Semi-parametric estimation of partially linear single-index models

Yingcun Xia\textsuperscript{a, *}, Wolfgang Härdle\textsuperscript{b}

\textsuperscript{a}Department of Statistics and Applied Probability, National University of Singapore, Singapore
\textsuperscript{b}CASE—Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Berlin, Germany

Received 9 June 2004
Available online 30 January 2006

Abstract

One of the most difficult problems in applications of semi-parametric partially linear single-index models (PLSIM) is the choice of pilot estimators and complexity parameters which may result in radically different estimators. Pilot estimators are often assumed to be root-$n$ consistent, although they are not given in a constructible way. Complexity parameters, such as a smoothing bandwidth are constrained to a certain speed, which is rarely determinable in practical situations.

In this paper, efficient, constructible and practicable estimators of PLSIMs are designed with applications to time series. The proposed technique answers two questions from Carroll et al. [Generalized partially linear single-index models, J. Amer. Statist. Assoc. 92 (1997) 477–489]: no root-$n$ pilot estimator for the single-index part of the model is needed and complexity parameters can be selected at the optimal smoothing rate. The asymptotic distribution is derived and the corresponding algorithm is easily implemented. Examples from real data sets (credit-scoring and environmental statistics) illustrate the technique and the proposed methodology of minimum average variance estimation (MAVE).

© 2005 Elsevier Inc. All rights reserved.

AMS 1991 subject classification: 62G07; 62G20; 62F10

Keywords: Asymptotic distribution; Generalized partially linear model; Local linear smoother; Optimal consistency rate; Single-index model

1. Introduction

Although the presence of nonlinearities in statistical data analysis is often modelled with non- and semi-parametric methods, there are still few noncritical semi-parametric techniques. One
argument that has been advanced is that—despite a reduction in dimensionality—the practical estimation still depends heavily on pilot estimators and complexity parameters. Another argument against finely tuned semi-parametrics is that mathematical tools for inferential decisions and software implementations are either missing or not readily accessible. The purpose of this paper is to show that such critiques may be refuted even for the very flexible class of partially linear single-index models (PLSIM):

\[
y = \beta_0^T Z + g(\theta_0^T X) + \varepsilon, \quad (1.1)
\]

where \(E(\varepsilon|X, Z) = 0\) almost surely, \(\beta_0\) and \(\theta_0\) are unknown parameters, \(g(\cdot)\) is an unknown link function. The PLSIM (1.1) was first analyzed by Carroll et al. [4] and contains the single-index models \((\beta_0 \equiv 0)\), generalized partially linear models \((X\) one-dimensional and \(y\) observed logits) and generalized linear models \((\beta_0 \equiv 0\) and \(g\) known). The advantage of the PLSIM lies in its generality and its flexibility. The wide spread application of PLSIMs though is somewhat obstructed by the facts described above: necessity of pilot estimators for \(\theta_0\) and complexity parameters such as bandwidths (to estimate the link function \(g\)). In this paper, we further assume \(||\theta_0|| = 1\) and that the first entry of \(\theta_0\) is positive for model identification; see [26].

The issue of the order of magnitude of the complexity parameter was addressed in [4, Eq. (18), p. 483]. The convenience of a root-\(n\) pilot estimator for \(\theta_0\) was employed in [7] but was found to severely influence the final estimate. In practical application, these two important questions will be addressed in this paper: we will show that a simple multi-dimensional kernel estimator suffices to ensure root-\(n\) consistency of the parametric parts of (1.1) and that no under-smoothing is required for the proposed algorithm.

One motivation of our work comes from credit scoring and the study of nonlinear effects in retail banking. Another motivation comes from the analysis of circulatory and respiratory problems in Hong Kong and the study of the complicated effect of weather conditions on the health problems. Credit scoring methods are designed to assess risk measures for potential borrowers, companies, etc. Typically, the scoring is reduced to a classification or (quadratic) discriminant analysis problem, see [1,10]. The credit data set of Müller and Rönz [17] consists of 6180 cases with 8 metric variables \((x_2, \ldots, x_9)\) and 15 categorical explanatory variables \((x_{10}, \ldots, x_{24})\). Covariates \(x_3, x_4, x_5\) are of the most interest and represent, respectively, duration of payment, amount of loan and customer’s age. The response variable \(y\) was 0 if customers paid their installments without problem, or 1 otherwise. There were 513 cases with a \(y\) value of 1. A scatterplot matrix of the observations \((x_2, x_3, x_4, x_5)\) is given in Fig. 1.

The distribution of the variable \(y\) (black points in Fig. 1) shows a clear nonlinear structure and speaks therefore against a linear discriminant analysis. A logit model

\[
\logit\{P(y = 1|X, Z)\} = \beta_0^T Z + \theta_0^T X \quad (1.2)
\]

(also of linear structure) shows clear nonlinearity in the residuals, see [17]. Here \(X\) denotes the vector of metric variables and \(Z\) the vector of categorical variables. Müller and Rönz [17] therefore applied a partially linear approach as in [22] by replacing one linear term in (1.2) operating on the metric variable \(x_5\) by a nonparametric function \(g(x_5)\) as shown in Fig. 2.

We partition the range of \(x_4\) (or \(x_5\)) into 50 intervals with equal lengths. We cluster the observations with \(x_4\) (or \(x_5\)) in the same interval as one class. We calculate the relative frequencies \(\hat{p}\) for \(y = 1\). In Fig. 2, the variable \(x_4\) (or \(x_5\)) is plotted against the logit \((\hat{p}) = \log(\hat{p}/(1 - \hat{p}))\). Using bootstrap, the nonlinearity was tested and found to be significant. The question of how to integrate further nonlinear influences by the other metric variables was analyzed in [17] at a
Fig. 1. Scatterplots: variables $x_2$ to $x_5$, observations corresponding to $y = 1$ are emphasized in black.

Fig. 2. Marginal dependency. Thicker bullets correspond to more observations in a class. The lines are local linear smoothers.

multi-dimensional kernel regression (e.g. on $(x_4, x_5)$, see Fig. 5.6 in their article) and found to be too difficult to implement due to the high-dimensional kernel smoothing. The technique that we develop here will make it possible to overcome the dimensionality issue and indicate nonlinear influences on the logits via the PLSIM. Based on these discussion, we can consider fit the relation between $Y = \log(p/(1-p))$ and $(X, Z)$ by

$$y = \beta^T \tilde{Z} + g(\theta_1 x_4 + \theta_2 x_5) + \varepsilon,$$

(1.3)

where $\tilde{Z} = (x_2, x_3, x_6, \ldots, x_{24})$.

The other motivation of this research comes from the investigation of the number of daily hospital admissions of patients suffering from the circulatory and respiratory (CR) problems in Hong Kong from 1994–1996. There is a jump in the numbers at the beginning of 1995 due to the additional hospital beds released to accommodate CR patients from the beginning of 1995. We remove this jump by a simple kernel smoothing over time and denote the remaining time series by $y_t$. The pollutants and weather conditions might cause the CR problems. The pollutants include sulfur dioxide ($x_{1t}$, in $\mu g m^{-3}$), nitrogen dioxide ($x_{2t}$, in $\mu g m^{-3}$), respirable suspended particulates ($x_{3t}$, in $\mu g m^{-3}$) and ozone ($x_{4t}$, in $\mu g m^{-3}$), and weather conditions include temperature ($x_{5t}$, in $^\circ C$)
and relative humidity ($x_{6t}$, in %). It is obvious that the higher the levels of air pollutants are, the stronger they tend to cause health problems. Furthermore, simple kernel smoothing suggests that

we can approximate the relations between $y_t$ and the pollution levels linearly; see Fig. 3. However, for the other covariates such as temperature and humidity, the relations are unknown and might be nonlinear. Fig. 3 is simple regression analyses based on kernel smoothing. The relation of $y_t$ with NO$_2$ is almost linear, but the relation of $y_t$ with humidity is nonlinear and hard to explain.

To explore the relation between $y_t$ and air pollutants and weather conditions, we may consider the following model:

$$y_t = \beta^T Z_t + g(\theta^T X_t) + \epsilon_t,$$

where $Z_t$ consists of levels of pollutants and their lagged variables, and $X_t$ consists of weather conditions and their lagged variables.

A nonparametric estimation method can be evaluated in the following four aspects: (1) computational cost, (2) efficiency, (3) conditions necessary for the consistency, and (4) the range of the bandwidth. Most nonparametric methods for the estimation of model (1.1) or its special cases are criticized in one or another aspects. The methods in [4,8,13,23,24] include complicated optimization techniques and no simple algorithm is available up to now. The method of Li [14] required symmetric distribution of the covariate; Härdle and Stoker [9] and Hristache et al. [12] required that $|E g'(\theta_0^T X)|$ is away from 0. If these conditions are violated, their methods cannot obtain useful estimators. The method of Härdle and Stoker [9] and the method of Hristache et al. [11,12] are not asymptotically efficient in the semi-parametric sense. Most of the methods mentioned above require a bandwidth that is much smaller than the data-driven bandwidth in order to allow the estimator of the parameters to achieve root-$n$ consistency, i.e. under-smoothing the link function is needed; see, [4,6,9,11,12] among others. More discussions on the selection of bandwidth for the partially linear model can be found in [15].

In this paper, we present the minimum average variance estimation (MAVE) method that will provide a remedy to these four weak points.

2. Estimation method and asymptotic results

The basic algorithm for estimating the parameters in (1.1) is based on observing that

$$(\beta_0, \theta_0) = \arg \min_{\beta, \theta} E \left[ y - \{\beta^T Z + g(\theta^T X)\} \right]^2$$

(2.1)
subject to $\theta^T \theta = 1$. By conditioning on $\zeta = \theta^T X$, we see that (2.1) equals $E_{\zeta} \sigma^2_{\beta,0}(\zeta)$ where

$$
\sigma^2_{\beta,0}(\zeta) = E \left[ (y - (\beta^T Z + g(\zeta)))^2 | \theta^T X = \zeta \right].
$$

It follows that

$$
E \left[ y - (\beta^T Z + g(\theta^T X)) \right]^2 = E_{\zeta} \sigma^2_{\beta,0}(\theta^T X).
$$

Therefore, minimization (2.1) is equivalent to,

$$
(\beta_0, \theta_0) = \arg \min_{\beta, \theta} E_{\zeta} \sigma^2_{\beta,0}(\zeta)
$$

subject to $\theta^T \theta = 1$. Let $\{(X_i, Z_i, y_i) \in (X_i, Z_i, y_i) i = 1, 2, \ldots, n \}$ be a sample from $(X, Z, y)$. The conditional expectation in (2.2) is now approximated by the sample analog. For $X_i$ close to $x$, we have the following local linear approximation:

$$
y_i - \beta_0^T Z - g(\theta_0^T X_i) \approx y_i - \beta_i^T Z_i - g(\theta_i^T x) - g'(\theta_0^T x) X_{i0} \theta_0,
$$

where $X_{i0} = X_i - x$. Following the idea of local linear smoothing [5], we may estimate $\sigma^2_{\beta,0}(\theta^T x)$ by

$$
\hat{\sigma}^2_{\beta,0}(\theta^T x) = \min_{\beta, \theta} \sum_{i=1}^n \left\{ y_i - \beta_i^T Z_i - a - d X_{i0} \theta \right\}^2 w_{i0}.
$$

Here, $w_{i0} \geq 0$, $i = 1, 2, \ldots, n$, are some weights with $\sum_{i=1}^n w_{i0} = 1$, typically centering at $x$.

Let $X_{i j} = X_i - X_j$. By (2.2) and (2.3), our estimation procedure is to minimize

$$
\frac{1}{n} \sum_{j=1}^n G(\theta^T X_j) I_n(X_j) \sum_{i=1}^n \left\{ y_i - \beta^T Z_i - a_j - d_j X_{i j} \theta \right\}^2 w_{ij}
$$

with respect to $(a_j, d_j)$ and $(\beta, \theta)$, where $G(\cdot)$ is another weight function that controls the contribution of $(X_j, Z_j, y_j)$ to the estimation of $\beta$ and $\theta$. For example, when the model is assumed to be heteroscedastic and $Var(y|X, Z) = V(\theta_0^T X)$, then $G(\cdot) = V(\cdot)$; see [4,7]. $I_n(x)$ is employed here for technical purpose to handle the boundary points. It is given in the next section. See also [7]. For simplicity, we can take $I_n(\cdot) = 1$ in practice. If the noise term is negligible, the true parameter $(\beta_0, \theta_0)$ is a solution to the minimization problem in (2.4) with $w_{ij} \to 0$ if $X_i - X_j \to 0$.

This observation can give us an intuition about why minimizing (2.4) can generate a consistent estimator. We call the estimation procedure the MAVE (conditional) method. The main difference between our method and those in [4,7] is that our method estimate all the parameters and the nonparametric function by minimizing a single common loss function as in (2.4) while those in the two papers estimated nonparametric functions and parameters by minimizing two different loss functions, respectively. This difference is the reason that under-smoothing is not necessary in our method. Similar spirit appeared in [7] for the single-index model, where they minimized a common loss function with respect to both smoothing parameter $h$ and the single-index, but employed another one for the nonparametric function.

Minimizing (2.4) can be decomposed to two typical quadratic programming problems by fixing $(a_j, d_j)$, $j = 1, \ldots, n$ and $(\beta, \theta)$ alternatively. They can be solved easily with simple analytic
expressions. Given $(\beta, \theta)$, we have from minimizing (2.4)

$$
\begin{pmatrix}
(a_j \\
d_j)
\end{pmatrix} = \left\{ \sum_{i=1}^{n} w_{ij} \left( \frac{1}{X_{ij}^T \theta} \right) \left( \frac{1}{X_{ij} \theta} \right)^T \right\}^{-1} \sum_{i=1}^{n} w_{ij} \left( \frac{1}{X_{ij}^T \theta} \right) (y_i - \beta^T Z_i).
$$

(2.5)

Given $(a_j, d_j)$, we have from minimizing (2.4)

$$
\begin{pmatrix}
(\beta) \\
(\theta)
\end{pmatrix} = \left\{ \sum_{j=1}^{n} G(\theta^T X_j) I_n(X_j) \sum_{i=1}^{n} w_{ij} \left( \frac{Z_i}{d_j X_{ij}} \right) \left( \frac{Z_i}{d_j X_{ij}} \right)^T \right\}^{-1} \times \sum_{j=1}^{n} G(\theta^T X_j) I_n(X_j) \sum_{i=1}^{n} w_{ij} \left( \frac{Z_i}{d_j X_{ij}} \right) (y_i - a_j)
$$

(2.6)

and standardize $\theta := sgn_1 \theta/|\theta|$, where $sgn_1$ is the sign of the first entry in $\theta$. Here and later, $|\gamma| = (\gamma^T \gamma)^{1/2}$ for any vector $\gamma$. The minimization in (2.4) can be solved by iterations between (2.5) and (2.6). We call the above procedure the PLSIM algorithm.

The choice of the weights $w_{ij}$ plays an important role in different estimation methods. See [11,12,25]. In this paper, we use two sets of weights. Suppose $H(\cdot)$ and $K(\cdot)$ are a $p$-variate symmetric density function and a univariate symmetric density function, respectively. The first set of weights is $w_{ij} = H_{b,i}(X_j)/\sum_{\ell=1}^{n} H_{b,\ell}(X_j)$, where $H_{b,i}(X_j) = b^{-p} H(X_{ij}/b)$ and $b$ is a bandwidth. This is a multivariate dimensional kernel weight. For this kind of weights, we set $I_n(x) = 1$ if $n^{-1} \sum_{\ell=1}^{n} H_{b,\ell}(x) > c_0$; 0 otherwise for some constant $c_0 > 0$. Iterating (2.5) and (2.6) until convergence, denote the estimators (i.e., the final values) of $\theta$ and $\beta$ by $\tilde{\theta}$ and $\tilde{\beta}$, respectively. Because of the so-called “curse of dimensionality”, the estimation based on this kind of weights is not efficient. However, the multivariate kernel weight can help us to find consistent initial estimators for $\theta_0$ and $\beta_0$. We then use single-index kernel weights $w_{i,j}^{0} = K_{h,i}^0(\theta^T X_j)/\sum_{\ell=1}^{n} K_{h,\ell}^0(\theta^T X_j)$, where $K_{h,i}^0(v) = h^{-1} K((\theta^T X_i - v)/h)$, $h$ is the bandwidth and $\theta$ is the previous estimate of $\theta_0$. Here, we take $I_n(x) = 1$ if $\tilde{f}_\theta(\theta^T X_j) > c_0$; 0 otherwise, where $\tilde{f}_\theta(v) = n^{-1} \sum_{\ell=1}^{n} K_{h,\ell}^0(v)$. Iterating (2.5) and (2.6) until convergence, denote the estimators (i.e., the final values) of $\theta$ and $\beta$ by $\hat{\theta}$ and $\hat{\beta}$, respectively. After obtaining estimates $\hat{\theta}$ and $\hat{\beta}$, we can then estimate $g(v)$ by the solution of $a_j$ in (2.5) with $\theta^T X_j$ replaced by $v$, denote the estimate by $\hat{g}(v)$.

**Lemma 1.** Let $\tilde{\theta}$ and $\tilde{\beta}$ be the estimators based on the multi-dimensional kernel weight. Suppose that (C1)–(C6) in Section 6 hold, $b \to 0$ and $nb^{p+2}/\log n \to \infty$. If we start the estimation procedure with $\theta$ such that $\theta^T \theta_0 \neq 0$, then

$$
\tilde{\theta} - \theta_0 = o_P(1), \quad \tilde{\beta} - \beta_0 = o_P(1).
$$

This lemma suggests that we can obtain consistent initial estimators of $\theta_0$ and $\beta_0$ using multi-dimensional kernel. Although these initial estimators have a slow consistent rate, they suffice for us to obtain root-$n$ consistent estimators eventually. If the design of $X$ is symmetric, the method of Li [14] can also be used to get consistent initial estimators. There are other numerical methods available for the choice of the initial values. See, for example, [2,20]. However, those methods cannot guarantee the consistent properties for the initial estimators. The consistency is needed for the technical purpose to get the following theorem. Let $\mu_\theta(x) = E(X|\theta^T X = \theta^T x)$,
\[ v_0(x) = E(Z|\theta^T X = \theta^T x), \text{and for } k = 0 \text{ and } 2, \]
\[ W_k = E \left\{ G(\theta_0^T X)I(f_{\hat{\theta}}(\theta_0^T X) > c_0) \left( \frac{Z - v_0(X)}{g'(\theta_0^T X)(X - \mu_{\theta_0}(X))} \right) \right\}. \]

**Theorem 1.** Let \((\hat{\beta}, \hat{\theta})\) be the estimators based on the single-index kernel weight starting with \((\beta, \theta) = (\hat{\beta}, \hat{\theta})\). Suppose (C1)–(C6) in Section 6 hold, \(h \sim n^{-\delta} \) with \(1/6 < \delta < 1/4\) and that \(E\{\varepsilon_i|(X_j, Z_j, y_j), j < i\} = 0\) almost surely. Then
\[
(nh)^{1/2} \left( \hat{\beta} - \beta_0 \right) \overset{D}{\longrightarrow} N(0, W_0^{-1} W_2 W_0^{-1}),
\]
where \(W_0^{-1}\) is the Moore–Penrose inverse of \(W_0\). If further the density function \(f_{\theta_0}(v)\) of \(\theta_0^T X\) is positive and the derivative of \(E(\varepsilon^2|\theta_0^T X = v)\) exists, then
\[
(nh)^{1/2} \left( \hat{g}(v) - g(v) - \frac{1}{2}\kappa_2 \hat{g}''(v)h^2 \right) \overset{D}{\longrightarrow} N(0, f_{\theta_0}^{-1}(v) \int (K(v))^2 dv E(\varepsilon^2|\theta_0^T X = v)),
\]
where \(\kappa_2 = \int K(v)v^2 dv\).

If \(E\{\varepsilon_i|(X_j, Z_j, y_j), j < i\} \neq 0\), then the asymptotic normal distribution still holds, but the covariance matrix in the distribution depends on the structure of the stochastic process of the observations. If \(E(\varepsilon^2|X, Z) = \sigma^2\) is constant, then the asymptotic distribution of \((\hat{\beta}, \hat{\theta})\) is the same as that obtained by Carroll et al. [4]. They further showed that their estimator is efficient in the semi-parametric sense if the conditional density of \(Y\) given \(X\) and \(Z\) belongs to the exponential distribution family. Therefore, our estimator is also efficient in the semi-parametric sense under the same conditions. If Theorem 1 is used for statistical inference about the parameters, we need to have a consistent estimator for the covariance matrix. Here, we provide such a consistent estimator for \(W_k\) as
\[
\hat{W}_k = n^{-1} \sum_{i=1}^n G(\hat{\theta}^T X_i)I(f_{\hat{\theta}}(\hat{\theta}^T X_i) > c_0) \left( \frac{Z_i - \hat{v}(X_i)}{d_i(X_i)} \right)^T \left( Y_i - \hat{\beta}^T Z_i - a_i \right)^k,
\]
where \(\hat{v}(x) = n^{-1} \sum_{i=1}^n K_{h,i}(X_i)X_i/f_{\hat{\theta}}(\hat{\theta}^T x)\), \(\hat{v}_0(x) = n^{-1} \sum_{i=1}^n K_{h,i}(X_i)Z_i/f_{\hat{\theta}}(\hat{\theta}^T x)\) and \(\hat{f}_{\hat{\theta}}(v)\) is defined above and \(k = 0\) and 2.

Bandwidth selection is always an important issue for nonparametric smoothing. One of the advantages of our method is that we do not need under-smoothing the link function. Therefore, most commonly used bandwidth selection methods can be employed here. Consider estimation of \(g\) at the final step of the iterations. For a given function \(w(\cdot)\) with compact support, minimizing the asymptotic weighted mean-squared error with weight \(f_{\theta_0}(\cdot)w(\cdot)\) yields the optimal global
bandwidth
\[
    h_0 = \left\{ \frac{\sigma^2 \int w(u) \, du \int (K(u))^2 \, du}{k_2^2 \int g''(u) f_\theta(u) w(u) \, du} \right\}^{1/5} n^{-1/5}.
\]

See also the discussion in Carroll et al. [4]. Both the cross-validation bandwidth selection method and the plug-in method [21] can be used to obtain bandwidths that are asymptotically consistent of \( h_0 \). Here, we give the details for the cross-validation bandwidth selection method. For any \( \theta \) and \( \beta \), let
\[
    CV_0(b) = n^{-1} \sum_{i=1}^{n} (Y_i - \beta^T Z_i - \hat{\alpha}_i^\dagger)^2 \quad \text{and} \quad CV(h) = n^{-1} \sum_{i=1}^{n} (Y_i - \beta^T Z_i - \hat{\alpha}_i^\dagger)^2,
\]
where \( \hat{\alpha}_i^\dagger \) and \( \hat{\alpha}_i^\dagger \) are, respectively, given in
\[
    \left( \hat{\alpha}_i^\dagger \right) = \left\{ \sum_{j=1 \atop j \neq i}^{n} w_{ji} \left( \frac{1}{X_j^T \theta} \right) \left( \frac{1}{X_j^T \theta} \right)^T \right\}^{-1} \sum_{j=1 \atop j \neq i}^{n} w_{ji} \left( \frac{1}{X_j^T \theta} \right) \left( Y_j - \beta^T Z_j \right)
\]
and
\[
    \left( \hat{\alpha}_i^\dagger \right) = \left\{ \sum_{j=1 \atop j \neq i}^{n} w^0_{ji} \left( \frac{1}{X_j^T \theta} \right) \left( \frac{1}{X_j^T \theta} \right)^T \right\}^{-1} \sum_{j=1 \atop j \neq i}^{n} w^0_{ji} \left( \frac{1}{X_j^T \theta} \right) \left( Y_j - \beta^T Z_j \right).
\]
The bandwidth in each step is then \( \hat{b} = \arg \min_b CV_0(b) \) for the weight function \( w_{ij} \) and \( \hat{h} = \arg \min_h CV(h) \) for the weight function \( w^0_{ij} \).

3. Numerical comparisons

In this section, we first use an example to demonstrate the relation between estimation errors and the bandwidth. We then use the examples in [4,7] to check the performance of our estimation method for finite data sets. In our simulations, kernel functions \( H(x) = 3(1 - |x|^2)I(|x| < 1)/4 \) and \( K(u) = 3(1 - u^2)I(|u| < 1)/4 \) are used. A computer code for the calculation is available at http://www.stat.nus.edu.sg/~staxyc/plsi.m

Example 3.1. Consider the following model:
\[
    y_t = \beta_{01} x_{1t} + \beta_{02} z_{2t} + 2 \exp\{-3(\theta_{01} x_{1t-1} + \theta_{02} x_{1t-2} + \theta_{03} x_{1t-3})^2\} + 0.5 \epsilon_t,
\]
where \( x_t = 0.4 x_{t-1} - 0.5 x_{t-2} + u_t \) with \( u_t, t = 1, 2, \ldots, \sim \text{IID } \text{Uniform}(-1, 1); z_{1t} \) and \( z_{2t}, t = 1, 2, \ldots, \sim \text{IID } \text{Normal}(0, 1); \) and that \( \{u_t\}, \{z_{1t}\}, \{z_{1t}\} \) and \( \{\epsilon_t\} \) are independent series. Here, \( Z_t = (z_{1t}, z_{2t})^T \) and \( X_t = (x_{t-1}, x_{t-2}, x_{t-3}, \ldots, x_{t-p})^T. \) The true parameters are \( \beta = (\beta_{01}, \beta_{02})^T = (1, 2)^T \) and \( \theta = (\theta_{01}, \theta_{02}, \theta_{03}, \ldots, \theta_{0p})^T = (-2/3, 1/3, 2/3, 0, \ldots, 0)^T. \) We define the estimation errors as \( e_\beta = (|\hat{\beta}_1 - \beta_{01}| + |\hat{\beta}_2 - \beta_{02}|)/2 \) and \( e_\theta = 1 - |\hat{\theta}^T \theta| \) for \( \hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)^T \)
Consider the following two models:

\[ y = 4\left(\frac{x_1 + x_2 - 1}{\sqrt{2}}\right)^2 + 4 + 0.2\varepsilon, \quad (3.1) \]

\[ y = \sin\{\pi(\frac{x_1 + x_2 + x_3}{\sqrt{3}} - A)/(B - A)\} + \beta Z + 0.1\varepsilon, \quad (3.2) \]

where \(x_1, x_2, x_3\) are independent uniformly distributed on \([0, 1]\), \(A = 0.3912\) and \(B = 1.3409\). Model (3.1) was used by Härdle et al. [7], in which \(\theta_0 = (\theta_{11}, \theta_{12})^T = (1/\sqrt{2}, 1/\sqrt{2})^T\). Model (3.2) was used by Carroll et al. [4], in which \(\theta_0 = (\theta_{21}, \theta_{22}, \theta_{23})^T = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^T\). We start the simulation for model (3.1) with \(\theta = (1, 3)^T/\sqrt{10}\) and model (3.2) with \(\theta = (0, 1, 2)^T/\sqrt{5}\). The cross-validation bandwidth is used. The number of replications is 100. With sample size \(n = 50, 100\) and 200, respectively, the simulation results are listed in Table 1.

For model (3.1), the corresponding simulation results of \(\phi = \arccos(\theta_{11})\) were 0.766(0.103), 0.792(0.084), 0.782(0.045) for \(n = 50, 100\) and 200, respectively, in of Härdle et al. [7]. Our results outperform theirs. A possible reason is that minimizing the cross-validation type of residuals was used to estimate the parameters in their paper, which reduces the estimation efficiency. See [24] for details. For model (3.2), the corresponding simulation results of Carroll et al. [4] for \(\theta_{21}, \theta_{22}\) and \(\theta_{23}\) were \((1.4e - 4), (1.6e - 4)\) and \((1.3e - 4)\), respectively, when \(n = 200\). Our results also improve theirs.
Table 1
Mean and mean-squared error (in parentheses) of the estimated parameters for models (3.1) and (3.2)

<table>
<thead>
<tr>
<th>n</th>
<th>$\theta_{11}$</th>
<th>$\theta_{12}$</th>
<th>$\phi$ = arccos($\theta_{11}$)</th>
<th>$\theta_{21}$</th>
<th>$\theta_{22}$</th>
<th>$\theta_{23}$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.7117</td>
<td>0.6965</td>
<td>0.7746</td>
<td>0.5793</td>
<td>0.5727</td>
<td>0.5785</td>
<td>0.2967</td>
</tr>
<tr>
<td></td>
<td>(0.0040)</td>
<td>(0.0045)</td>
<td>(0.0918)</td>
<td>(5.5e-4)</td>
<td>(5.7e-4)</td>
<td>(6.5e-4)</td>
<td>(1.1e-3)</td>
</tr>
<tr>
<td>100</td>
<td>0.7074</td>
<td>0.7047</td>
<td>0.7835</td>
<td>0.5785</td>
<td>0.5780</td>
<td>0.5748</td>
<td>0.2972</td>
</tr>
<tr>
<td></td>
<td>(0.0015)</td>
<td>(0.0015)</td>
<td>(0.0541)</td>
<td>(2.8e-4)</td>
<td>(2.6e-4)</td>
<td>(2.2e-4)</td>
<td>(4.7e-4)</td>
</tr>
<tr>
<td>200</td>
<td>0.7071</td>
<td>0.7059</td>
<td>0.7845</td>
<td>0.5776</td>
<td>0.5770</td>
<td>0.5772</td>
<td>0.2992</td>
</tr>
<tr>
<td></td>
<td>(0.0008)</td>
<td>(0.0008)</td>
<td>(0.0403)</td>
<td>(1.2e-4)</td>
<td>(1.3e-4)</td>
<td>(1.2e-4)</td>
<td>(2.5e-4)</td>
</tr>
</tbody>
</table>

4. Real data analysis

Now we return to our real data sets in Section 1. The Epanechnikov kernel and the cross-validation bandwidths are used in the calculations below.

4.1. Credit scoring

We partition $x_4$ and $x_5$ into 20 intervals separately. All the observations in a combination of one interval of $x_4$ and one interval of $x_5$ form a group. In all the groups, the proportion $\hat{p}_i$, of customers who paid their installments without problem, and $y_i = \log(\hat{p}_i/(1-\hat{p}_i))$ are calculated. We consider model (1.3) with all the covariates by taking $Z = (x_2, x_3, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{22}, x_{23}, x_{24})^T$, $X = (x_4, x_5)^T$. Because $x_{10}, \ldots, x_{24}$ are categorical variables, they are denoted by dummy variables. Here, $x_4$ and $x_5$ are standardized, respectively, for ease of calculations. For simplicity, we assume $E(\varepsilon^2|X, Z) = \sigma^2$ is a constant. Applying the estimation procedure to the data set, we obtain the estimates of the parameters as listed in Table 2. See [17] for more explanations. The estimate of the unknown function is shown in the right panel of Fig. 5. The nonlinearity in $x_4$ and $x_5$, i.e. $\hat{\theta}^T X = 0.249 x_4 + 0.969 x_5$, is clear as shown in Fig. 5.

Based on the estimation results as shown in Fig. 5, we have the following conclusion. Recall that $x_4$ is the amount of loan. Fig. 2 in Section 1 shows that as the loan increases, the customer’s ability to pay the installment increases, i.e. $y$ decreases. This seems misleading. For age $x_5$, Fig. 2 suggests that there is a range for the age at which the customer has less problem to pay their loan. In comparison, the estimated model can give us a more reasonable explanation. Fig. 5 suggests that as the age and the loan, $\hat{\theta}^T X = 0.249 x_4 + 0.969 x_5$, increases, the ability to pay the installments decreases. However, there is an age range, in which the customer has high ability to pay the loan.

4.2. Circulatory and respiratory (CR) problems in Hong Kong

Due to the hospital booking system, the day-of-the-week can affect $y_i$. We use dummy variables to describe the day of the $t$th observation by a 6-dimension vector $(D_{t1}, \ldots, D_{t6})$, where $D_{tk} = 1$ if the observation is taken on the $k$th day of a week; 0 otherwise. Together with lagged variables of pollutants and weather conditions in 1 week, we take $Z_t = (D_{t1}, \ldots, D_{t6}, x_{1,t-1}, \ldots, x_{1,t-7}, x_{2,t-1}, \ldots, x_{2,t-7}, x_{3,t-1}, \ldots, x_{3,t-7}, x_{4,t-1}, \ldots, x_{4,t-7})^T$ and $X_t = (x_{5,t-1}, \ldots, x_{5,t-7}, x_{6,t-1}, \ldots, x_{6,t-7})^T$. These models are estimated by using xkernels with the cross-validation bandwidths.
Fig. 5. Estimation results of the credit scoring data. The left panel is \( y - \hat{\beta}^T Z \) plotted against \( \hat{\theta}^T X \). The right panel is the estimated \( g \) and 95% symmetric pointwise confidence interval.

Table 2
Estimation results of the credit scoring data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coeff.</th>
<th>SE</th>
<th>Variable</th>
<th>Coeff.</th>
<th>SE</th>
<th>Variable</th>
<th>Coeff.</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.159</td>
<td>0.163</td>
<td>17#2</td>
<td>-1.718</td>
<td>0.472</td>
<td>20#3</td>
<td>-0.082</td>
<td>0.294</td>
</tr>
<tr>
<td>3</td>
<td>0.021</td>
<td>0.114</td>
<td>17#3</td>
<td>-1.211</td>
<td>0.433</td>
<td>20#4</td>
<td>0.263</td>
<td>0.251</td>
</tr>
<tr>
<td>6</td>
<td>-0.109</td>
<td>0.105</td>
<td>17#4</td>
<td>1.977</td>
<td>0.576</td>
<td>21#2</td>
<td>-2.194</td>
<td>0.683</td>
</tr>
<tr>
<td>7</td>
<td>-0.454</td>
<td>0.119</td>
<td>17#5</td>
<td>4.715</td>
<td>0.932</td>
<td>21#3</td>
<td>-1.490</td>
<td>0.363</td>
</tr>
<tr>
<td>8</td>
<td>0.189</td>
<td>0.120</td>
<td>17#6</td>
<td>-1.327</td>
<td>0.316</td>
<td>22#2</td>
<td>-1.102</td>
<td>0.582</td>
</tr>
<tr>
<td>9</td>
<td>0.032</td>
<td>0.091</td>
<td>18#2</td>
<td>2.145</td>
<td>0.528</td>
<td>22#3</td>
<td>-0.785</td>
<td>0.490</td>
</tr>
<tr>
<td>10#2</td>
<td>0.817</td>
<td>0.302</td>
<td>18#3</td>
<td>1.037</td>
<td>0.413</td>
<td>22#4</td>
<td>0.753</td>
<td>0.715</td>
</tr>
<tr>
<td>11#2</td>
<td>0.188</td>
<td>0.293</td>
<td>18#4</td>
<td>0.878</td>
<td>0.447</td>
<td>22#5</td>
<td>0.770</td>
<td>0.584</td>
</tr>
<tr>
<td>12#2</td>
<td>0.635</td>
<td>0.303</td>
<td>18#5</td>
<td>1.756</td>
<td>0.359</td>
<td>22#6</td>
<td>-3.837</td>
<td>0.957</td>
</tr>
<tr>
<td>13#2</td>
<td>-0.815</td>
<td>0.276</td>
<td>18#6</td>
<td>1.876</td>
<td>0.449</td>
<td>22#7</td>
<td>2.253</td>
<td>0.608</td>
</tr>
<tr>
<td>14#2</td>
<td>1.680</td>
<td>0.544</td>
<td>18#7</td>
<td>1.770</td>
<td>0.551</td>
<td>22#8</td>
<td>0.838</td>
<td>0.531</td>
</tr>
<tr>
<td>15#2</td>
<td>1.416</td>
<td>0.347</td>
<td>19#2</td>
<td>0.416</td>
<td>0.369</td>
<td>22#9</td>
<td>1.441</td>
<td>0.526</td>
</tr>
<tr>
<td>15#3</td>
<td>2.411</td>
<td>0.469</td>
<td>19#3</td>
<td>1.287</td>
<td>0.307</td>
<td>22#10</td>
<td>-1.519</td>
<td>1.199</td>
</tr>
<tr>
<td>15#4</td>
<td>3.247</td>
<td>0.520</td>
<td>19#4</td>
<td>-0.966</td>
<td>0.539</td>
<td>22#11</td>
<td>-0.644</td>
<td>0.510</td>
</tr>
<tr>
<td>15#5</td>
<td>2.782</td>
<td>0.617</td>
<td>19#5</td>
<td>1.343</td>
<td>0.673</td>
<td>23#2</td>
<td>0.087</td>
<td>0.350</td>
</tr>
<tr>
<td>15#6</td>
<td>0.987</td>
<td>0.374</td>
<td>19#6</td>
<td>1.691</td>
<td>0.465</td>
<td>23#3</td>
<td>0.787</td>
<td>0.499</td>
</tr>
<tr>
<td>16#2</td>
<td>0.214</td>
<td>0.476</td>
<td>19#7</td>
<td>0.992</td>
<td>0.539</td>
<td>24#2</td>
<td>1.717</td>
<td>0.612</td>
</tr>
<tr>
<td>16#3</td>
<td>0.680</td>
<td>0.431</td>
<td>19#8</td>
<td>-1.170</td>
<td>0.566</td>
<td>24#4</td>
<td>0.249</td>
<td>0.026</td>
</tr>
<tr>
<td>16#4</td>
<td>1.714</td>
<td>0.539</td>
<td>19#9</td>
<td>0.173</td>
<td>0.608</td>
<td>5</td>
<td>0.969</td>
<td>0.007</td>
</tr>
<tr>
<td>16#5</td>
<td>1.218</td>
<td>0.442</td>
<td>19#10</td>
<td>1.070</td>
<td>0.348</td>
<td>23#2</td>
<td>0.087</td>
<td>0.350</td>
</tr>
<tr>
<td>16#6</td>
<td>1.588</td>
<td>0.465</td>
<td>20#2</td>
<td>2.021</td>
<td>0.539</td>
<td>24#3</td>
<td>1.441</td>
<td>0.526</td>
</tr>
</tbody>
</table>

The bold coeff. are statistically significant with “t-value” greater than 2.0.

The estimated link function \( \hat{g} \) is shown in Fig. 6.
5. Conclusion

This paper proposes an estimation method for the partially linear single-index model. Good properties of the method have been demonstrated theoretically and by simulations. These properties are: (1) a $\sqrt{n}$-consistent pilot estimator is not needed, see Theorem 1. This solves the problems addressed in [4]; (2) convergence of the PLSIM algorithm is proved, see the proof of Theorem 1 in Section 6; (3) a wide range of bandwidth can be used including the cross-validated bandwidth. This makes the algorithm stable and frees it from the audible critique on “the necessity of uncontrollable hyperparameters”; (4) under normal noise assumptions, the estimators of the parameters are asymptotically efficient in semi-parametric sense, see [4]; and (5) the PLSIM algorithm is applicable to time series. Of course, there are some limitations to the method. Because the estimators are based on minimizing the sum of residual squares, they are not the most efficient in the semi-parametric sense when the noise is not normally distributed. Extending the approach in this paper to more complicated models such as the generalized partially linear single-index model given in [4], the simple algorithm proposed in this paper cannot be used.

Some important problems for the model and the estimation method need to be investigated in further work. These problems include variable selection and model checking. However, since the estimation has been simplified in this paper to solving simple regression problems as in (2.5) and (2.6). It is possible that the problems can be solved based on the methods in this paper.

6. Assumptions and proofs

Let $U = (X^T, Z^T)^T$. Suppose $\{(U_i, y_i), i = 1, \ldots, n\}$ is a set of observations. We make the following assumptions on the stochastic nature of the observations, the link function and the kernel functions.

(C1) The observations are a strongly mixing and stationary sequence with geometric decaying mixing rate $x(k)$. 

Based on this model, the effects of weather conditions on the CR problems are as follows. The coefficients of temperatures $x_{5,t-2}$ and $x_{5,t-5}$ forms a contrast. Together with Fig. 6, it suggests that a rapid temperature variation (rather than the temperature itself) will increase the hospital admission $y_t$. The coefficients of humidity $x_{6,t-4}$ and $x_{6,t-7}$ have about the same value, which can be taken as an average. Together with Fig. 6, it suggests that extreme dry or wet weather will increase the hospital admission in Hong Kong.
(C2) With Probability 1, $X$ lies in a compact set $D$; the marginal density functions $f$ of $X$ and $f_0$ of $\theta^T X$ for any $|\theta| = 1$ have bounded derivatives; regions $\{x : f(x) \geq c_0\}$ and $\{x : f_0(\theta^T x) > c_0\}$ for all $\theta : |\theta| = 1$ are nonempty.

(C3) For any perpendicular unit norm vectors $\theta$ and $\vartheta$, the joint density function $f(u_1, u_2)$ of $(\theta^T X, \vartheta^T X)$ satisfies $f(u_1, u_2) < c_{f\theta} f_{\theta^T X}(u_1) f_{\vartheta^T X}(u_2)$, where $c$ is a constant.

(C4) $g$ has bounded, continuous third-order derivative; the conditional expectations $E(Z|X) = x$, $E(ZZ^T|X = x)$, $E(U|\theta^T X = v)$ and $E(UU^T|\vartheta^T X = v)$ have bounded derivatives; $E(\vartheta^T X = x)$, $E(\vartheta^T X|X = x)$, $E((Z_\ell||Z_1)|X_1 = x_1, X_\ell = x_\ell)$ and $E((Z_\ell||Z_1)|\theta^T X_1 = u, \vartheta^T X_\ell = v)$ are bounded by a constant for all $\ell > 0, x_1, x_\ell, x, u$ and $v$, where $r > 3$.

(C5) $H$ is a density function with bounded derivative and compact support $\{|x| \leq a_0\}$ for some $a_0 > 0$; $K$ is a symmetric density function with bounded derivative and compact support $[-b_0, b_0]$ for some $b_0 > 0$ and that the Fourier transform of $K$ is absolutely integrable.

(C6) Matrix $E((Z - E(Z|X))(Z - E(Z|X))^T)$ is a positive definite matrix.

The mixing rate [3] in (C1) can be relaxed to algebraic rate $\omega(k) = O(k^{-p})$. Suppose the bandwidth $h \sim n^{-\delta}$. Then the mixing rate satisfying the following equation is sufficient:

$$\sum_{n=1}^{\infty} n^{-\frac{1}{2} - \frac{1}{r} - \delta(\frac{1}{r} + \frac{1}{p})\rho + 2p + \frac{1}{r} + \frac{1}{p} + \frac{1}{2} + \frac{1}{2}\delta} (\log n)^{p/2} < \infty.$$

The regions with positive densities in (C2) are needed to avoid zero values of the denominator of the kernel estimator of regression. There are different approaches for this purpose. See, e.g. [7,9,15]. However, their ideas are the same. We can further assume that $c_0$ decreases to 0 with $n$ at a slow speed, but it makes no difference in practice. Assumption (C3) ensures successful searching for the direction $\theta$ globally. If we restrict the searching area, the assumption can be removed; see [7]. The third-order derivatives in (C4) is needed for higher-order expansion. Actually, existence of second-order derivative is sufficient for the root-$n$ consistency. In this paper, we only employ kernel functions with compact support as in (C5). We further assume that $\kappa_2 \equiv \int K(v) v^2 = 1$ and $H_2 \equiv \int H(U)U^T dU = I_{p \times p}$; otherwise we may take $K(v) = K(v/\sqrt{\kappa_2})/\sqrt{\kappa_2}$ and $H(U) = H(H_2^{-1/2}U)(det(H_2))^{-1/2}$. (C6) is imposed for identification. Similarly, if we search for the direction $\theta$ in a small neighborhood of $\theta_0$ as in [4,7], (C6) can be removed.

6.1. Proof of the main results

The basic tools are given in Lemmas A.1–A.3. Some simple calculation results are listed in Lemmas 6.1–6.3. Based on these lemmas, Lemma 1 and Theorem 1 are proved. Let $\delta_\theta = |\theta - \theta_0|$, $\delta_\beta = |\beta - \beta_0|$ and $\delta_\gamma = \delta_\theta + \delta_\beta$. In a bounded parameter space, $\delta_\theta$, $\delta_\beta$ and $\delta_\gamma$ are bounded. Let $\delta_{pn} = (\log n/(nb^p))^1/2$, $\tau_{pn} = b^2 + \delta_{pn}$, $\delta_n = (\log n/nh)^1/2$, $\tau_n = h^2 + \delta_n$ and $\delta_0n = (\log n/n)^1/2$. By the condition $h \sim n^{-\delta}$ with $1/6 < \delta < 1/4$, we have $\delta_0n \ll h^2 \ll h^{-1}\delta_n$ and $\delta_n \ll h$. We shall use these relations frequently in our calculations. Let $\Theta = \{\theta : |\theta| = 1\}$. Suppose $A_n$ is a matrix. $\bar{A}_n = O(a_n)$ (or $o(a_n)$) means every element in $A_n$ is $O(a_n)$ (or $o(a_n)$) almost surely. We adopt the consistency in the sense of “almost surely” because we need to prove the convergence of the algorithm, which theoretically need infinite iterations. Let $c, c_1, c_2, \ldots$ be a set of constants. For ease of exposition, $c$ may have different values at different places. We abbreviate $K_h(\theta^T X_{i0})$ and $H_h(X_{i0})$ as $K_{h,i}^\theta(x)$ (or $K_{h,i}(x)$) and $H_{h,i}(x)$ (or $H_{h,i}$), respectively, in the following context. We take $G(\cdot) \equiv 1$ in the proofs for simplicity.
In the following context, we abbreviate $L$ for any function $L(x)$, and $L_\theta$ or $L_\theta(x)$ for any function $L_\theta(\theta^T x)$. Let $\nu_0$ and $\mu_0$ be defined as in Section 2, and

$$v = E(Z|X = x), \quad \pi = E(ZZ^T|X = x), \quad \pi_\theta = E(ZZ^T|\theta^T X = \theta^T x),$$

$$\Sigma_\theta = E(XX^T|\theta^T X = \theta^T x) - \mu_0 x^T - x \mu_\theta^T + xx^T.$$ 

Let

$$\zeta_0 = \frac{1}{n} \sum_{i=1}^{n} H_{b,i}, \quad S_1 = \frac{1}{n} \sum_{i=1}^{n} H_{b,i} X_{i0}, \quad S_2 = \frac{1}{n} \sum_{i=1}^{n} H_{b,i} X_{i0} X_{i0}^T,$$

$$T_1 = \frac{1}{n} \sum_{i=1}^{n} H_{b,i} Z_i, \quad T_2 = \frac{1}{n} \sum_{i=1}^{n} H_{b,i} Z_i Z_i^T, \quad C_2 = \frac{1}{n} \sum_{i=1}^{n} H_{b,i} X_{i0} Z_i^T,$$

$$E_1 = \frac{1}{n} \sum_{i=1}^{n} H_{b,i} Z_i y_i, \quad D_1 = \frac{1}{n} \sum_{i=1}^{n} H_{b,i} X_{i0} y_i, \quad W_n = \zeta_0 S_2 - S_1 S_1^T$$

and

$$\tilde{w}_{a,i}^0(x) = (\theta^T S_2 \theta) H_{b,i} - \theta^T S_1 H_{b,i} \theta^T X_{i0}, \quad \tilde{w}_{d,i}^0(x) = \zeta_0 H_{b,i} \theta^T X_{i0} - \theta^T S_1 H_{b,i}.$$

Based on (2.4), we can obtain initial estimators of \( \zeta_0 \) and any vector $\beta$. Let $\tilde{w}_j^0 = \tilde{\theta}^T W_n(X_j) \theta$ and calculate

$$\tilde{\theta} = (\tilde{\theta}_j^0)^{-1} \sum_{i=1}^{n} \tilde{w}_{a,i}^0(X_j) \{ y_i - \beta^T Z_i \}, \quad \tilde{\theta}_j = (\tilde{\theta}_j^0)^{-1} \sum_{i=1}^{n} \tilde{w}_{d,i}^0(X_j) \{ y_i - \beta^T Z_i \}, \quad (6.1)$$

$$\tilde{\theta} = (\tilde{\theta}_j^0)^{-1} \sum_{j=1}^{n} I_n(X_j) \left( \begin{array}{c} E_1(X_j) - \tilde{\theta}^0 T_1(X_j) \\ D_1(X_j) - \tilde{\theta}^0 D_2(X_j) \end{array} \right) / \zeta_0(X_j), \quad (6.2)$$

where $sgn_1$ is the sign of first entry in $\tilde{\theta}$ and $A^-$ denotes the Moore–Penrose inverse of matrix $A$. Repeat the calculations in (6.1) and (6.2) with $(\theta, \beta)$ replaced by $(\tilde{\theta}, \tilde{\beta})$ until convergence. Denote the final value by $(\tilde{\theta}, \tilde{\beta})$. Next, we shall improve the efficiency of the estimators using a univariate kernel. Let

$$\zeta_k^0 = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^0(\theta^T X_{i0})^k, \quad k = 0, 1, 2, 3,$$

$$w_{a,i}^0 = \zeta_2^0 K_{h,i}^0 - \zeta_1^0 K_{h,i}^0 \theta^T X_{i0}, \quad w_{d,i}^0 = \zeta_0^0 K_{h,i}^0 \theta^T X_{i0} - \zeta_1^0 K_{h,i}^0,$$

$$w = \frac{1}{n} \sum_{i=1}^{n} w_{a,i}^0, \quad S_1 = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^0 X_{i0}, \quad S_2 = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^0 X_{i0} X_{i0}^T.$$
\[ T_1^0 = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^T Z_i, \quad E_1^0 = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^T Z_i y_i, \quad D_1^0 = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^T X_i y_i, \]

\[ T_2^0 = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^T Z_i Z_i^T, \quad C_2^0 = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^T y_i^T X_i Z_i^T. \]

\[ S_{1,1}^0 = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^T \{ (\theta^T X_{i0}) X_{i0} \}, \quad S_{2,1}^0 = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^T \{ (\theta^T X_{i0}) \}^2 X_{i0}, \]

\[ S_{1,2}^0 = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^T \{ (\theta^T X_{i0}) X_{i0} X_{i0}^T \}, \quad S_{2,2}^0 = \frac{1}{n} \sum_{i=1}^{n} K_{h,i}^T X_{i0} \{ (\theta - \theta_0)^T X_{i0} \}^2. \]

Based on (2.4), we improve the estimators \( \hat{\theta} \) and \( \hat{\beta} \) as follows. Let \( w_j^0 = w^0(X_j) \). Starting with \( (\theta, \beta) = (\hat{\theta}, \hat{\beta}) \), calculate

\[ \hat{a}_j^0 = (w_j^0)^{-1} \sum_{i=1}^{n} w_{a,i}(X_j) \{ y_i - \beta^T Z_i \}, \quad \hat{d}_j^0 = (w_j^0)^{-1} \sum_{i=1}^{n} w_{d,i}(X_j) \{ y_i - \beta^T Z_i \}, \quad (6.3) \]

\[ \begin{pmatrix} \hat{\beta} \\ \hat{\theta} \end{pmatrix} = \left( \begin{pmatrix} \hat{a}_j^0 \\ \hat{d}_j^0 \end{pmatrix} \right)^{-1} \sum_{j=1}^{n} I_n(X_j) \left( E_j^0(X_j) - \frac{1}{\hat{a}_j^0 D_j^0(X_j)} - \frac{1}{\hat{d}_j^0 C_j^0(X_j)} S_j^0(X_j) \right) \hat{\theta}(X_j), \quad (6.4) \]

where \( sgn_1 \) is the sign of first entry of \( \hat{\theta} \) and

\[ D_n^0 = \sum_{j=1}^{n} I_n(X_j) \left( \begin{pmatrix} T_j^0(X_j) \\ C_j^0(X_j) \end{pmatrix} \right)^T \left( \begin{pmatrix} D_j^0 \end{pmatrix} \right)^{-1} \left( \begin{pmatrix} S_j^0 \end{pmatrix} \right). \]

Repeat the procedure (6.3) and (6.4) with \( (\theta, \beta) \) replaced by \( (\hat{\theta}, \hat{\beta}) \) until convergence. Denote the final value by \( (\hat{\beta}, \hat{\theta}) \).

Let \( \tilde{\Delta}_i(x) = y_i - \tilde{a} - \beta_0^T Z_i - \tilde{a} X_{i0}^T \theta_0 \) and \( \tilde{\Delta}_i(x) = y_i - \tilde{a} - \beta_0^T Z_i - \tilde{a} X_{i0}^T \theta_0 \). We have

\[ \begin{pmatrix} \tilde{\beta} \\ \tilde{\theta} \end{pmatrix} = \left( \begin{pmatrix} \theta_0 \\ \theta_0 \end{pmatrix} \right) + \tilde{D}_n^{-1}(\theta) \sum_{j=1}^{n} I_n(X_j) \sum_{i=1}^{n} H_{h,i}(X_j) \left( \begin{pmatrix} Z_i \\ X_{ij} \end{pmatrix} \right) \Delta_i(X_j)/\hat{\theta}_0(X_j), \quad (6.5) \]

\[ \begin{pmatrix} \tilde{\beta} \\ \tilde{\theta} \end{pmatrix} = \left( \begin{pmatrix} \theta_0 \\ \theta_0 \end{pmatrix} \right) + \tilde{D}_n^{-1}(\theta) \sum_{j=1}^{n} I_n(X_j) \sum_{i=1}^{n} K_{h,i}(\theta^T X_j) \left( \begin{pmatrix} Z_i \\ X_{ij} \end{pmatrix} \right) \Delta_i(X_j)/\hat{\theta}_0(X_j), \quad (6.6) \]

By Lemmas A.1 and A.3, we have

\[ \hat{\theta}_0 = f(x) + O(J_0 + \tau_{pn}), \quad S_1 = O(bJ_0 + b\tau_{pn}), \]

\[ S_2 = f(x) I_{p_X} b^2 + O(b^2 J_0 + b^2 \tau_{pn}), \quad T_1 = f(x) v(x) + O(J_0 + \tau_{pn}), \]

\[ T_2 = f(x) \pi(x) + O(J_0 + \tau_{pn}), \quad \frac{1}{n} \sum_{i=1}^{n} H_{h,i} Z_i \varepsilon_i = O(\delta_{pn}), \]
\[ \frac{1}{n} \sum_{i=1}^{n} H_{b,i} \varepsilon_i = O(\delta_{pn}), \quad \frac{1}{n} \sum_{i=1}^{n} H_{b,i} X_{i0} (\theta^T X_{i0})^k \varepsilon_i = O(b^{k+1} \delta_{pn}), \]
\[ \frac{1}{n} \sum_{i=1}^{n} H_{b,i} |X_{i0}|^k = O(b^k), \quad C_2 = O(bJ_0 + b^2 + b\delta_n), \tag{6.7} \]
and
\[ \zeta_0^0 = f_\theta + O(J_0 + \tau_n), \quad \zeta_1^0 = O(hJ_0 + h^2 + h\delta_n), \quad \zeta_2^0 = f_\theta h^2 + O(h^2 J_0 + h^2 \tau_n), \]
\[ \zeta_3^0 = O(h^4 + b^3 J_0 + h^3 \delta_n), \quad S_1^0 = f_\theta (\mu_0 - x) + O(J_0 + \tau_n), \quad S_2^0 = \tilde{\Sigma}_0 f_\theta + O(J_0 + \tau_n), \]
\[ w_0^0 = f_\theta^2 h^2 + O(h^2 J_0 + h^2 \tau_n), \quad T_1^0 = f_\theta v_0 + O(J_0 + \tau_n), \quad T_2^0 = f_\theta \pi_0 + O(J_0 + \tau_n), \]
\[ C_2^0 = O(hJ_0 + h^2 + h\tau_n), \quad S_{1,1}^0 = O(hJ_0 + h^2 + h\tau_n), \quad S_{1,2}^0 = O(hJ_0 + h^2 + h\tau_n), \]
\[ S_{2,1}^0 = f_\theta (\mu_0 - x) h^2 + O(h^2 J_0 + h^2 \tau_n), \quad S_3^0 = O(\delta_0^2). \tag{6.8} \]

Let \( \tilde{a}, \tilde{a}, \tilde{\alpha} \) and \( \tilde{d} \) be, respectively, the values of \( \tilde{a}_j, \tilde{d}_j, \tilde{\alpha}_j \) and \( \tilde{d}_j \) with \( X_j \) replaced by \( x \). For simplicity, we further assume that \( f(x) > c_0 \) and \( f_\theta(\theta^T x) > c_0 \) for all \( x \in \mathcal{D} \) (otherwise, we only need to change \( \mathcal{D} \) to \( \{ x : f(x) > c_0 \} \) or \( \{ x : f_\theta(\theta^T x) > c_0 \} \) in the proofs). Thus, \( I_n(X_j) \equiv 1 \) when \( n \) is sufficiently large.

**Lemma 6.1.** Let \( \beta_d = \beta_0 - \beta \) and \( \theta_d = \theta_0 - \theta \). Suppose assumptions (C1)–(C5) hold. We have
\[ \tilde{a} = g(\theta_0^T x) + v^T \beta_d + O(J_0 + b + \delta_{pn}), \]
\[ \tilde{d} = \theta^T \theta_0 g'((\theta_0^T x) + O((1 + b^{-1} J_0) \delta_\beta + b^{-1} \delta_{pn} + b), \]
\[ \tilde{a} = g(\theta_0^T x) + g'((\theta_0^T x) \mu_0 - x) + v^T \beta_d + \frac{1}{2} g''((\theta_0^T x) h^2 + R_{n,3} + O(\delta_0^2 + J_0 \delta_\gamma + \tau_n \delta_\gamma + h \tau_n), \]
\[ \tilde{d} = g'((\theta_0^T x) + h^{-1} R_{n,4} + O(\delta_0^2 + (h^{-1} J_0 + 1 + h^{-1} \delta_\gamma) \delta_\gamma + \tau_n) \]
uniformly for \( x \in \mathcal{D} \) and \( \theta \in \Theta \), where \( R_{n,3} = \lfloor nf_\theta \rfloor^{-1} \sum_{i=1}^{n} K_{h,i}^T \varepsilon_i \) and \( R_{n,4} = \lfloor nhf_\theta \rfloor^{-1} \sum_{i=1}^{n} K_{h,i}^T \tilde{X}_{i0} \varepsilon_i \).

**Lemma 6.2.** Suppose assumptions (C1)–(C5) hold. We have
\[ \frac{1}{n} \tilde{D}_n(\theta) = \begin{pmatrix} E(ZZ^T) + O(b + \delta_{pn}) & O(b^2 + b^2 \delta_{pn}) \\ O(b^2 + b^2 \delta_{pn}) & (\theta^T \theta_0)^2 E[g'(\theta_0^T x)]^2 I_{p \times p} b^2 + O(b \delta_{pn} + b^2 \delta_\beta) \end{pmatrix} \]
and
\[ \frac{1}{n} \tilde{D}_n(\theta) = \begin{pmatrix} E(ZZ^T) & \tilde{C}_{12} \\ \tilde{C}_{12}^T & 2\tilde{W}_0 \end{pmatrix} + O(h^{-1} \delta_n + \delta_\gamma) \]
uniformly for \( \theta \in \Theta \), where \( \tilde{C}_{12} = E[g'(\theta_0^T x) Z(\mu_0(X) - X)^T] \) and \( \tilde{W}_0 = E[g'(\theta_0^T x)]^2 \{ X - \mu_0(X) \} \{ X - \mu_0(X) \}^T \).
Lemma 6.3. Suppose assumptions (C1)–(C5) hold. Then

\[
\frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} H_{b,i}(X_j)Z_i \bar{\Delta}_i(X_j) / \bar{\zeta}_0(X_j) = E\{v(X)\nu^T(X)\}(\beta - \beta_0) + O(b + \delta_{pn}),
\]

\[
\frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \bar{d}_j \sum_{i=1}^{n} H_{b,i}(X_j)X_{ij} \bar{\Delta}_i(X_j) / \bar{\zeta}_0(X_j) = b^2(\theta^T \theta_0)(1 - \theta^T \theta_0)E\{g'(\theta_0^T X)\}^2 \theta_0
\]

\[+ O(b^3 + b\delta_{pn} + b^2 \delta_\beta), \]

\[
\frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} K_{b,i}^0(X_j)Z_i \bar{\Delta}_i^0(X_j) / \bar{\zeta}_0^0(\theta^T X_j) = E\{v_0(X)\nu_0^T(X)\} \beta_d + \frac{1}{n} \sum_{i=1}^{n} \{Z_i - v_0(X_i)\} \varepsilon_i
\]

\[+ O((\delta_\theta + h + h^{-1} \delta_n) \delta_\gamma + h \tau_n), \]

\[
\frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} K_{b,i}^0(X_j)X_{ij} \bar{\Delta}_i^0(X_j) / \bar{\zeta}_0^0(\theta^T X_j)
\]

\[= \tilde{W}_0 \theta_d + \frac{1}{n} \sum_{i=1}^{n} g'(\theta_0^T X_i) \{\mu_0(X_j) - X_i\} \varepsilon_i
\]

\[+ O((\delta_\gamma + h^{-1} \delta_n + h) \delta_\gamma + h \tau_n + h^{-1} \delta_n^2), \]

uniformly for $\theta \in \Theta$.

**Proof of Lemma 1.** We shall prove that the equations in the Lemma 1 hold with probability 1. From Lemmas 6.2 and 6.3 and (6.5), we have for any $\beta$ and $\theta$ with $\theta^T \theta = 1$,

\[
\tilde{\beta} - \beta_0 = \{E(ZZ^T)\}^{-1} E\{v(Z)\nu^T(Z)\}(\beta - \beta_0) + O(b + b^{-1} \delta_{pn}). \tag{6.9}
\]

Note that the above equation does not depend on the choice of $\theta$. This is because we use a multivariate kernel, i.e. we use a more general multivariate function to replace $g(\theta_0^T X)$. In the algorithm, (6.9) can be written as

\[
\tilde{\beta}_{k+1} - \beta_0 = \{E(ZZ^T)\}^{-1} E\{v(X)\nu^T(X)\}(\tilde{\beta}_k - \beta_0) + O(b + b^{-1} \delta_{pn}), \tag{6.10}
\]

where the sub-index $k$ indicates that the corresponding values are the results of the $k$th iteration in the algorithm; see (6.1) and (6.2). By assumption (C6), $E(ZZ^T) - E\{v(X)\nu^T(X)\}$ is a positive definite matrix. Note that $E\{v(X)\nu^T(X)\}$ is a semi-positive matrix. Hence the eigenvalues of $\{E(ZZ^T)\}^{-1} E\{v(X)\nu^T(X)\}$ are all less than 1. After sufficiently many steps, we have from (6.10)

\[
\tilde{\beta}_k - \beta_0 = O(b + b^{-1} \delta_{pn}). \tag{6.11}
\]

See the proof of Theorem 1 below for more details. Therefore

\[
\tilde{\beta} - \beta_0 = O(b + b^{-1} \delta_{pn}). \tag{6.12}
\]

If $\theta^T \theta_0 \neq 0$, then it follows from Lemmas 6.2 and 6.3 and (6.5) that

\[
\tilde{\theta} - \theta_0 = (\theta^T \theta_0)^{-1}(1 - \theta^T \theta_0)\theta_0 + O(\delta_\beta + b + b^{-1} \delta_{pn}),
\]
i.e. \( \tilde{\theta} = (\theta^T \theta_0)^{-1} \theta_0 + O(\delta_\beta + b + b^{-1} \delta_{pn}) \). By (6.12), we may assume \( \delta_\beta \) is small enough (otherwise, take \( \beta = \beta_0 \)). Thus

\[
\tilde{\theta} := sgn_1 \frac{\tilde{\theta}}{|\tilde{\theta}|} = \theta_0 + O(\delta_\beta + b + b^{-1} \delta_{pn}).
\]

where \( sgn_1 \) is the sign of first entry of \( \tilde{\theta} \). In the algorithm, we have

\[
\tilde{\theta}_{k+1} - \theta_0 = O(\delta_{\beta_k} + b + b^{-1} \delta_{pn}). \tag{6.13}
\]

Combining (6.11) and (6.13), we have,

\[
\tilde{\theta} - \theta_0 = O(b + b^{-1} \delta_{pn}). \tag{6.14}
\]

The proof is completed. \( \square \)

**Proof of Theorem 1.** It follows from Lemmas 6.2 and 6.3 and (6.6) that

\[
\begin{pmatrix}
\tilde{\beta} - \beta_0 \\
\tilde{\theta} - \theta_0
\end{pmatrix} = \tilde{D} - N_n + \tilde{D}^{-1} \tilde{C} \begin{pmatrix}
\beta - \beta_0 \\
\theta - \theta_0
\end{pmatrix}
+ O\{(\delta_\gamma + h + h^{-1} \delta_n) \delta_\gamma + h \tau_n + h^{-1} \delta_n^2\},
\]

where

\[
\tilde{C} = \begin{pmatrix}
E\{v_{\theta_0}(X)v_{\theta_0}^T(X)\} & 0 \\
0 & \tilde{W}_0
\end{pmatrix}, \quad \tilde{D} = \begin{pmatrix}
E(ZZ^T) & \tilde{C}_{12} \\
\tilde{C}_{12}^T & 2\tilde{W}_0
\end{pmatrix},
\]

\[
N_n = \frac{1}{n} \sum_{i=1}^{n} \left( g'(\theta_0^T X_i)\{\mu_{\theta_0}(X_i) - X_i\} \right) \epsilon_i.
\]

Following the proof of Lemma 1 of Xia et al. [24], we have \( \tilde{C}, \tilde{D} \) and \( W_0 = \tilde{D} - \tilde{C} \) are all semi-positive matrices with rank \( p + q - 1 \). Therefore, \( D \overset{\text{def}}{=} (\tilde{D}^{-1})^{1/2} \tilde{C} (\tilde{D}^{-1})^{1/2} \) is a semi-positive matrix with all eigenvalues less than 1. There exist \( 1 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{p+q-1} > 0 \) and orthogonal matrix \( \Gamma \) such that

\[
D = \Gamma \text{diag}(\lambda_1, \ldots, \lambda_{p+q-1}, 0) \Gamma^T.
\]

Let \( (\tilde{\beta}_k, \tilde{\theta}_k) \) be the calculation results of the \( k \)th iteration in the algorithm; see (6.3) and (6.4). For any \( k \), Eq. (6.15) holds with \( (\tilde{\beta}, \tilde{\theta}) \) replaced by \( (\tilde{\beta}_{k+1}, \tilde{\theta}_{k+1}) \) and \( (\beta, \theta) \) by \( (\tilde{\beta}_k, \tilde{\theta}_k) \). Let

\[
\tilde{\gamma}_{k+1} = (\tilde{D}^{-1})^{1/2} N_n + D \tilde{\gamma}_k + O\{(\delta_{\tilde{\gamma}_k} + h + h^{-1} \delta_n) \delta_{\tilde{\gamma}_k} + h \tau_n + h^{-1} \delta_n^2\}. \tag{6.16}
\]

It follows that

\[
\delta_{\tilde{\gamma}_{k+1}} \leq \delta_{0n} + \lambda_1 \delta_{\tilde{\gamma}_k} + c(\delta_{\tilde{\gamma}_k} + h + h^{-1} \delta_n) \delta_{\tilde{\gamma}_k} + c(h \tau_n + h^{-1} \delta_n^2)
\]

\[
= \delta_{0n} + (\lambda_1 + c \delta_{\tilde{\gamma}_k} + c(h + h^{-1} \delta_n)) \delta_{\tilde{\gamma}_k} + c(h \tau_n + h^{-1} \delta_n^2) \tag{6.17}
\]

almost surely, where \( c \) is a constant. We can further take \( c > 1 \). Because \( h, h^{-1} \delta_n, \tau_n, \delta_{0n} \to 0 \) as \( n \to \infty \), we may assume that

\[
c(h + h^{-1} \delta_n) \leq (1 - \lambda_1)/3, \quad \delta_{0n} + c(h \tau_n + h^{-1} \delta_n^2) \leq (1 - \lambda_1)^2/(9c). \tag{6.18}
\]
By (6.12) and (6.14), we may assume
\[ \delta_{\gamma_1} \leq (1 - \lambda_1)/(3c). \] (6.19)
Therefore, it follows from (6.17)–(6.19) that
\[ \delta_{\gamma_2} \leq \{\lambda_1 + 2(1 - \lambda_1)/3\}(1 - \lambda_1)/(3c) + (1 - \lambda_1)^2/(9c) = (1 - \lambda_1)/(3c). \] (6.20)
From (6.17), (6.18) and (6.20), we have
\[ \delta_{\gamma_3} \leq (1 - \lambda_1)/(3c). \]
Consequently, \( \delta_{\gamma_k} \leq (1 - \lambda_1)/(3c) \) for all \( k \). Therefore we have from (6.17) that
\[ \delta_{\gamma_{k+1}} \leq \lambda_0 \delta_{\gamma_k} + \delta_{0n} + c(h \tau_n + h^{-1} \delta_n^2) \]
almost surely, where \( 0 \leq \lambda_0 < (2 + \lambda_1)/3 < 1 \). It follows that
\[ \delta_{\gamma_k} \leq \lambda_0^k \delta_{\gamma_1} + \{\delta_{0n} + c(h \tau_n + h^{-1} \delta_n^2)\} \sum_{j=1}^{\infty} \lambda_0^j = O(\delta_{0n} + h \tau_n + h^{-1} \delta_n^2) \]
for sufficiently large \( k \). By (6.16), we have
\[
\begin{align*}
D^{1/2} \left( \hat{\beta} - \beta_0 \over \hat{\theta} - \theta_0 \right) &= (\tilde{D}^{-})^{1/2} N_n + D \tilde{D}^{1/2} \left( \hat{\beta} - \beta_0 \over \hat{\theta} - \theta_0 \right) + O(\delta_{0n}^2 + h \tau_n + h^{-1} \delta_n^2) \\
&= (\tilde{D}^{-})^{1/2} N_n + D \tilde{D}^{1/2} \left( \hat{\beta} - \beta_0 \over \hat{\theta} - \theta_0 \right) + o(n^{-1/2}). \quad (6.21)
\end{align*}
\]
The facts that \( n^{1/2} h^3 \to 0 \) and \( n^{1/2} h^{-1} \delta_n^2 \to 0 \) are used in the last step above. It follows from (6.21) that
\[
(\tilde{D} - \tilde{D}^{1/2} D \tilde{D}^{1/2}) \left( \hat{\beta} - \beta_0 \over \hat{\theta} - \theta_0 \right) = N_n + o(n^{-1/2})
\]
or
\[
W_0 \left( \hat{\beta} - \beta_0 \over \hat{\theta} - \theta_0 \right) = N_n + o(n^{-1/2}).
\]
The first part of Theorem 1 follows from the above equation and the central limiting theorem of dependent data, see e.g. [19]. The second part follows immediately from the first part and Theorem 1 of Carroll et al. [4]. □

6.2. Proofs of Lemmas 6.1–6.3

In this subsection, we first give some basic lemmas. Based on these basic lemmas, the proofs of Lemmas 6.1–6.3 are simple algebraic calculations. We will state the main ideas of the calculations without going into the details. However, the details can be obtained on request from the authors. It can also be downloaded from http://www.stat.nus.edu.sg/~staxyc/plsi.pdf
Lemma A.1. Suppose that \( m_1(\theta, x, z) \) and \( \varphi(x, z, v) \) are measurable functions with \( \sup_{\theta \in \Theta} E \{ |m_1(\theta, X, Z)|^r \} < \infty \) for some \( r > 2 \) and \( \sup_{x,z} |m_1(\theta, x, z) - m_1(\theta_0, x, z)| < c |\theta - \theta_0| \). Let \( \varphi_i = \varphi(X_i, Z_i, y_i) \). Assume \( \sup_{\theta \in \Theta, v} E(|\varphi_i|^r) \left| \theta^T X = v \right| < \infty \) and \( \sup_{\theta \in \Theta, u, v} E(|\varphi_i|) \left| \theta^T X_i = u, \theta^T X_i = v \right| < c \) for all \( i > 1 \). Let \( g(v) \) be any function with continuous second-order derivative, \( m(u, v) = g(u) - g(v) - g'(v)(u-v) - g''(v)(u-v)^2/2 \) and \( \xi^k_1 = m(\theta_0^T X_i, \theta_0^T x)z_i (\theta^T X_{i0})^\ell \), where \( z_i \) is any component of \( Z_i \), \( k = 0, 1 \) and \( \ell = 0, 1 \). If (C1) hold, then

\[
\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \{ m_1(\theta, X_i, Z_i) - E m_1(\theta, X_i, Z_i) \} \right| = O(\delta_{0n}).
\]

\[
\sup_{|\theta - \theta_0| < a_n} \left| \frac{1}{n} \sum_{i=1}^n \{ m_1(\theta, X_i, Z_i) - m_1(\theta_0, X_i, Z_i) \} \epsilon_i \right| = O(a_n \delta_{0n}),
\]

where \( a_n \to 0 \) as \( n \to \infty \). If further (C2)–(C5) hold, \( h \sim n^{-\delta} \) with \( 0 < \delta < 1 - 2/r \), then

\[
\sup_{x \in \mathcal{D}} \left| \frac{1}{n} \sum_{i=1}^n \{ H_{b,i} \varphi_i - E(H_{b,i} \varphi_i) \} \right| = O(\delta_{pn}),
\]

\[
\sup_{\theta \in \Theta, x \in \mathcal{D}} \left| \frac{1}{n} \sum_{i=1}^n \{ K_{h,i}^\theta \varphi_i - E(K_{h,i}^\theta \varphi_i) \} \right| = O(\delta_n),
\]

\[
\sup_{|\theta - \theta_0| < a_n} \left| \frac{1}{n} \sum_{i=1}^n \{ K_{h,i}^\theta \xi_{i1}^{k,\ell} - E(K_{h,i}^\theta \xi_{i1}^{k,\ell}) \} \right| = O(\delta_n h^\ell (a_n^2 + h^2)).
\]

The proof of Lemma A.1 is quite standard. The details can be found in [8,16,23].

Lemma A.2. Let \( \varphi_i \) be defined in Lemma A.1 and \( f(x, z, y) \) be the density function of \( (X, Z, y) \). If (C1) and (C5) hold, then

\[
\sup_{\theta \in \Theta} \left| \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \left\{ K_{h,i}^\theta(x_j) \varphi_j - \int K_{h,i}^\theta(x) \varphi(x, z, y) f(x, z, y) \, dx \, dz \, dy \right\} \varepsilon_i \right| = O(\delta_n^2).
\]

Proof. Let \( \Delta_n(\theta) \) be the value in the absolute symbols. By the continuity of \( K_{h,i}^\theta \) in \( \theta \), there are \( n_1 < cn^{2p} \) points \( \theta_{n,1}, \ldots, \theta_{n,n_1} \) in \( \Theta \) such that \( \cup_{k=1}^{n_1} \{ \theta : |\theta - \theta_{n,k}| < h^2 \delta_n^2 \} \supset \Theta \) and

\[
\max_{1 \leq k \leq n_1} \sup_{|\theta - \theta_{n,k}| < h^2 \delta_n^2} |\Delta_n(\theta) - \Delta_n(\theta_{n,k})| = O(\delta_n^2). \tag{6.22}
\]

The Fourier transform \( \phi(s) = \int \exp(isv)K(v) \, dv \) will be used in the following, where \( i \) is the imaginary unit. Thus \( K(v) = \int \exp(-isv)\phi(s) \, ds \). We have

\[
\Delta_n(\theta_{n,k}) = \frac{1}{n^2} h^{-1} \sum_{j=1}^n \sum_{i=1}^n \int \left[ \exp(-is\theta^T_{n,k} X_{ij})/h \right] \varphi_j \left. 
- \int \exp(-is\theta^T_{n,k} X_{i0}/h) \varphi(x, z, y) f(x, z, y) \, dx \, dz \, dy \right\} \phi(s) \, ds \, \varepsilon_i.
\]
Similar to the proof of Lemma A.1, we have almost surely, where \(c_{1182}\)

\[
\begin{align*}
\text{Proof.} & \\
\text{We here only give the details for the first part. If } & \\
\text{uniformly for } & \\
\text{min} & \\
\text{From (C5), we have } & \\
\int & \\
\text{Note that } & \\
\therefore & \\
\text{Therefore, the second part of Lemma A.2 follows from (6.22) to (6.24). } \\
\end{align*}
\]

almost surely, where \(c_8\) and \(c_9\) are constants which do not depend on \(s\). Hence

\[
\max_{1 \leq k \leq n_1} |\Delta_n(\theta_{n,k})| \leq h^{-1} \int c_8 \delta_0 s c_9 \delta_0 n \phi(s) ds = O(h^{-1} \delta_0^2) = O(\delta_n^2). \tag{6.23}
\]

Note that

\[
\sup_{\theta \in \Theta} |\Delta_n(\theta)| \leq \max_{1 \leq k \leq n_1} |\Delta_n(\theta_{n,k})| + \max_{1 \leq k \leq n_1} \sup_{|\theta - \theta_{n,k}| < h^2 \delta_n} |\Delta_n(\theta) - \Delta_n(\theta_{n,k})|. \tag{6.24}
\]

Therefore, the second part of Lemma A.2 follows from (6.22) to (6.24).

For easy of exposition, we abuse \(D\) as the positive support of the \(f(x)\). Let \(d(x, D^c) = \min_{x' \in [\mathbb{R}^p - D]} |x - x'|\) and define bounded functions \(J_0(x), J_0(\nu)\) such that \(J_0(x) = 0\) if \(d(x, \mathbb{R}^p - D) > a_0 b\) and \(J_0(\theta^T x) = 0\) if \(d(\theta^T x, \theta^T (\mathbb{R}^p - D)) > b_0 h\). By the definition, we have

\[
\frac{1}{n} \sum_{j=1}^{n} J_0(X_j) = O(b), \quad \frac{1}{n} \sum_{j=1}^{n} J_0(\theta) = O(h). \tag{6.25}
\]

To cope with the boundary points, we give the following nonuniform rate of convergency.

**Lemma A.3.** Suppose assumptions (C3) and (C5) hold. Then

\[
\begin{align*}
EH_b(x - x)(\theta^T (X - x)/b)^k \{\partial^T (X - x)/b\}^\ell & = v_{k,\ell}^0 f(x) + J_0(x) + O(h), \\
EK_h(\theta^T (X - x))(\theta^T (X - x)/h)^\ell & = \tau_\ell f_0(\theta^T x) + J_0(x) + O(h)
\end{align*}
\]

uniformly for \(\theta, \nu \in \Theta\) and \(x \in D\), where \(v_{k,\ell}^0 \in \int_{\mathbb{R}^p} H(U)(\theta^T U)^k (\theta^T U)^\ell dU\) and \(\tau_\ell = \int K(u) u^\ell dU\). 

**Proof.** We here only give the details for the first part. If \(d(x, D^c) > a_0 b\), we define \(J_0(x) = 0\). From (C5), we have

\[
\int_D H_b(U - x)(\theta^T (U - x)/b)^k \{\partial^T (U - x)/b\}^\ell f(U) dU = \int_{\mathbb{R}^p} H(U)(\theta^T U)^k \{\theta^T U\}^\ell f(x + hU) dU = v_{k,\ell}^0 f(x) + O(h).
\]
If $d(x, D^c) < a_0 b$, we have by (C3)
\[
J_0(x) \overset{\text{def}}{=} \int_D H_b(U - x)\{\theta^T(U - x)/b\}^k \{\theta^T(U - x)/b\}^\ell f(U) dU \\
\leq \int_{\mathbb{R}^p} H(U)\{\theta^T(U)/b\}^k \{\theta^T(U)/b\}^\ell f(x + hU) dU = O(1).
\]
Therefore, the first part of Lemma A.3 follows. \qed

Outline of Proofs of Lemmas 6.1–6.3. The proofs are actually algebraic calculations based on Lemmas A.1–A.3. In the proofs, we need to apply Taylor expansions to the model at $\theta^T_0 x$ and have
\[
y_i = \beta_0^T Z_i + g(\theta^T_0 x) + g'(\theta^T_0 x)\theta^T_0 X_{i0} + \frac{1}{2} g''(\theta^T_0 x)\{\theta^T_0 X_{i0}\}^2 + O(\{\theta^T_0 X_{i0}\}^3) + \varepsilon_i \\
= \beta_0^T Z_i + g(\theta^T_0 x) + g'(\theta^T_0 x)\theta^T_0 X_{i0} + \frac{1}{2} g''(\theta^T_0 x)\{\theta^T X_{i0}\}^2 \\
+ O(\{\theta^T X_{i0}\}^3 + \delta_0^2|\theta^T X_{i0}| + \delta_0^3) + \varepsilon_i.
\]
We then handle each term in the expansions separately. For this purpose, we need to handle terms of summations like
\[
\frac{1}{n} \sum_{i=1}^n K_h(\theta^T X_{i0}) z_i = E\{K_h(\theta^T X_{i0}) z_i\} \\
+ \frac{1}{n} \sum_{i=1}^n \left[ K_h(\theta^T X_{i0}) z_i - E\{K_h(\theta^T X_{i0}) z_i\} \right], \tag{6.26}
\]
and
\[
\frac{1}{n^2} \sum_{i,j=1}^n K_h(\theta^T X_{ij}) z_i = E\{K_h(\theta^T X_{ij}) z_i\} \\
+ \frac{1}{n} \sum_{i,j=1}^n \left[ K_h(\theta^T X_{ij}) z_i - E\{K_h(\theta^T X_{ij}) z_i\} \right], \tag{6.27}
\]
where $z_i$ is defined in Lemma A.1. For (6.26), we can apply Lemma A.1 to the second term on the right-hand side and Lemma A.3 to the first term. Similarly, we can handle the first and the second terms on the right-hand side of (6.27) by Lemmas A.3 and A.2, respectively.

Acknowledgements

We thank two referees and an associate editor for very helpful comments. The research was partially support by the Alexander von Humboldt foundation, FRG R-155-000-048-112 of the National University of Singapore, the National Natural Science Foundation of China (Grant No. 10471061) and Sonderforschungsbereich 373, Berlin.

References