE. Table 3 also suggests that the proposed method prefers a smaller bandwidth.In addition, the improvement of the proposed method is much more significant for Model

4 than that for Model 3 when compared to the LLS estimator. Again, this is due to the error in Model 4 deviating further from the normal distribution.

6 Example

To study how prestige depends on education levels and income, Fox (1997) collected a data set from the 1971 Canada Census that contains prestige (Pineo-Porter prestige score for occupation from a social survey in the mid-1960s), income (average income of incumbents, in dollars) and education (average education time of occupation incumbents, in years) of 102 individuals. Fox (1997) pointed out that linear regression analysis of prestige on income and education is not proper. He suggested the use of nonparametric methods to obtain prior knowledge, regarding the relationship between prestige and education and income.

Here, we investigate the dependence of prestige on income and education by using the nonparametric regression model (1.1). We estimate the nonparametric regression function $g(\cdot)$ using both the LLS and the proposed methods. The bandwidth of each method is chosen by leave-one-out cross validation (CV(1)). The variances are estimated by the resampling method proposed by Jin, Ying and Wei (2001) with 500 bootstrap samples. The choice of the 500 sample size is determined by monitoring the stability of the standard errors.

Example 1: $Y = g(X) + \varepsilon$, where Y is prestige and X is income. The plots of the prediction errors (PE) vs the bandwidth are displayed in Figure 1, suggesting that $h = 13500$ and $h = 10100$ are optimal bandwidths for the LLS and the proposed methods, respectively. With the corresponding optimal bandwidths, the resulting estimates of $g(\cdot)$ and the pointwise 95% bootstrap confidence interval are presented in Figure 2.

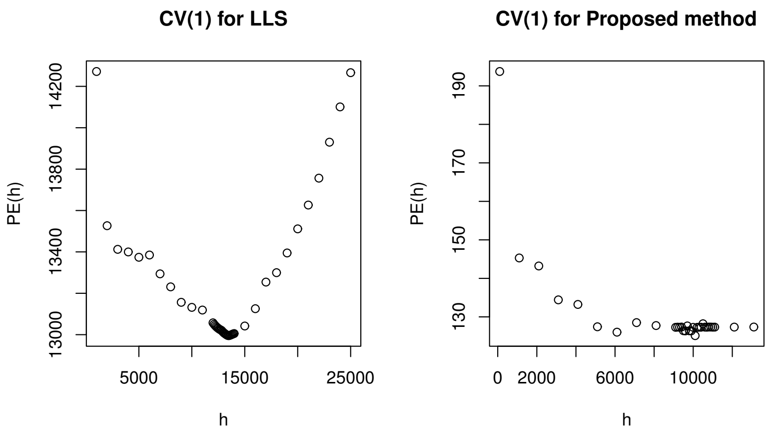

Figure 1: The plot of PE vs h for the LLS estimator and the proposed method when analyzing the relationship between prestige and income.

From Figure 2, we see that the estimated functions based on the LLS estimator and the proposed method are similar. Both methods suggest that the function $g(\cdot)$ is highly non-constant and non-linear. However, the confidence band of the LLS estimator is much wider than that of the proposed estimator, suggesting that the proposed method is more efficient than the LLS estimator. This is consistent with our simulation findings and the theory described earlier stating that the proposed method is semiparametrically efficient. Moreover, Figure 2 shows that prestige increases linearly when income increases from \$0 to \$14000, but prestige only increases slightly when income increases beyond \$14000.

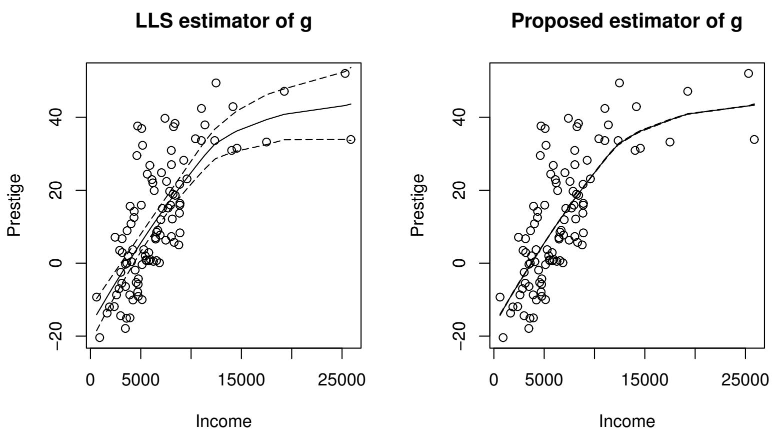

Figure 2: The estimated regression functions (solid-line) and their 95% confidence limits (dashed-line) by the LLS estimator and the proposed estimator when analyzing the relationship between prestige and income.

Example 2: $Y = g(X) + \varepsilon$, where Y is prestige and X is education. The plots of the prediction errors (PE) vs the bandwidth are displayed in Figure 3 for Example 2, suggesting that bandwidths $h = 5.5$ and $h = 3.9$ should be chosen for the LLS and the proposed methods, respectively. The resulting estimates of $g(\cdot)$ and the pointwise 95% bootstrap confidence intervals are presented in Figure 4. Figure 4 again shows that the estimates based on the two methods are similar, but the proposed method is more efficient than the LLS estimator. Figure 4 indicates that prestige basically increases linearly with education. Combining the results in Figure 2 and Figure 4, we hence consider the following partial linear model

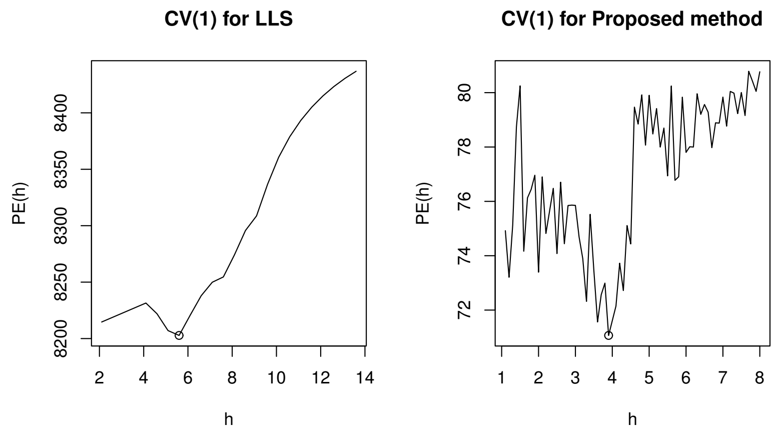

Figure 3: The plot of PE vs h for the LLS estimator and the proposed method when analyzing the relationship between prestige and education.

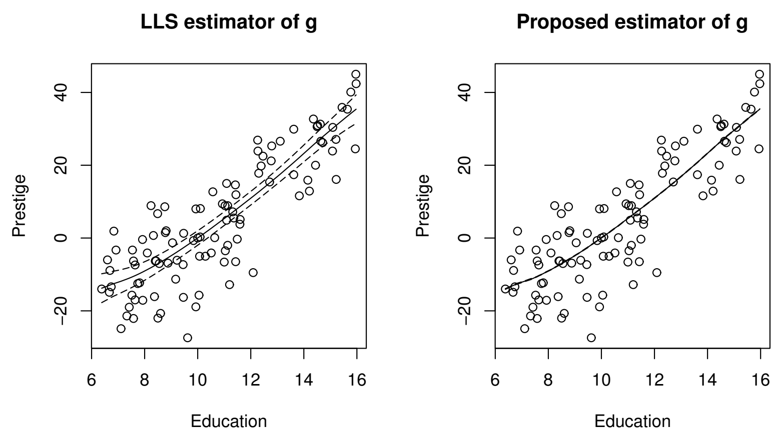

Figure 4: The estimated regression functions (solid lines) and their 95% confidence limits (dashed lines) by the LLS estimator and the proposed estimator when analyzing the relationship between prestige and education.

Example 3: $Y = g(X) + \gamma Z + \varepsilon$, where Y is prestige, X is the income and Z is the education. The plots of the prediction errors (PE) vs the bandwidth are displayed in Figure 5 for Example 3, suggesting that $h = 13600$ and $h = 7200$ are optimal bandwidths for the LLS and the proposed methods, respectively. The estimates for γ are given in Table 4, and the estimates for $g(\cdot)$ and the pointwise 95% bootstrap confidence interval are shown in Figure 6. Table 4 indicates that high education can significantly increase prestige. Figure 6 suggests that prestige increases linearly with income increasing from \$0 to \$11000, and is stable when income varies from \$11000 to \$18000, then increases linearly when income is larger than \$18000. Once again, Table 4 and Figure 6 show that the two estimators give similar estimators for the regression function and coefficient, but the proposed method is more efficient than the LLS estimator for the parameters and non-parametric functions.

Table 4 : The parameter estimator of the partial linear model for the prestige data.

Proposed		LLS	
Est.	sd	Est.	sd
4.9719	0.0021	4.9714	0.4750

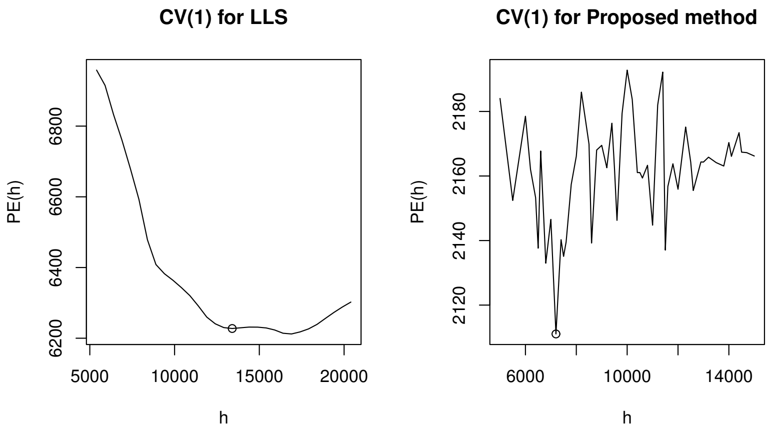

Figure 5: The plot of PE vs h for the LLS estimator and the proposed method when analyzing the relationship between prestige, income and education.

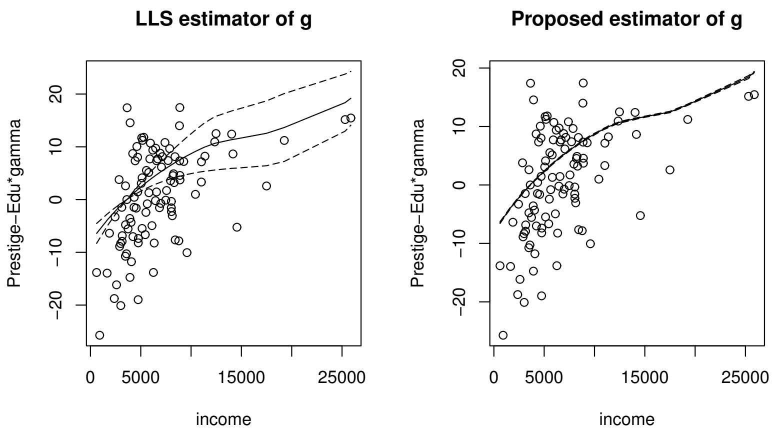

Figure 6: The estimated regression functions (solid lines) and their 95% confidence limits (dashed lines) by the LLS estimator and the proposed estimator when analyzing the relationship between prestige, income and education.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). Efficient and adaptive estimation for semiparametric models. John Hopkins University Press, Baltimore.

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

Carter, C. K. and Eagleson, G. K. (1992). A comparison of variance estimators in nonparametric regression. *J. Roy. Statist. Soc. Ser. B. (Methodological)* **54**, 773-780.

Carter, C. K., Eagleson, G. K. and Silverman, B. W. (1992). A comparison of the Reinsch and Speckman splines, *Biometrika*, **79**, 81-91.

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

Chen, K., Guo, S., Sun, L. and Wang, J. (2010). Global partial likelihood for nonparametric proportional hazards models. *J. Am. Statist. Assoc.* **105**, 750–760.

Chen, K., Lin, H. Z. and Zhou, Y. (2012). Efficient estimation for the Cox model with varying coefficients. *Biometrika*, **2**, 379-392.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, **74**, 829-836.

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

Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.*, **81**, 310-320.

Eubank, R. L. (1988). Spline smoothing and nonparametric regression. Marcel Dekker, New York, second edition.

Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and its Applications. London: Chapman and Hall.

Fox, J. (1997). Applied Regression, Linear Models, and Related Methods. Sage.

Gasser, T. and Muller, H. G. (1979). kernel estimation of regression functions. In: *Lecture Notes in Mathematics,* **757**, 23-68. New York: Springer.

Green, P., Jennison, C. and Seheult, A. (1985). Analysis of field experiments by least squares smoothing. *J. Roy. Statist. Soc. Ser. B.*, **47**, 299-315.

Green, P. J. and Silverman, B. W. (1994) Nonparametric regression and generalized linear models. London: Chapman and Hall.

Hall, P. and Carroll, R. J. (1989). Variance function estimation in regression: The effect of estimating the mean. *J. Roy. Statist. Soc. Ser. B. (Methodological)*, **51**, 3-14.

Hall, P. and Marron, J. S. (1990). On variance estimation in nonparametric regression. *Biometrika*, **77**, 415-419.

Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *J. Multiv. Analysis*, **29**, 163-179.

Härdle, W., Liang, H. and Gao, J. (2000). Partially linear models. Springer-Verlag.

Hastie, T. J. and Tibshirani, R. (1990). Generalized additive models, London: Chapman and Hall.

Heckman, N. E. (1986). Spline smoothing in a partly linear model. *J. Roy. Statist. Soc. Ser. B.*, **48**, 244-248.

Heckman, N. E. (1988). Minimax estimates in a semiparametric model. *J. Amer. Statist. Assoc.*, **83**, 1090-1096.

Horowitz, J. L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica.*, **64**, 103-137.

Jin, Z., Ying, Z. and Wei, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika*, **88**, 381-390.

Ma, Y. and Carroll, R. J. (2006). Locally efficient estimators for semiparametric models with measurement error. *J. Amer. Statist. Assoc.*, **101**, 1465-1474.

Muller, H. G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *Ann. Statist.*, **12**, 766-774.

Muller, H. G. and Stadtmuller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.*, **15**, 610-625.

Nadaraya, E. A. (1964). On estimating regression. *Theory of probability and its applications*, **9**, 141-142.

Neumann, M. H. (1994). Fully data-driven nonparametric variance estimators. *Statistics: A Journal of Theoretical and Applied Statistics*, **25**, 189-212.

Pollard, D. (1990). Empirical Processes: Theory and Applications. Institute of Mathematical Statistics, Hayward, California.

Shiau, J-J., Wahba, G. and Johnson, D. R. (1986). Partial spline models for the inclusion of tropopause and frontal boundary information in otherwise smooth two- and three- dimensional objective analysis. *Journal of Atmospheric and Oceanic Technology*, **3**, 714-725.

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B*, **50**, 413-436.

Stone, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.*, **5**, 595-645.

Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B. (Methodological)*, **40**, 364-372.

Wahba, G. (1984a). Cross validated spline methods for the estimation of multivariate functions from data on functionals. *In Statistics: An Appraisal, Proc. 50th Anniversary Conf. Iowa State Statistical Laboratory (Edited by David, H. A. and David, H. T.)*, 205-235, Iowa State University Press, Ames.

Wahba, G. (1984b). Partial spline models for the semiparametric estimation of functions of several variables. *In Statistical Analysis for Time Series, Japan-Us Joint Seminar* , 312-329, Tokyo: Institute of Statistical Mathematics.

Wang, F. T. and Scott, D. W. (1994). The L_1 method for robust nonparametric regression. *J. Amer. Statist. Assoc.*, **89**, 65-76.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian of Journal of Statistics, Series A*, **26**, 359-372.

Appendix A: Conditions and Proofs of Theorems 1 to 3

Denote the density function of X to be $p(x)$,

$$\dot{f}(x) = \partial f(x)/\partial x, \;\; \ddot{f}(x) = \partial^2 f(x)/\partial x^2,$$

$$f(w;\alpha) = Ef(w - g(X_i) + \alpha(X_i)), \;\; \dot{f}(w;\alpha) = E\dot{f}(w - g(X_i) + \alpha(X_i)),$$

$$\ddot{f}(w;\alpha) = E\ddot{f}(w - g(X_i) + \alpha(X_i)), \;\; \rho_1(\alpha;x) = E\left\{\frac{\dot{f}(g(x)+\varepsilon_i-\alpha(x);\alpha)}{f(g(x)+\varepsilon_i-\alpha(x);\alpha)}\right\},$$

$$\rho_2(\alpha;x) = E\left(\frac{\ddot{f}(g(x)+\varepsilon_i-\alpha(x);\alpha)}{f(g(x)+\varepsilon_i-\alpha(x);\alpha)}\right), \;\; u(\alpha;x) = \rho_1(\alpha;x)p(x),$$

$$\mathcal{C}_0 = \{\alpha(x) : x \in [0,1], \alpha(x) \text{ is continuous on } [0,1], \alpha(0) = 0 \text{ and } \sup_x |\alpha(x)| \le C\},$$

where C is a constant.

Condition:

1. The kernel functions $K(\cdot)$ and $\mathcal{K}(\cdot)$ is symmetric function with a compact support $[-1,1]$. $K(\cdot)$ has bounded derivative and $\mathcal{K}(\cdot)$ has second derivative. We assume that K and $\mathcal{K}$ have orders of 2 and r, respectively. The kernel function of order r satisfies: $\int_{-1}^{1} u^j \mathcal{K}(u)du$ is 1 when $j=0$, is 0 when $1 \le j \le r-1$, and is not zero when $j=r$.

2. X is bounded with compact support $[0,1]$. X and ε are independent.

3. The density functions $f(\cdot)$ and $p(\cdot)$ are positive, and have continuous $(r+2)$th and second derivatives on corresponding supports, respectively. $f(x) \to 0$ as $|x| \to \infty$.

4. The function g has continuous second derivative on $[0,1]$ and $g \in \mathcal{C}_0$.

5. $E\left\{ \frac{\int_0^1 \dot{f}(\varepsilon_i - h(x) + h(z))p(z)dz}{\int_0^1 f(\epsilon_i - h(x) + h(z))p(z)dz} \right\} = 0$ if and only if $h(x) \equiv 0$ over $x \in [0,1]$.

6. $n\hbar^5/\log(n) \to \infty$, $h^2 = o(\hbar)$, $\hbar^r = o(h^2)$, $h^2 \log(n) \to 0$ and $nh/(\log(n))^2 \to \infty$, as $n \to \infty$.

The assumption on h and $\hbar$ can be satisfied, for example, if K is a second-order kernel, $\mathcal{K}$ is fourth-order kernel functions with $h \propto n^{-1/5}$ and $\hbar \propto n^{-a}$, $1/10 < a < 1/5$. The second-order and fourth-order kernel functions can be taken from Muller (1984). Since the derivative of f is required to estimate g, and derivative functional converge relatively slowly, the higher-order kernel for $\mathcal{K}$ is needed to insure sufficiently rapid convergence. Condition 5 is used to to insure the identification of the model.

Lemma 1 *Denote* $\dot{\mathcal{K}}_\hbar(w) = \dot{\mathcal{K}}(w/\hbar)/\hbar^2$, $\ddot{\mathcal{K}}_\hbar(w) = \ddot{\mathcal{K}}(w/\hbar)/\hbar^3$,

$$\widehat{f}(w;\alpha) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_\hbar\left(Y_i - \alpha(X_i) - w\right), \quad \widehat{\dot{f}}(w;\alpha) = -\frac{1}{n}\sum_{i=1}^{n} \dot{\mathcal{K}}_\hbar\left(Y_i - \alpha(X_i) - w\right),$$

and $\widehat{\ddot{f}}(w;\alpha) = \frac{1}{n}\sum_{i=1}^{n} \ddot{\mathcal{K}}_\hbar\left(Y_i - \alpha(X_i) - w\right)$. *Under Conditions 1-6, we have*

$$\sup_{w\in(-\infty,\infty),\alpha\in\mathcal{C}_0} |\widehat{f}(w;\alpha) - f(w;\alpha)| = O_p\left(\hbar^r + \log(n)^{1/2}(n\hbar)^{-1/2}\right), \tag{6.1}$$

$$\sup_{w\in(-\infty,\infty),\alpha\in\mathcal{C}_0} |\widehat{\dot{f}}(w;\alpha) - \dot{f}(w;\alpha)| = O_p\left(\hbar^r + \log(n)^{1/2}(n\hbar^3)^{-1/2}\right), \tag{6.2}$$

$$\sup_{w\in(-\infty,\infty),\alpha\in\mathcal{C}_0} |\widehat{\ddot{f}}(w;\alpha) - \ddot{f}(w;\alpha)| = O_p\left(\hbar^r + \log(n)^{1/2}(n\hbar^5)^{-1/2}\right). \tag{6.3}$$

Proof of Lemma 1. See Supplementary Material.

Lemma 2. Let $\theta = (\alpha,\beta)'$. Denote $m(X_i;x) = g(X_i) - g(x) - \dot{g}(x)(X_i - x)$, $\mathcal{Y}_i(\theta,x) = Y_i - \alpha(x) - \beta(x)(X_i - x)$, $\widehat{\dot{f}}_i(\theta,x) = \widehat{\dot{f}}(\mathcal{Y}_i(\theta,x);\alpha)$, $f_i(\theta,x) = f(\mathcal{Y}_i(\theta,x);\alpha)$, $\widehat{f}_i(\theta,x) =$

$\widehat{f}(\mathcal{Y}_i(\theta,x);\alpha)$, $\dot{f}_i(\theta,x) = \dot{f}(\mathcal{Y}_i(\theta,x);\alpha)$ and $W_i(x) = (1,(X_i - x)/h)'$. Under Conditions 1-6 stated in Appendix A, we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\dot{f}}_i(\theta_0,x)}{\widehat{f}_i(\theta_0,x)}K_h(X_i-x)W_i(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{\dot{f}_i(\theta_0,x)}{f_i(\theta_0,x)}K_h(X_i-x)W_i(x)$$
$$-\frac{1}{n^2}\sum_{i\neq j}^{n}V(\varepsilon_j,\varepsilon_i,X_i;x) + O_p(\hbar^r) + o_p(h^2+n^{-1/2}),$$

where $\theta_0 = (g,\dot{g})'$ is the true value of θ.

Proof of Lemma 2. See Supplementary Material.

Proof of Theorem 1.

For any vector functions $\theta(\cdot) = (\alpha(\cdot),\beta(\cdot))'$, set

$$U(\theta;x) = \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\dot{f}}(Y_i-\alpha(x)-\beta(x)(X_i-x);\alpha)}{\widehat{f}(Y_i-\alpha(x)-\beta(x)(X_i-x);\alpha)}K_h\left(X_i-x\right)W_i(x). \tag{6.4}$$

Under model (1.1), we have

$$U(\theta;x) = u(\alpha;x)(1,0)' + o_p(1).$$

Suppose there exist two functions g and $g+h$ in $\mathcal{C}_0$ such that $u(g;x)\equiv 0$ and $u(g+h;x)\equiv 0$. Then,

$$0 = \left[u(g+h;x)-u(g;x)\right]/p(x) = E\left\{\frac{\int_0^1 \dot{f}(\varepsilon_i-h(x)+h(y))p(y)dy}{\int_0^1 f(\varepsilon_i-h(x)+h(y))p(y)dy}\right\}. \tag{6.5}$$

By $U(\widehat{\theta};x) = 0$ and $u(g;x)=0$ as well as Condition 5, g is unique root of $u(\alpha;x)=0$ in $\alpha\in\mathcal{C}_0$.

Define

$$\mathcal{B}_n = \left\{\alpha : \|\alpha\|\le D, \|\alpha(x_1)-\alpha(x_2)\|\le d[|x_1-x_2|+b_n], x_1,x_2\in[0,1]\right\},$$

for some constants $D>0$ and $d>0$, where $b_n = h + \log(n)^{1/2}(nh)^{-1/2} + \hbar^r + \log(n)^{1/2}(n\hbar^5)^{-1/2}$. To show the uniform consistency of $\widehat{\theta}$, it suffices to prove the following (i)-(iii):

(i) For each continuous function vector α and any bounded function vector β, $\sup_{0\le x\le 1}$ $\|U(\alpha,\beta;x)-u(\alpha;x)\| = o_p(1)$.

(ii) $\sup_{0\le x\le 1,\alpha\in\mathcal{B}_n,\beta\in\mathcal{R}}\|U(\alpha,\beta;x)-u(\alpha;x)\|=o_p(1)$, where $\mathcal{R}$ is the set of functions on $[0,1]$, which are bounded uniformly.

(iii) $P\{\widehat{g}\in\mathcal{B}_n\cap\mathcal{C}_0\}\to 1$.

Once (i)-(iii) are established, using the similar idea to the Arzela-Ascoli theorem and (iii), we can show that for any subsequence of $\{\widehat{g}\}$, there exists a further convergent subsequence $\widehat{g}_n$ such that uniformly in $w\in[0,1]$, $\widehat{g}_n(x)\to g^*(x)$ in probability. It is easily seen that $g^*\in\mathcal{C}_0$. Note that $u(g^*;x)=U(\widehat{g},\widehat{\dot{g}};x)-\Big[U(\widehat{g},\widehat{\dot{g}};x)-u(\widehat{g};x)\Big]-[u(\widehat{g};x)-u(g^*;x)]$, and $U(\widehat{g},\widehat{\dot{g}};x)=0$, where $\widehat{\dot{g}}(\cdot)$ is the estimator of $\dot{g}(\cdot)$. It follows from (ii) that $u(g^*;x)=0$. Since $u(\alpha;x)=0$ has a unique root at g, we have $g^*=g$, which yields the uniform consistency of $\widehat{g}$.

Proof of (i). Using Lemma 2 and noting that $\frac{1}{n^2}\sum_{i,j}^n V(\varepsilon_j,\varepsilon_i,X_i;x)=O_p(n^{-1/2})$ by the central limit theorem of U-statistics, we have

$$\begin{aligned}&\sup_x\left|U(\theta;x)-u(\alpha;x)\right|\\&\qquad=\sup_x\Big|\frac{1}{n}\sum_{i=1}^n\frac{\dot{f}_i(\theta,x)}{f_i(\theta,x)}K_h(X_i-x)W_i(x)-u(\alpha;x)\Big|+o_p(1).\end{aligned}\tag{6.6}$$

It can be shown (e.g., Horowitz, 1996) that the first term of the right side of (6.6) is $O_p\left\{h^2+\log(n)^{1/2}(nh)^{-1/2}\right\}$. Hence, (i) follows.

Proof of (ii). For any given continuous function $\eta(x)$, we have

$$\begin{aligned}&U(\theta;x)-U(\eta,\beta;x)\\&\quad=\frac{1}{n}\sum_{i=1}^n\frac{\{\widehat{\dot{f}}_i(\theta,x)-\widehat{\dot{f}}_i(\eta,\beta,x)\}\widehat{f}_i(\eta,\beta,x)}{\widehat{f}_i(\theta,x)\widehat{f}_i(\eta,\beta,x)}K_h(X_i-x)W_i(x)\\&\quad\quad+\frac{1}{n}\sum_{i=1}^n\frac{\{\widehat{f}_i(\eta,\beta,x)-\widehat{f}_i(\theta,x)\}\widehat{\dot{f}}_i(\eta,\beta,x)}{\widehat{f}_i(\theta,x)\widehat{f}_i(\eta,\beta,x)}K_h(X_i-x)W_i(x).\end{aligned}\tag{6.7}$$

It follows from Condition 1 that

$$\begin{aligned}&\sup_{x,\alpha\in\mathcal{B}_n,\beta\in\mathcal{R}}\left|\widehat{\dot{f}}_i(\theta,x)-\widehat{\dot{f}}_i(\eta,\beta,x)\right|\\&=\sup_{x,\alpha\in\mathcal{B}_n,\beta\in\mathcal{R}}\left|\frac{1}{n}\sum_{j=1}^n\dot{\mathcal{K}}_{\hbar}(Y_j-\alpha(X_j)-Y_i+\alpha(x)+\beta(x)(X_i-x))\right.\\&\qquad\qquad\left.-\frac{1}{n}\sum_{j=1}^n\dot{\mathcal{K}}_{\hbar}(Y_j-\eta(X_j)-Y_i+\eta(x)+\beta(x)(X_i-x))\right|\\&\le M\sup_{x,\alpha\in\mathcal{B}_n,\beta\in\mathcal{R}}\left|\eta(x)-\alpha(x)\right|.\end{aligned}\tag{6.8}$$

where M is a constant and may be different in different place. Similarly,

$$\sup_{x,\alpha\in\mathcal{B}_n,\beta\in\mathcal{R}}\left|\widehat{f}_i(\eta,\beta,x)-\widehat{f}_i(\theta,x)\right|\le M\sup_{x,\alpha\in\mathcal{B}_n,\beta\in\mathcal{R}}\left|\eta(x)-\alpha(x)\right|.$$

Then, it follows that

$$\sup_{x,\alpha\in\mathcal{B}_n,\beta\in\mathcal{R}}\left|U(\theta;x)-U(\eta,\beta;x)\right|\le M\sup_{x,\alpha\in\mathcal{B}_n,\beta\in\mathcal{R}}\left|\eta(x)-\alpha(x)\right|. \tag{6.9}$$

For any $\varepsilon>0$, let $\mathcal{H}$ denote the finite set of points $mD/M, m=0,\pm 1,\cdots,M$, where M is an integer such that $D/M<\varepsilon/3$ ($\mathcal{H}$ is an ε-net for the linear interval $[-D,D]$). Choose k and n large enough so that $d(1/k+b_n)<\varepsilon/3$, let $\mathcal{Q}$ consist of those elements of the space of continuous functions on $[0,1]$ that are linear on each subinterval $I_i=[(i-1)/k,i/k], i=1,\cdots,k$, and assume values in $\mathcal{H}$ at points $i/k, i=0,1,\cdots,k$. The size of the set $\mathcal{Q}$ is $(2M+1)^{k+1}=O((1/\varepsilon)^{1/\varepsilon})$. For any $\alpha(\cdot)\in\mathcal{B}_n$, there exists an element $\eta(\cdot)$ of $\mathcal{Q}$ such that $|\eta(i/k)-\alpha(i/k)|<\varepsilon/3, i=0,1,\cdots,k$. Since $\eta(\cdot)$ is linear on each subinterval I_i, it follows that $sup_{x\in[0,1],\alpha\in\mathcal{B}_n}|\eta(x)-\alpha(x)|<\varepsilon$. That is , for all n sufficiently large, $\mathcal{Q}$ is an ε-net for $\mathcal{B}_n$. As a result, $logN(\varepsilon,\mathcal{B}_n,\|\cdot\|_\infty)\le O(\frac{1}{\varepsilon}log(\frac{1}{\varepsilon}))=o(n)$, where $N(\varepsilon,\mathcal{B}_n,\|\cdot\|_\infty)$ is the covering number with respect to the norm $\|\cdot\|_\infty$ of the class $\mathcal{B}_n$.

Hence, with (i) and (6.9), as well as $sup_{x,\alpha\in\mathcal{B}_n}\left|u(\alpha;x)-u(\eta;x)\right|\le Msup_{x,\alpha\in\mathcal{B}_n}\left|\eta(x)-\alpha(x)\right|$, by the uniform law of large numbers (Pollard, 1990, p.39), (ii) is proved.

Proof of (iii). Note that $U(\widehat{\theta};x)=0$ for any x, using the first component of $U(\widehat{\theta};x)=0$, we get that

$$\begin{aligned}
0=U_1(\widehat{\theta};x_1)-U_1(\widehat{\theta};x_2)&=\frac{1}{n}\sum_{i=1}^n\frac{\widehat{\dot{f}}_i(\widehat{\theta},x_1)-\widehat{\dot{f}}_i(\widehat{\theta},x_2)}{\widehat{f}_i(\widehat{\theta},x_1)}K_h(X_i-x_1)\\
&\quad-\frac{1}{n}\sum_{i=1}^n\frac{\left\{\widehat{f}_i(\widehat{\theta},x_1)-\widehat{f}_i(\widehat{\theta},x_2)\right\}\widehat{\dot{f}}_i(\widehat{\theta},x_2)}{\widehat{f}_i(\widehat{\theta},x_1)\widehat{f}_i(\widehat{\theta},x_2)}K_h(X_i-x_1)\\
&\quad+\frac{1}{n}\sum_{i=1}^n\frac{\widehat{\dot{f}}_i(\widehat{\theta},x_2)}{\widehat{f}_i(\widehat{\theta},x_2)}\left\{K_h(X_i-x_1)-K_h(X_i-x_2)\right\}\\
&=G_1(\widehat{\theta},x_0,x_1,x_2)\left\{\widehat{g}(x_1)-\widehat{g}(x_2)\right\}+G_2(\widehat{\theta},x_0,x_1,x_2)(x_2-x_1)+O_p(b_n),
\end{aligned}$$

where x_0 is between x_1 and x_2, and may be different in different functions,

$$G_1(\theta, x_0, x_1, x_2) = -p(x_1)E\left[\left\{\frac{\ddot{f}_i(\theta, x_0)}{f_i(\theta, x_1)} - \frac{\dot{f}_i(\theta, x_0)\dot{f}_i(\theta, x_2)}{f_i(\theta, x_1)f_i(\theta, x_2)}\right\} \mid X_i = x_1\right],$$

$$\begin{aligned}
&G_2(\theta, x_0, x_1, x_2) \\
&= \beta(x_2)\left(G_1(\theta, x_0, x_1, x_2) + p(x_0)E\left[\left\{\frac{\ddot{f}_i(\theta, x_2)}{f_i(\theta, x_2)} - \frac{\dot{f}_i^2(\theta, x_2)}{f_i^2(\theta, x_2)}\right\} \mid X_i = x_0\right]\right) \\
&\qquad - \dot{p}(x_0)E\left[\dot{f}_i(\theta, x_2)/f_i(\theta, x_2) \mid X_i = x_0\right].
\end{aligned}$$

Hence, *(iii)* follows.

Proof of Theorem 2. Denote $d_{n,1} = \sup_x |\widehat{g}(x) - g(x)|$ and $d_{n,2} = h \sup_x |\widehat{\dot{g}}(x) - \dot{g}(x)|$. Write

$$\begin{aligned}
&U(\widehat{\theta}; x) - U(\theta_0; x) \\
&= \{\widehat{g}(x) - g(x)\}\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\dot{f}}_i^2(\theta_0, x) - \widehat{\ddot{f}}_i(\theta_0, x)\widehat{f}_i(\theta_0, x)}{\widehat{f}_i(\widehat{\theta}, x)\widehat{f}_i(\theta_0, x)}K_h(X_i - x)W_i(x) \\
&+ \{\widehat{\dot{g}}(x) - \dot{g}(x)\}\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\dot{f}}_i^2(\theta_0, x) - \widehat{\ddot{f}}_i(\theta_0, x)\widehat{f}_i(\theta_0, x)}{\widehat{f}_i(\widehat{\theta}, x)\widehat{f}_i(\theta_0, x)}(X_i - x)K_h(X_i - x)W_i(x) \\
&- \int\{\widehat{g}(u) - g(u)\}p(u)du\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\dot{f}}_i^2(\theta_0, x) - \widehat{\ddot{f}}_i(\theta_0, x)\widehat{f}_i(\theta_0, x)}{\widehat{f}_i(\widehat{\theta}, x)\widehat{f}_i(\theta_0, x)}K_h(X_i - x)W_i(x) \\
&\qquad + O_p\left\{d_{n,1}^2 + d_{n,2}^2\right\}.
\end{aligned}$$

By Lemma 2, we get

$$\begin{aligned}
U(\widehat{\theta}; x) - U(\theta_0; x) &= \{\widehat{g}(x) - g(x)\}E\left\{\frac{\dot{f}^2(\varepsilon_i) - \ddot{f}(\varepsilon_i)f(\varepsilon_i)}{f^2(\varepsilon_i)}\right\}p(x)\,(1,0)' \\
&+ h\{\widehat{\dot{g}}(x) - \dot{g}(x)\}E\left\{\frac{\dot{f}^2(\varepsilon_i) - \ddot{f}(\varepsilon_i)f(\varepsilon_i)}{f^2(\varepsilon_i)}\right\}p(x)\mu_2(0,1)' \\
&- \int\{\widehat{g}(u) - g(u)\}p(u)duE\left\{\frac{\dot{f}^2(\varepsilon_i) - \ddot{f}(\varepsilon_i)f(\varepsilon_i)}{f^2(\varepsilon_i)}\right\}p(x)(1,0)' \\
&+ O_p\left\{d_{n,1}^2 + d_{n,2}^2 + (\hbar^r + (n\hbar^5)^{-1/2}(\log n)^{1/2})(d_{n,1} + d_{n,2})\right\} + o_p(h^2), \quad (6.10)
\end{aligned}$$

On the other hand, by Lemma 2,

$$U(\theta_0;x) = \frac{h^2\ddot{g}(x)\mu_2}{2}E\left\{\frac{\ddot{f}(\varepsilon)f(\varepsilon)-\dot{f}^2(\varepsilon)}{f^2(\varepsilon)}\right\}p(x)(1,0)'$$
$$+\frac{1}{n}\sum_{i=1}^{n}\frac{\dot{f}(\varepsilon_i)}{f(\varepsilon_i)}K_h(X_i-x)W_i(x)-\frac{1}{n^2}\sum_{i\neq j}^{n}V(\varepsilon_j,\varepsilon_i,X_i;x)+O_p\left(\hbar^r\right)$$
$$+o_p(h^2+n^{-1/2}). \tag{6.11}$$

Denote $V_1(\varepsilon_j,\varepsilon_i,X_i;x)$ to be the first component of $V(\varepsilon_j,\varepsilon_i,X_i;x)$. Substituting (6.11) into (6.10), noting that $U(\widehat{\theta};x)=0$, we obtain that

$$\widehat{g}(x)-g(x)-\int\{\widehat{g}(u)-g(u)\}p(u)du$$
$$=\frac{h^2\ddot{g}(x)\mu_2}{2}-\frac{1}{n\tau p(x)}\sum_{i=1}^{n}\frac{\dot{f}(\varepsilon_i)}{f(\varepsilon_i)}K_h(X_i-x)+\frac{1}{n^2\tau p(x)}\sum_{i\neq j}^{n}V_1(\varepsilon_j,\varepsilon_i,X_i;x)$$
$$+O_p\left\{\hbar^r+d_{n,1}^2+d_{n,2}^2+(n\hbar^5)^{-1/2}(\log n)^{1/2}(d_{n,1}+d_{n,2})\right\}+o_p(h^2+n^{-1/2}), \tag{6.12}$$

and

$$d_{n,2}=O_p\left(\log(n)^{1/2}(nh)^{-1/2}+\hbar^r\right)+o_p(h^2). \tag{6.13}$$

Using the central limit theorem of U-statistics, we have

$$\frac{1}{n^2\tau p(x)}\sum_{i\neq j}^{n}V_1(\varepsilon_j,\varepsilon_i,X_i;x)=O_p(n^{-1/2}).$$

Taking supremum norm on both side of (6.12), and using Condition 6, we get $d_{n,1}=O_p\{h^2+(nh)^{-1/2}\}$. Then Theorem 2 follows from (6.12).

Proof of Theorem 3.

Using Lemma 2 and the conditions on the bandwidths, we have

$$\widehat{g}(x)-g(x)-\int\{\widehat{g}(u)-g(u)\}p(u)du-\frac{h^2\ddot{g}(x)\mu_2}{2}+\frac{1}{n\tau p(x)}\sum_{i=1}^{n}\frac{\dot{f}(\varepsilon_i)}{f(\varepsilon_i)}K_h(X_i-x)$$
$$=\frac{1}{n^2\tau p(x)}\sum_{i\neq j}^{n}V_1(\varepsilon_j,\varepsilon_i,X_i;x)+O_p(\hbar^r)+o_p(n^{-1/2}+h^2).$$

Given $\phi\in\mathcal{D}_0$, we firstly consider the asymptotic variance of $\int_0^1\widehat{g}(x)\phi(x)dx$. Note that $nh^4 \to 0$, we have

$$\int_0^1 \{\widehat{g}(x) - g(x)\}\phi(x)dx = \frac{1}{n^2\tau}\sum_{i\neq j}^{n}\left\{\int_0^1 V_1(\varepsilon_j, \varepsilon_i, X_i; x)\phi(x)/p(x)dx \right.$$
$$\left. -\frac{n^2}{n(n-1)}\int_0^1 \frac{\dot{f}(\varepsilon_i)}{f(\varepsilon_i)}K_h(X_i - x)\phi(x)/p(x)dx\right\} + o_p(n^{-1/2}). \tag{6.14}$$

On the other hand, note that

$$E\left[\int_0^1 \{V_1(\varepsilon_j, \varepsilon_i, X_i; x) + V_1(\varepsilon_i, \varepsilon_j, X_j; x)\}\,\phi(x)/p(x)dx \mid \varepsilon_j, X_j\right] = O(h^2),$$

and $E\{V_1(\varepsilon_j, \varepsilon_i, X_i; x)\} = 0$ when $i \neq j$. Using the central limit of U-statistic, we have $\frac{1}{n^2}\sum_{i\neq j}^{n}\int_0^1 V_1(\varepsilon_j, \varepsilon_i, X_i; x)\phi(x)/p(x)dx = o_p(n^{-1/2})$. Coupling with $nh^4 \to 0$, we obtain,

$$n^{1/2}\int_0^1 \{\widehat{g}(x) - g(x)\}\phi(x)dx$$
$$= \frac{1}{\sqrt{n}\tau}\sum_{i=1}^{n}\int_0^1\left\{-\frac{\dot{f}(\varepsilon_i)}{f(\varepsilon_i)}K_h(X_i - x)\right\}\phi(x)/p(x)dx + o_p(1)$$
$$\to N(\left\{0, \tilde{\sigma}^2\right\}, \tag{6.15}$$

where $\tilde{\sigma}^2 = \frac{1}{\tau}\int_0^1 \frac{\phi^2(x)}{p(x)}dx$. To show the asymptotic efficiency of $\int_0^1 \widehat{g}(x)\phi(x)dx$, we consider a parametric submodel with paramter γ, such that $g(x;\gamma) = g(x) + \frac{\gamma\phi(x)}{\tau p(x)}$, where γ is an unknown parameter. Obviously, $\gamma_0 = 0$ is the true value of γ. The score function for this parameter submodel is

$$\frac{1}{\sqrt{n}\tau}\sum_{i=1}^{n}\left\{\frac{\dot{f}(\varepsilon_i - \frac{\gamma\phi(X_i)}{\tau p(X_i)})}{f(\varepsilon_i - \frac{\gamma\phi(X_i)}{\tau p(X_i)})}\right\}\frac{\phi(X_i)}{p(X_i)}, \tag{6.16}$$

where the variance of the score at γ_0 is $\tilde{\sigma}^2 \equiv \frac{1}{\tau}\int_0^1 \frac{\phi^2(x)}{p(x)}dx$. Thus, the maximum likelihood estimator of γ, denoted as $\tilde{\gamma}$, satisfies

$$\sqrt{n}(\tilde{\gamma} - \gamma_0) \to N(0, \tilde{\sigma}^{-2}). \tag{6.17}$$

Observe that

$$\int_0^1 \{g(x;\tilde{\gamma}) - g(x;\gamma_0)\}\phi(x)dx = (\tilde{\gamma} - \gamma_0)\int_0^1 \frac{\phi^2(x)}{\tau p(x)}dx \tag{6.18}$$

Thus, it follows from (6.17) and (6.18) that

$$\sqrt{n}\int_0^1 \{g(x;\tilde{\gamma}) - g(x;\gamma_0)\}\phi(x)dx \to N(0, \tilde{\sigma}^2). \tag{6.19}$$

This, together with (6.14) and (6.15), shows that the asymptotic variance of $\int_0^1 \widehat{g}(x)\phi(x)dx$ is the same as that of $\int_0^1 g(x;\tilde{\gamma})\phi(x)dx$. As explained in Bickel, Klaassen, Ritov and Wellner(1993, p.46), $\int_0^1 \widehat{g}(x)\phi(x)dx$ is asymptotically efficient for the estimate of $\int_0^1 g(x)\phi(x)dx$. The proof of Theorem 3 is complete.

Appendix B: Conditions for Theorems 4 to 7

Denote $m(X_i; x) = g(X_i) - g(x) - \dot{g}(x)(X_i - x)$,

$$f(w; \alpha, \gamma) = E\left[f\left\{\alpha(X_i) - g(X_i) + Z_i'(\gamma - \gamma_0) + w\right\}\right],$$

$$\widehat{f}(w; \alpha, \gamma) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_\hbar\left\{Y_i - \alpha(X_i) - Z_i^T\gamma - w\right\},$$

$$\mathcal{Y}_i(\theta, \gamma, x) = Y_i - Z_i^T\gamma - \alpha(x) - \beta(x)(X_i - x),\ \mathcal{Y}_i(\alpha, \gamma) = Y_i - Z_i^T\gamma - \alpha(X_i),$$

$$f_i(\theta, \gamma, x) = f(\mathcal{Y}_i(\theta, \gamma, x); \alpha, \gamma),\ \widehat{f}_i(\theta, \gamma, x) = \widehat{f}(\mathcal{Y}_i(\theta, \gamma, x); \alpha, \gamma),$$

$$\dot{f}_i(\theta, \gamma, x) = \dot{f}(\mathcal{Y}_i(\theta, \gamma, x); \alpha, \gamma), \widehat{\dot{f}}_i(\theta, \gamma, x) = \widehat{\dot{f}}(\mathcal{Y}_i(\theta, \gamma, x); \alpha, \gamma),$$

$$f_i(\alpha, \gamma) = f(\mathcal{Y}_i(\alpha, \gamma); \alpha, \gamma),\ \dot{f}_i(\alpha, \gamma) = \dot{f}(\mathcal{Y}_i(\alpha, \gamma); \alpha, \gamma),$$

$$\widehat{f}_i(\alpha, \gamma) = \widehat{f}(\mathcal{Y}_i(\alpha, \gamma); \alpha, \gamma),\ \widehat{\dot{f}}_i(\alpha, \gamma) = \widehat{\dot{f}}(\mathcal{Y}_i(\alpha, \gamma); \alpha, \gamma).$$

Condition:

1. The kernel functions $K(\cdot)$ and $\mathcal{K}(\cdot)$ is symmetric function with a compact support $[-1, 1]$. $K(\cdot)$ has bounded derivative and $\mathcal{K}(\cdot)$ has second derivative. We assume that K and $\mathcal{K}$ have orders of 2 and r, respectively. The kernel function of order r satisfies: $\int_{-1}^{1} u^j \mathcal{K}(u)du$ is 1 when $j = 0$, is 0 when $1 \leq j \leq r - 1$, and is not zero when $j = r$.

2. X and Z are bounded with compact support $[0, 1]$. The covariate (X, Z) and ε are independent. Denote the joint density function of (X, Z) to be $p(x, z)$. The density function $p(x, z)$ is positive and has continuous second partial derivative on corresponding supports.

3. Denote the density function of X to be $p(x)$. The density functions $f(\cdot)$ and $p(\cdot)$ are positive, and have continuous $(r + 2)$th and second derivatives on corresponding supports, respectively. $f(x) \to 0$ as $|x| \to \infty$.

4. The function g has continuous second derivative on $[0, 1]$.

5. $E\left\{\frac{\int_0^1\int_0^1 \dot{f}(\varepsilon_i - h(x) - Z_i'w + h(u) + v'w)p(u,v)dudv}{\int_0^1\int_0^1 f(\epsilon_i - h(x) - Z_i'w + h(u) + v'w)p(u,v)dudv}\right\} = 0$ if and only if $h(x) \equiv 0$ over $x \in [0, 1]$.

6. $n\hbar^5/\log(n) \to \infty$, $h^2 = o(\hbar)$, $\hbar^r = o(h^2)$, $h^2\log(n) \to 0$ and $nh/(\log(n))^2 \to \infty$, as $n \to \infty$.