Statistica Sinica **18**(2008), 515-534

DESIGN-ADAPTIVE MINIMAX LOCAL LINEAR REGRESSION FOR LONGITUDINAL/CLUSTERED DATA

Kani Chen[1], Jianqing Fan[2] and Zhezhen Jin[3]

[1]*Hong Kong University of Science and Technology,* [2]*Princeton University and* [3]*Columbia University*

Abstract: This paper studies a weighted local linear regression smoother for longitudinal/clustered data, which takes a form similar to the classical weighted least squares estimate. As a hybrid of the methods of Chen and Jin (2005) and Wang (2003), the proposed local linear smoother maintains the advantages of both methods in computational and theoretical simplicity, variance minimization and bias reduction. Moreover, the proposed smoother is optimal in the sense that it attains linear minimax efficiency when the within-cluster correlation is correctly specified. In the special case that the joint density of covariates in a cluster exists and is continuous, any working within-cluster correlation would lead to linear minimax efficiency for the proposed method.

Key words and phrases: Asymptotic bias, generalized estimating equations, kernel function, linear minimax efficiency, mean squared error, nonparametric curve estimation.

1. Introduction

Recently, nonparametric curve estimation with clustered data has attracted considerable attention. Because of within-cluster correlation, the extension of nonparametric techniques is not straightforward.

The search for simple and reliable nonparametric estimators began with the important work of Lin and Carroll (2000). They showed that correctly accounting for within-cluster correlation does not necessarily yield a better estimator when a specific kernel method is used. Welsh, Lin and Carroll (2002) demonstrated further that a spline estimator whose weight is more global can have a smaller variance than a kernel method, and that the asymptotic variance is smaller when the within-cluster correlation is correctly specified. Wang (2003) proposed a kernel method which has the same variance as that of Welsh, Lin and Carroll (2002). Hence, it enjoys the same merits as that of the spline method when the within-cluster variance is known. Wang (2003) uses seemingly unrelated observations, even though they might induce bias and even, as pointed out in Wang (2003), may not be more accurate than the estimator of Lin and Carroll

(2000). Generally, the performance of these aforementioned estimators is difficult to compare theoretically. In particular, the bias term of the estimator of Wang (2003) can only be expressed as the solution of a Fredholm-type equation; this cannot be easily evaluated, see equation (5) of Lin, Wang, Welsh and Carroll (2004). As a result, mean squared errors are hard to evaluate. Going further, comparisons become moot when the alternative estimators have slightly different assumptions. For example, if the regression function is assumed to have a continuous second derivative at a point, the bias can be made of order $o(h^2)$ with h being the bandwidth. As a result, one can transfer it into variance improvement even without inflating biases. Hence, uniform results on minimax risk play a crucial role in comparing various methods.

In addition to the aforementioned proposals, Yao, Müller and Wang (2005a, b) adapt the functional data analysis techniques to the analysis of longitudinal data. Welsh, Lin and Carroll (2002) provide insightful discussions about spline and kernel methods, including their locality and efficiency. All these estimators are linear in the response variables. While different approaches have their own merits and deal with different aspects of longitudinal data, questions arise naturally about the benchmark performance of the linear estimators. While variance minimization has been the central subject in the related literature, the accuracy of curve estimation is generally measured by the mean squared error (MSE) at a point or by mean integrated squared error (MISE). An important and widely adopted criterion for studying the optimality of smoothing methodologies is linear minimax efficiency. It arises naturally if the linear minimax result of Fan (1992) and Chen (2003) can be extended to the analysis of longitudinal data. The question is important from both theoretical and methodological points of view, as its answer provides useful insights to semiparametric models for longitudinal data (Lin and Carroll (2001, 2006), Lin and Ying (2001) and Fan and Li (2004)).

Our approach here is to combine the ideas of Chen and Jin (2005) and Wang (2003) to adapt to various designs and achieve minimax efficiency among a proper linear class. In the special case that the joint density of covariates in a cluster exists and is continuous, any working within-cluster correlation leads to linear minimax efficiency for the proposed method. Thanks to the theoretical developments of the past decades, major issues about local polynomial smoothing, such as bandwidth selection, kernel function or weighting scheme selection, model complexity and minimax efficiencies, are thoroughly understood (see Fan (1992), Fan and Gijbels (1992, 1995, 1996), Ruppert and Wand (1994), Fan, Heckman and Wand (1995) and Ruppert (1997), among many others). Whether these results can be carried over to clustered data analysis depends critically on the extension. In this regard, the proposed method is a natural extension of classical local polynomial regression smoothing; it has a closed form weighted least squares type expression; it has both computational and theoretical simplicity.

The next section introduces the nonparametric regression model for clustered data and the proposed local linear estimator. Section 3 presents an asymptotic expansion and demonstrates linear minimax efficiency. Section 4 describes analogous results for generalized linear models. Section 5 presents simulation studies, and Section 6 contains some closing remarks.

2. A Nonparametric Regression Model and Local Linear Smoothers

Suppose (X_{ij}, Y_{ij}), $j = 1, \ldots, J_i$, are J_i covariate-response pairs of subject i for $i = 1, \ldots, n$. The marginal nonparametric regression model assumes that

$$Y_{ij} = m(X_{ij}) + \epsilon_{ij}, \qquad j = 1, \ldots, J_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

where $m(\cdot)$ is the unknown function to be estimated, ϵ_{ij} is an error term with conditional mean 0, and finite marginal variance. Let $\mathbf{y}_i = (Y_{i1}, \ldots, Y_{iJ_i})^T$, $\mathbf{x}_i = (X_{i1}, \ldots, X_{iJ_i})^T$, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iJ_i})^T$, and $\Sigma_i = \mathrm{Var}\,\{\boldsymbol{\epsilon}_i | \mathbf{x}_i\}$, a $J_i \times J_i$ matrix. The cluster sizes J_is are assumed to be bounded. To facilitate presentation, we assume that $J_i \equiv J$ throughout the paper. We also assume that the $\{(\boldsymbol{\epsilon}_i, \mathbf{x}_i),\ i \geq 1\}$ are independent and identically distributed, and that the marginal densities of covariates exist.

Our proposed estimate is a delicate consequence of the ideas of Chen and Jin (2005) and Wang (2003). Heuristically, this estimate uses global observations and global variances. The main idea might be illustrated as follows. Suppose $J = 3$ and, for cluster i, (X_{i1}, Y_{i1}) is a local observation to a given point x_0 (i.e., X_{i1} is near x_0), while (X_{i2}, Y_{i2}) and (X_{i3}, Y_{i3}) are not. If the latter two observations are of partial cluster level, i.e., $X_{i2} = X_{i3}$, then $Y_{i2} - Y_{i3}$ has conditional mean 0. Therefore, one can view $Y_{i1} + \lambda(Y_{i2} - Y_{i3}) I(X_{i2} = X_{i3})$ as a candidate to estimate $m(x_0)$, where λ can be chosen by minimizing the variance of the estimator. This idea is basically the same as "the use of control variables" in the simulation literature (Ross (1997)), and is reflected in the local weights (2.3) below.

Here is some notation. The Moore-Penrose generalized inverse of a matrix is adopted throughout paper. The generalized inverse of any symmetric $J \times J$ matrix A is defined to be a symmetric matrix, denoted still by A^{-1}, such that $AA^{-1}A = A$ and $A^{-1}AA^{-1} = A^{-1}$. Specifically, if we let $A = \Gamma\ \mathrm{diag}(\lambda_1, \ldots, \lambda_J)\Gamma^T$ with Γ being an orthonormal matrix, then, $A^{-1} = \Gamma\ \mathrm{diag}(1/\lambda_1, \ldots, 1/\lambda_J)\Gamma^T$, where $1/0$ denotes 0.

Throughout the paper, x_0 is an arbitrary but fixed interior point of the domain of X_{ij}. Let $K(\cdot)$ be a symmetric density function with bounded support which is assumed, without loss of generality, to be $[-1, 1]$. Define $K_h(t) =$

$K(t/h)/h$, where h is a bandwidth. Typical choices of $K(\cdot)$ are, for example, the Epanechnikov kernel $K_0(t) = 0.75(1-t^2)I(|t| \le 1)$ and the uniform kernel $K_1(t) = 0.5I(|t| \le 1)$, where $I(\cdot)$ is the indicator function. Let $\mathbf{K}_i = \text{diag}\{K_h(X_{i1}-x_0), \ldots, K_h(X_{iJ}-x_0)\}$, $\mathbf{I}$ be the $J \times J$ identity matrix and $\mathbf{1}$ be the J-vector with all elements 1. Let $A_i(j) = \{l : X_{il} = X_{ij}\}$ and $|A_i(j)|$ denote the size of the set $A_i(j)$. Define a $J \times J$ matrix

$$\bar{\mathbf{1}}_i = \begin{pmatrix} e_{11} & \cdots & e_{1J} \\ \vdots & \vdots & \vdots \\ e_{J1} & \cdots & e_{JJ} \end{pmatrix} \text{ where } e_{l,j} = \begin{cases} \dfrac{1}{|A_i(j)|} & \text{if } l \in A_i(j) \text{ and } |X_{ij}-x_0| > h, \\ 0 & \text{otherwise.} \end{cases}$$

Observe that $\bar{\mathbf{1}}_i$ is a symmetric $J \times J$ matrix such that, for any function $g(\cdot)$ with $g(t) = 0$ for all $t \in [x_0-h, x_0+h]$,

$$(\mathbf{I}-\bar{\mathbf{1}}_i)g(\mathbf{x}_i) = 0, \quad \bar{\mathbf{1}}_i\bar{\mathbf{1}}_i = \bar{\mathbf{1}}_i \text{ and } (\mathbf{I}-\bar{\mathbf{1}}_i)(\mathbf{I}-\bar{\mathbf{1}}_i) = \mathbf{I}-\bar{\mathbf{1}}_i. \tag{2.2}$$

For any function $g(\cdot)$ defined on real line, we use $g(\mathbf{x}_i)$ to denote $\{g(X_{i1}), \ldots, g(X_{iJ})\}^T$. Set

$$\mathbf{W}_i = \mathbf{K}_i\{(\mathbf{I}-\bar{\mathbf{1}}_i)\mathbf{V}_i(\mathbf{I}-\bar{\mathbf{1}}_i)\}^{-1}, \tag{2.3}$$

where $\mathbf{V}_i$ is the modeled/estimated Σ_i, the conditional covariance matrix of the response $\mathbf{y}_i$ given covariates $\mathbf{x}_i$. We assume that $\mathbf{V}_i$ is measurable to the σ-algebra generated by $\mathbf{x}_i$. When marginal variances are known, modeling/estimating the variance matrix is the same as modeling/estimating the correlation matrix. In other words, $\mathbf{V}_i$ is a working matrix.

The weighted least squares type estimator of $\{m(x_0), m'(x_0)\}^T$ is defined as

$$\begin{pmatrix} \hat{m}(x_0) \\ \hat{m}'(x_0) \end{pmatrix} = \Big(\sum_{i=1}^n \mathbf{P}_i^T\mathbf{W}_i\mathbf{P}_i\Big)^{-1}\sum_{i=1}^n \mathbf{P}_i^T\mathbf{W}_i\mathbf{y}_i, \tag{2.4}$$

where

$$\mathbf{P}_i = \begin{pmatrix} 1 & (X_{i1}-x_0) \\ \vdots & \vdots \\ 1 & (X_{iJ}-x_0) \end{pmatrix}_{J\times 2}.$$

In theory, the analysis of the proposed estimator becomes relatively simple since the rich theoretical results established for local polynomial smoothing can be largely carried over, see the propositions and corollaries in Section 3. More importantly, the estimator is easy to compute.

Remark 1. Chen and Jin (2005) use only local observations (i.e., only Y_{i1} in the illustration at the second paragraph of this section) and weight them by their variances $\{\mathbf{I}_i\mathbf{V}_i\mathbf{I}_i\}^{-1}$, where $\mathbf{I}_i = \text{diag}\{I(|X_{i1}-x_0| \le h), \ldots, |X_{iJ}-x_0| \le h)\}$;

Lin and Carroll (2000) also use local observations but weight them by the global variances $\mathbf{V}_i^{-1}$. The estimator of Chen and Jin (2005) is more accurate than that of Lin and Carroll (2000), but less so than that proposed here. On the other hand, Wang (2003) uses global observation and weights them by global variances. In the illustration at the beginning of Section 2, Wang (2003) essentially uses $Y_{i1} + \lambda_1\{Y_{i2} - \hat{a}(X_{i2})\} + \lambda_2\{Y_{i3} - \hat{a}(X_{i3})\}$ as a datum to estimate $m(x_0)$, where $\hat{a}(\cdot)$ is a preliminary estimator of $m(\cdot)$ and λ_1 and λ_2 are chosen for variance minimization. Such a method can indeed lead to smaller variances. However, as the "control variables" are synthetically created here, the preliminary estimator $\hat{a}(\cdot)$ might induce large bias. In contrast, the proposed estimator not only takes care of variance minimization but avoids possible bias inflation.

3. Asymptotic Properties and Optimality

Let $\{\Omega_k, 1 \le k \le 2^J - 1\}$ be the collections of distinct subsets of $\{1, \ldots, J\}$, except for the empty set. Let $B(x,h)$ denote the interval $[x-h, x+h]$. We assume there exists a $\delta_0 > 0$ such that for all $x \in B(x_0, \delta_0)$ and all $k = 1, \ldots, 2^J - 1$,

$$\begin{aligned}
&P[X_{1j} \in B(x,h),\ \{X_{1j},\ j \in \Omega_k\} \text{ are all equal, and } X_{1l} \ne X_{1j} \text{ for any } l \notin \Omega_k \\
&\text{and } j \in \Omega_k] \\
&\quad = \int_{-h}^{h} f_k(x+t)dt \\
&\quad = P[X_{1j} \in B(x,h) \text{ for all } j \in \Omega_k, \text{ and } X_{1j} \notin B(x,h) \text{ for all } j \notin \Omega_k] + o(h),
\end{aligned}$$

for all $h \in (0, \delta_0)$, where $f_k(\cdot)$, $1 \le k \le 2^J - 1$, are nonnegative continuous functions on $B(x_0, 2\delta_0)$ such that $\sum_{k=1}^{2^J-1} f_k(t) > 0$ for all $t \in B(x_0, 2\delta_0)$.

Remark 2. The above condition is referred to as "the existence of (local) partial density" of the covariates $\mathbf{x}_i$ at x_0, introduced in Chen and Jin (2005). Heuristically, for every $k = 1, \ldots, 2^J - 1$, $f_k(\cdot)$ can be viewed as a partial density of the covariates $\{X_{1j}, j \in \Omega_k\}$ at partial cluster level, i.e., the X_{1j} are equal for all $j \in \Omega_k$. Essentially, the condition ensures that two covariates take values in a small neighborhood of x_0 with a negligible chance unless they are of partial cluster level. This condition features various types of covariates of interest: cluster level covariates, partial cluster level covariates, and covariates with joint density. The marginal density of X_{1l} is the sum of $f_k(\cdot)$ over all Ω_k which contain l, see Chen and Jin (2005) for some special cases.

For every fixed $k = 1, \ldots, 2^J - 1$, let $\mathcal{S}_k(h) = \{X_{1j} \in B(x_0, h)$ for all $j \in \Omega_k$, and $X_{1j} \notin B(x_0, h)$ for all $j \notin \Omega_k\}$, and $\mathcal{S}_k(0) = \{X_{1j} = x_0$ for all $j \in \Omega_k$, and $X_{1j} \ne x_0$ for all $j \notin \Omega_k\}$. Define

$$\begin{aligned}
\xi_k &= E[\mathbf{1}^T\{(\mathbf{I} - \bar{\mathbf{I}}_{10})\mathbf{V}_1(\mathbf{I} - \bar{\mathbf{I}}_{10})\}^{-1}\mathbf{1}|\mathcal{S}_k(0)] \\
\text{and} \quad \bar{\xi}_k &= E[\mathbf{1}^T\{(\mathbf{I} - \bar{\mathbf{I}}_{10})\mathbf{V}_1(\mathbf{I} - \bar{\mathbf{I}}_{10})\}^{-1}\Sigma_1\{(\mathbf{I} - \bar{\mathbf{I}}_{10})\mathbf{V}_1(\mathbf{I} - \bar{\mathbf{I}}_{10})\}^{-1}\mathbf{1}|\mathcal{S}_k(0)],
\end{aligned}$$

where $\bar{\mathbf{1}}_{10}$ is the limit of $\bar{\mathbf{1}}_1$ as $h \to 0$. Moreover, let ξ_{k0} be defined in the same way as ξ_k, except with $\mathbf{V}_1$ replaced by Σ_1. Notice that $\bar{\xi}_k$ with $\mathbf{V}_1$ replaced by Σ_1 equals ξ_{k0}, by (2.2) and the properties of the generalized inverse. Throughout the paper, we assume that elements of $\mathbf{V}_1$ and Σ_1 are continuous functions of $\mathbf{x}_1$, and that the eigenvalues of $\mathbf{V}_1$ and Σ_1 are uniformly bounded, and bounded away from 0.

Proposition 1. *Let $\mathcal{F}_n^X$ denote the σ-algebra generated by $\{\mathbf{x}_i,\ i = 1, \ldots, n\}$. If the condition of the existence of partial density holds, and $h \to 0$, $nh \to \infty$ as $n \to \infty$, then the following results hold.*

(i) *The conditional variance of $\hat{m}(x_0)$ is*

$$\mathrm{Var}\,\{\hat{m}(x_0)|\mathcal{F}_n^X\} = \frac{\gamma(K)\sum_{k=1}^{2^J-1} f_k(x_0)\bar{\xi}_k}{nh[\sum_{k=1}^{2^J-1} f_k(x_0)\xi_k]^2}\{1 + o_P(1)\}, \tag{3.1}$$

where $\gamma(K) = \int K^2(t)dt$.

(ii) *Assume $m(\cdot)$ is twice continuously differentiable. The conditional bias of $\hat{m}(x_0)$ is*

$$\mathrm{Bias}\{\hat{m}(x_0)|\mathcal{F}_n^X\} = \frac{h^2}{2}\gamma_*(K)m''(x_0) + o_P(h^2), \tag{3.2}$$

where $\gamma_*(K) = \int t^2 K(t)dt$.

The following Corollary presents the answers to the problem of minimization of asymptotic variances or MSE. It shows that the best working covariance matrices are the true ones.

Corollary 1. *If the conditions of Proposition 1 hold, then the following results hold.*

(1) *Given a bandwidth and a kernel, the conditional variance of $\hat{m}(x_0)$ is minimized when the working covariance matrices equal the true ones, i.e., $\mathbf{V}_i = \Sigma_i$ for $i \geq 1$, and the minimized asymptotic variance is*

$$\frac{\gamma(K)}{\Big\{nh\sum_{k=1}^{2^J-1} f_k(x_0)\xi_{k0}\Big\}\{1 + o_P(1)\}}. \tag{3.3}$$

(2) *Given a bandwidth, the uniform kernel with the true covariance matrix minimizes the asymptotic conditional variance.*

(3) *Suppose $m''(x_0) \neq 0$. The conditional asymptotic mean squared error is minimized when the working covariance matrices equal the true ones, the*

smooth symmetric nonnegative kernel is the Epanechnikov kernel $K_0(t) = 3/4(1-t^2)I(|t| \le 1)$ *and the bandwidth is*

$$h = \left[\frac{15}{n\{m''(x_0)\}^2 \sum_{k=1}^{2^J-1} f_k(x_0)\xi_{k0}}\right]^{\frac{1}{5}}. \tag{3.4}$$

The minimum asymptotic mean squared error is

$$\frac{3}{4}15^{-\frac{1}{5}}\{m''(x_0)\}^{\frac{2}{5}}\left[\frac{1}{n\sum_{k=1}^{2^J-1} f_k(x_0)\xi_{k0}}\right]^{\frac{4}{5}}. \tag{3.5}$$

Remark 3. One can compare the asymptotic variances of the proposed estimator with those of Chen and Jin (2005) and Lin and Carroll (2000) when the working covariance matrices equal the true ones. In this case, the leading term of the asymptotic variance of the estimator of Chen and Jin (2005) is the same as that given in (3.3) except with ξ_{k0} replaced by $E[\mathbf{1}^T\{\mathbf{I}_1\Sigma_1\mathbf{I}_1\}^{-1}\mathbf{1}|S_k(0)]$, where $\mathbf{I}_1 = \text{diag}\{I(|X_{11}-x_0| \le h), \ldots, I(|X_{1J}-x_0| \le h)\}$, and that of Lin and Carroll (2000) is the same as that given in (3.1), except with $\bar{\xi}_k$ replaced by $E[\mathbf{1}^T\mathbf{I}_1\Sigma_1^{-1}\mathbf{I}_1\Sigma_1\mathbf{I}_1 \Sigma_1^{-1}\mathbf{I}_1\mathbf{1}|S_k(0)]$ and ξ_k replaced by $E[\mathbf{1}^T\mathbf{I}_1\Sigma_1^{-1}\mathbf{I}_1\mathbf{1}|S_k(0)]$. Because $\mathbf{1}^T\{(\mathbf{I}-\bar{\mathbf{I}}_1)\Sigma_1 (\mathbf{I}-\bar{\mathbf{I}}_1)\}^{-1}\mathbf{1} \ge \mathbf{1}^T\{\mathbf{I}_1\Sigma_1\mathbf{I}_1\}^{-1}\mathbf{1}$, it can be shown that, when the working covariance matrices equal the true ones, the asymptotic variances of the present estimator are smaller than or equal to those of Chen and Jin (2005). Moreover, both estimators have asymptotic variances smaller than or equal to those of Lin and Carroll (2000). The proof is similar to that of (3.3) and the details are omitted. These three estimators have the same asymptotic bias, while the estimator of Wang (2003) might induce a sizable bias although its asymptotic variances could be smaller.

We next establish the linear minimax efficiency of the proposed local linear estimators, which shows that our proposed estimator cannot be improved further by using other linear procedures. Define $\mathcal{C}_2 = \{m(\cdot) : |m(x)-m(x_0)-m'(x_0)(x-x_0)| \le C(x-x_0)^2/2\}$, where C is a fixed positive constant. An estimate $\hat{S}$ of $m(x_0)$ is linear if $\hat{S} = \sum_{i=1}^n W_i^{\mathrm{T}} Y_i$ where W_i is of J dimensions and is measurable to $\mathcal{F}_n^X$. Set

$$R_{0,\mathcal{L}}(n, \mathcal{C}_2) = \min_{\hat{S} \text{ is linear}} \max_{m(\cdot)\in\mathcal{C}_2} E[\{\hat{S}-m(x_0)\}^2|\mathcal{F}_n^X],$$

which is the linear minimax risk. Let

$$R_{0,1}^*(n, \mathcal{C}_2) = \min_{\hat{m}(x_0) \text{ is defined in (2.4)}} \max_{m(\cdot)\in\mathcal{C}_2} E[\{\hat{m}(x_0)-m(x_0)\}^2|\mathcal{F}_n^X]$$

be the minimax risk of all local linear smoothers defined in (2.4). Since the local linear smoothers defined in (2.4) are linear estimators, it follows that $R_{0,1}^*(n, \mathcal{C}_2) \ge R_{0,\mathcal{L}}(n, \mathcal{C}_2)$.

Proposition 2. *Assume that the conditions of Proposition* 1 *hold. Then,*

(i) *The estimator* $\hat{m}(x_0)$ *defined in* (2.4) *with the Epanechnikov kernel* $K_0(t) = 3/4(1-t^2)I(|t| \leq 1)$, $\mathbf{V}_i = \Sigma_i$, *and the bandwidth* $h_o = \{15/[nC^2 \sum_{k=1}^{2^J-1} f_k(x_0)\xi_{k0}]\}^{1/5}$ *satisfies*

$$\max_{m(\cdot)\in\mathcal{C}_2} E[\{\hat{m}(x_0) - m(x_0)\}^2 | \mathcal{F}_X^n] = R_{0,\mathcal{L}}(n, \mathcal{C}_2)\{1 + o_P(1)\}.$$

(ii) *Moreover,*

$$R^*_{0,1}(n, \mathcal{C}_2) = R_{0,\mathcal{L}}(n, \mathcal{C}_2)\{1 + o_P(1)\}$$
$$= \frac{3}{4} 15^{-\frac{1}{5}} C^{\frac{2}{5}} \left\{ n \sum_{k=1}^{2^J-1} f_k(x_0)\xi_{k0} \right\}^{-\frac{4}{5}} \{1 + o_P(1)\}.$$

Remark 4. The proposed estimator is locally linear minimax efficient. Under the pointwise linear minimax criterion, the estimator is better than all linear estimators, including those of Wang (2003) and Chen and Jin (2005). It is also noted that the linear form of Wang's (2003) kernel estimator in Lin et al. (2004) requires the use of a 'global bandwidth' rather than a 'local bandwidth'.

As the classical result in the theoretical development of the local polynomial smoothing methodology, Fan (1992) established the linear minimax efficiency of the local linear estimates for nonclustered data; see also Chen (2003) for linear minimax efficiency for local polynomial smoothers of all orders. Such a result demonstrates one of the most important superiorities of the local polynomial smoothing over other smoothing methodologies, as far as the pointwise estimation is concerned. Proposition 2 shows that the local linear smoothers defined in (2.4) for clustered data are indeed the proper generalizations of the classical local linear smoothers to nonclustered data.

The following corollary illustrates the linear minimax efficiency in some special cases.

Corollary 2. *Assume that the conditions of Proposition* 1 *hold, and that the joint density of* $(X_{11}, \ldots, X_{1J})^{\mathrm{T}}$ *exists and is continuous. Let* $\sigma_j^2(x_0) = \mathrm{Var}\,(Y_{1j}|X_{1j} = x_0)$ *and let* $f_j^*(\cdot)$ *be the marginal density of* X_{1j}, $1 \leq j \leq J$.

(i) *The local linear smoother* $\hat{m}(x_0)$ *defined in* (2.4) *is linear minimax efficient when the modelled marginal variances equal the true ones, the kernel is the Epanechnikov kernel, and the bandwidth* $h = [15/\{nC^2 \sum_{j=1}^J f_j^*(x_0)/\sigma_j^2(x_0)\}]^{1/5}$. *Moreover,*

$$R^*_{0,1}(n, \mathcal{C}_2) = R_{0,\mathcal{L}}(n, \mathcal{C}_2)\{1+o_P(1)\} = \frac{3}{4} 15^{-\frac{1}{5}} C^{\frac{2}{5}} \left[n \sum_{j=1}^{J} \frac{f_j^*(x_0)}{\sigma_j^2(x_0)} \right]^{-\frac{4}{5}} \{1+o_P(1)\}.$$

(ii)(Fan (1992)) *In particular, if* $J = 1$*, the local linear smoother* $\hat{m}(x_0)$ *defined in* (2.4) *is linear minimax efficient when the kernel is the Epanechnikov kernel and the bandwidth* $h = [15\sigma_1^2(x_0)/\{nC^2 f_1^*(x_0)\}]^{1/5}$*. Moreover,*

$$R^*_{0,1}(n, \mathcal{C}_2) = R_{0,\mathcal{L}}(n, \mathcal{C}_2)\{1 + o_P(1)\} = \frac{3}{4} 15^{-\frac{1}{5}} C^{\frac{2}{5}} \Big[\frac{\sigma_1^2(x_0)}{n f_1^*(x_0)}\Big]^{\frac{4}{5}} \{1 + o_P(1)\}.$$

Corollary 2 addresses the minimax efficiency under the existence of the joint density. In this case, part (i) shows that only the correct specification of the conditional marginal variances is needed. Specification of the within-cluster correlation, correct or incorrect, is irrelevant to the accuracy of curve estimation. In other words, any working correlation matrix will lead to the same accuracy of curve estimation. Specifically, Suppose $\mathbf{V}_i = \Phi_i C_i \Phi_i$ where Φ_i is the diagonal matrix containing the marginal variances of $\mathbf{y}_i$ and C_i is the working correlation matrix. As long as Φ_i is correctly specified, no matter what working correlation matrix is used, the variance of the curve estimate is minimized. This is mainly due to the minimax formulation and the assumption of existence of joint density that excludes the possibility of any two covariates being equal with positive probability. This is different from the method of Lin and Carroll (2000), where only the working independence correlation would lead to such asymptotic accuracy, and any other working correlation, correct or incorrect, would have an adverse effect on curve estimation. Part (ii) addresses the issue for nonclustered data, which is a classical result of the linear minimax efficiency of local linear smoothers, initially given in Fan (1992). Notice that, in the case of nonclustered data ($J = 1$), modelling of the conditional variance of $E(Y_{11}|X_{11} = x)$ is not necessary because of its continuity in x.

Corollary 3. *If the conditions of Proposition* 1 *hold and there exists a* $\delta > 0$ *such that* $P(X_{11} = \cdots = X_{1J} || X_{1j} - x_0| \leq \delta) = 1$, $1 \leq j \leq J$*, with* $f(x_0)$ *being the* (*common*) *marginal density of the* X_{1j} *at* x_0*, then the local linear smoother* $\hat{m}(x_0)$ *defined in* (2.4) *is linear minimax efficient when the working covariance matrices equal the true ones, the kernel is the Epanechnikov kernel, and the bandwidth* $h = [15/\{nC^2 f(x_0)\xi_*\}]^{1/5}$*. Moreover,*

$$R^*_{0,1}(n, \mathcal{C}_2) = R_{0,\mathcal{L}}(n, \mathcal{C}_2)\{1 + o_P(1)\} = \frac{3}{4} 15^{-\frac{1}{5}} C^{\frac{2}{5}} [n f(x_0)\xi_*]^{-\frac{4}{5}} \{1 + o_P(1)\},$$

where $\xi_* = E(\mathbf{1}^T \Sigma_1^{-1} \mathbf{1} \mid X_{11} = \cdots = X_{1J} = x_0)$.

Many more classical results established for local linear smoothing for nonclustered data can be carried over with only formal modifications.

Remark 5. The aforementioned linear minimax risk is defined for estimating $m(x_0)$, the regression function at a point. It is also possible to define the linear minimax risk under the mean integrated square loss:

$$R_{\mathcal{L}}(n,\mathcal{C}) = \min_{\hat{S}\text{ is linear}} \max_{m(\cdot)\in\mathcal{C}} E\Big[\|\hat{S}-m\|_w^2|\mathcal{F}_n^X\Big],$$

where $\|a(x)\|_w^2 = \int a(x)^2 w(x)dx$ is a weighted L_2-norm for a given weight function $w(\cdot)$ on the support of the marginal density of X, and $\mathcal{C}$ is a function class such as $\mathcal{C} = \{\|m''\|_w \le C\}$. For such a global loss, it would be interesting to study whether Wang's estimator yields minimax efficiency gain.

4. Generalized Linear Models

Suppose the responses Y_{ik} depend on covariates X_{ik} via

$$E(Y_{ik}|X_{ik}=x) = u\{\theta(x)\}, \quad \text{for} \quad k=1,\ldots,J,$$

where $u(\cdot)$ is a known smooth link function, and $\theta(\cdot)$ is the unknown function to be estimated. If $\theta(\cdot)$ is assumed to belong to a parametric family, the parameters have clear interpretation and can be estimated by the parametric GEE method. In the nonparametric setting, $\theta(\cdot)$ is arbitrary except with certain differentiability. Consequently, the above regression model can be equivalently formulated as model (2.1) by letting $m(\cdot) = u\{\theta(\cdot)\}$.

The estimation of $m(\cdot)$ by $\hat{m}(\cdot)$ defined in (2.4) has been addressed in preceding sections. If $u(\cdot)$ is the identity function, the estimation of $\theta(\cdot)$ is the same as that of $m(\cdot)$. In general, the estimator of $\theta(\cdot)$ is naturally obtained as $\hat{\theta}(\cdot) = u^{-1}\{\hat{m}(\cdot)\}$, where $\hat{m}(\cdot)$ is defined in (2.4). Unlike in the parametric setting, it might be $u\{\theta(\cdot)\}$ rather than $\theta(\cdot)$ that is of real interest. However, the function $\theta(\cdot)$ can have an advantage of no constraints on its range, as in the logistic regression problem.

Proposition 1 can be used to obtain asymptotic properties for $\hat{\theta}(x_0)$. By a Taylor expansion,

$$\begin{aligned}\hat{\theta}(x_0)-\theta(x_0) &= \frac{1}{u'\{\theta(x_0)\}}\{\hat{m}(x_0)-m(x_0)\} \\ &\quad -\frac{u''\{\theta(x_0)\}}{2[u'\{\theta(x_0)\}]^3}\{\hat{m}(x_0)-m(x_0)\}^2\{1+o_P(1)\}.\end{aligned}$$

Under some regularity conditions, one can show that,

$$\begin{aligned}\text{Abias}\{\hat{\theta}(x_0)\} &= \left[\frac{1}{u'\{\theta(x_0)\}}\text{bias}\{\hat{m}(x_0)|\mathcal{F}_n^X\} - \frac{u''\{\theta(x_0)\}}{2[u'\{\theta(x_0)\}]^3}\text{Var}\,\{\hat{m}(x_0)|\mathcal{F}_n^X\}\right] \\ &\quad \times\{1+o_P(1)\}, \\ \text{Avar}\{\hat{\theta}(x_0)\} &= [u'\{\theta(x_0)\}]^{-2}\text{Var}\,\{\hat{m}(x_0)|\mathcal{F}_n^X\}\{1+o_P(1)\}, \\ \text{and} \quad \text{AMSE}\{\hat{\theta}(x_0)\} &= [u'\{\theta(x_0)\}]^{-2}\text{MSE}\{\hat{m}(x_0)|\mathcal{F}_n^X\}\{1+o_P(1)\},\end{aligned}$$

where Abias, Avar and AMSE stand for the asymptotic bias, asymptotic varaince and asymptotic MSE, respectively. Applying Proposition 1, one can obtain a closed from expression of the bias, variance and MSE of $\hat{\theta}(x_0)$. Corollary 1 can also be carried over. In particular, the MSE of $\hat{\theta}(x_0)$ is minimized when the modeled variances equal the true ones, the bandwidth is the same as that given in (3.4), and the kernel is the Epanechnikov kernel. The minimized MSE is the same as in (3.5), except with a multiplier $[u'\{\theta(x_0)\}]^{-2}$. Optimalities analogous to Proposition 2 can also be established for $\hat{\theta}(x_0)$ in a similar fashion.

An appealing alternative is to extend our idea along with the local quasi-likelihood method of Fan, Heckman and Wand (1995). We will not pursue this issue further in the present paper.

5. Simulation Study

Simulation studies are carried out to evaluate the performance of the proposed linear smoother. The data are generated from the model

$$y_{ij} = m(x_{ij}) + \epsilon_{ij}, \quad j = 1, 2, 3, 4, \quad i = 1, \ldots, n,$$

where $m(x) = 1 - 60x \exp\{-20x^2\}$, x_{i1} and x_{i3} are independently generated as $U[-1, 1]$, $x_{i2} = x_{i1}$ and $x_{i4} = x_{i3}$, and errors $(\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3}, \epsilon_{i4})$ are generated from the multivariate normal with mean 0, correlation 0.6, and marginal variances 0.04, 0.09, 0.01 and 0.16, respectively.

The sample size n is 150 and the number of simulations is 1,000. The curve estimate $\hat{m}_0(\cdot)$ is computed on the grid points $x_j = -0.8 + 0.016j$, $j = 0, \ldots, 100$, with various global fixed bandwidths. Six different estimation methods are used: the proposed local linear smoother; the local linear method of Chen and Jin (2005); the working independence method of Lin and Carroll (2000); the one-step estimation method of Wang (2003); the estimation method of Wang (2003) with iterations; and the closed-form estimation method of Lin and Carroll (2006). The Epanechnikov kernel was used in all methods.

For each of the grid points, the bias and variance were computed based on the 1,000 simulation runs. Also, the integrated squared error D_i was obtained for the ith simulation, where $D_i = \int_{-0.8}^{0.8} \{m(x) - \hat{m}_i(x)\}^2 dx$ $(i = 1, \ldots, 1,000)$ with the integration replaced by summation over $x_j = -0.08 + 0.016j$ $(j = 0, \ldots, 100)$. Table 1 summarizes the results. In the table, 'Bias' stands for the average of the absolute values of biases over the 101 grid points, 'SD' stands for the average of the sample standard deviations over the 101 grid points and 'MISE' stands for the average of integrated squared errors. The table also reports the relative values of MISE for the four other estimators to that for the proposed estimator: a ratio greater than 1 indicates that the new estimator performs better.

Table 1. Comparison of methods based on 1,000 simulations.

	Proposed estimator			Chen and Jin's estimator			Lin-Carroll's estimator			Wang's first-step estimator			Wang's estimator after iterations			Lin-Carroll's 2006 estimator		
h	Bias	SD	$MISE_1$	Bias	SD	RMISE	Bias	SD	RMISE	Bias	SD	RMISE	Bias	SD	RMISE	Bias	SD	RMISE
0.02	0.027	0.863	25.17	0.036	1.217	2.364	0.041	1.247	2.876	0.029	0.711	0.587	0.027	0.625	0.474	0.031	0.778	0.831
0.03	0.012	0.093	0.045	0.012	0.109	1.545	0.012	0.115	1.215	0.013	0.082	0.756	0.014	0.076	0.674	0.012	0.084	1.049
0.04	0.021	0.046	0.005	0.021	0.049	1.096	0.021	0.056	1.340	0.024	0.041	0.976	0.026	0.039	1.000	0.022	0.041	0.905
0.05	0.033	0.040	0.007	0.033	0.042	1.055	0.034	0.047	1.178	0.038	0.035	1.147	0.040	0.035	1.231	0.034	0.035	0.986
0.06	0.047	0.038	0.011	0.048	0.040	1.043	0.048	0.044	1.110	0.056	0.035	1.264	0.058	0.035	1.368	0.049	0.035	1.050

Bias: average of absolute values of biases at 101 grid points

SD: average of standard deviations at 101 grid points

$MISE_1$: average of integrated squared errors D_i $(i = 1, \cdots, 1,000)$ for proposed method

RMISE: MISE as a multiple of $MISE_1$

All MISE ratios of the estimators of Chen and Jin (2005) and Lin and Carroll (2000) are greater than 1, indicating that the proposed method outperforms the two methods. When the bandwidth is 0.02 or 0.03, the MISE ratios show that Wang's method outperforms the proposed method. This is due to the fact that when the bandwidth is small, the biases in the preliminary estimates in Wang's method are small and her method utilizes more correlation than ours. However, the proposed method outperforms Wang's method as the bandwidth increases. This suggests that when the bandwidth is large, the effect of bias in Wang's method becomes more significant and contributes more to the MSE. It is interesting to notice that the performance of the method of Lin and Carroll (2006) is comparable with that of the proposed method.

6. Concluding Remarks

This paper proposes a weighted least squares type of local linear smoother for clustered data that improves on that of Chen and Jin (2005) and Lin and Carroll (2000), and achieves linear minimax efficiency. The key idea is the proper use of the working covariance matrices so that the resulting estimator has minimal asymptotic variance without bringing in additional bias. The estimator also has the theoretical and computational simplicity of that of Chen and Jin (2005). When a non-identity link function is used to relate the mean response to a function of covariates, the method discussed in Section 4 retains the simplicity of estimation.

This paper only discusses local linear smoothers. The estimator can obviously be extended to local polynomial smoothers of arbitrary orders. However, optimal properties such as linear minimax efficiency seem to be technically nontrivial to establish.

Appendix: proofs

A.1. Proof of Proposition 1

(i). Let $\mathbf{A}_n = \sum_{i=1}^n \mathbf{P}_i^T \mathbf{W}_i \mathbf{P}_i$ and $\mathbf{B}_n = \sum_{i=1}^n \mathbf{P}_i^T \mathbf{W}_i \Sigma_i \mathbf{W}_i^T \mathbf{P}_i$. It is easy to see that

$$\text{Var}\left\{ \begin{pmatrix} \hat{m}(x_0) \\ \hat{m}'(x_0) \end{pmatrix} \Big| \mathcal{F}_n^X \right\} = \mathbf{A}_n^{-1} \mathbf{B}_n \{\mathbf{A}_n^T\}^{-1}.$$

For $0 \le m, l \le 1$, let $a_{m+1,l+1}$ denote the $(m+1,l+1)$-th element of $\mathbf{A}_n$. Let j_k be an element of Ω_k. Recall that $\xi_k = E[\mathbf{1}^T\{(\mathbf{I}-\bar{\mathbf{I}}_{10})\mathbf{V}_i(\mathbf{I}-\bar{\mathbf{I}}_{10})\}^{-1}\mathbf{1}|\mathcal{S}_k(0)]$. With the condition of the existence of partial density at x_0 and change of variables, one can show that

$$\begin{aligned}
E(a_{m+1,l+1}) &= \sum_{i=1}^{n} E\Big[\Big\{(X_{i1}-x_0)^m,\ldots,(X_{iJ}-x_0)^m\Big\} \\
&\quad \times \mathbf{W}_i\Big\{(X_{i1}-x_0)^l,\ldots,(X_{iJ}-x_0)^l\Big\}^{\mathrm{T}}\Big] \\
&= n\sum_{k=1}^{2^J-1} E\Big[(X_{1j_k}-x_0)^{m+l}K_h(X_{1j_k}-x_0)I\{\mathcal{S}_k(h)\}\Big] \\
&\quad \times E\Big[\mathbf{1}^T\{(\mathbf{I}-\bar{\mathbf{1}}_{10})\mathbf{V}_1(\mathbf{I}-\bar{\mathbf{1}}_{10})\}^{-1}\mathbf{1}|\mathcal{S}_k(0)\Big]\{1+o(1)\} \\
&= n\sum_{k=1}^{2^J-1} \xi_k \int (x-x_0)^{m+l} f_k(x)\frac{1}{h}K\Big(\frac{x-x_0}{h}\Big)dx\{1+o(1)\} \\
&= nh^{m+l}\sum_{k=1}^{2^J-1} f_k(x_0)\xi_k\Big\{\int t^{m+l}K(t)dt+o(1)\Big\}.
\end{aligned}$$

It is analogous to show that $\{\mathrm{Var}\,(a_{m+1,l+1})\}^{1/2}=o(nh^{m+l})$. Then,

$$a_{m+1,l+1}=nh^{m+l}\sum_{k=1}^{2^J-1} f_k(x_0)\xi_k\Big\{\int t^{m+l}K(t)dt+o_P(1)\Big\}$$

since $a_{m+1,l+1}=E(a_{m+1,l+1})+O_P[\{\mathrm{Var}\,(a_{m+1,l+1})\}^{1/2}]$. Therefore,

$$\mathbf{A}_n=n\Big\{\sum_{k=1}^{2^J-1} f_k(x_0)\xi_k\Big\}\begin{pmatrix}1 & 0\\ 0 & h^2\int t^2K(t)dt\end{pmatrix}\{1+o_P(1)\}$$

by the symmetry of $K(\cdot)$. With a similar calculation, it follows that

$$\mathbf{B}_n=nh^{-1}\Big\{\sum_{k=1}^{2^J-1} f_k(x_0)\bar{\xi}_k\Big\}\begin{pmatrix}\int tK^2(t)dt & h\int tK^2(t)dt\\ h\int tK^2(t)dt & h^2\int t^2K^2(t)dt\end{pmatrix}\{1+o_P(1)\}.$$

Therefore,

$$\begin{aligned}
\mathrm{Var}\,\{\hat{m}(x_0)|\mathcal{F}_n^X\} &= (1\quad 0)\mathbf{A}_n^{-1}\mathbf{B}_n\{\mathbf{A}_n^T\}^{-1}\begin{pmatrix}1\\0\end{pmatrix} \\
&= \frac{\gamma(K)\sum_{k=1}^{2^J-1} f_k(x_0)\bar{\xi}_k}{nh[\sum_{k=1}^{2^J-1} f_k(x_0)\xi_k]^2}\{1+o_P(1)\}.
\end{aligned}$$

(ii). By the Taylor expansion, the conditional bias of $\hat{\boldsymbol{\beta}}$ is

$$E\left\{\begin{pmatrix}\hat{m}(x_0)\\ \hat{m}'(x_0)\end{pmatrix}\Big|\mathcal{F}_n^X\right\}-\begin{pmatrix}m(x_0)\\ m'(x_0)\end{pmatrix}=\mathbf{A}_n^{-1}\sum_{i=1}^n \mathbf{P}_i^T\mathbf{W}_i\left\{m(\mathbf{x}_i)-\mathbf{P}_i\begin{pmatrix}m(x_0)\\ m'(x_0)\end{pmatrix}\right\}$$

$$=\mathbf{A}_n^{-1}\sum_{i=1}^n \mathbf{P}_i^T\mathbf{W}_i\{(X_{i1}-x_0)^2,\ldots,(X_{iJ}-x_0)^2\}^{\mathrm{T}}$$

$$\times\{\frac{m''(x_0)}{2}+o_P(1)\}.$$

Similar to the asymptotic expansion of $\mathbf{A}_n$, one can also show that

$$\sum_{i=1}^n \mathbf{P}_i^T\mathbf{W}_i\{(X_{i1}-x_0)^2,\cdots,(X_{iJ}-x_0)^2\}^{\mathrm{T}}$$

$$=nh^2\sum_{k=1}^{2^J-1}\{f_k(x_0)\xi_k\}\left\{\begin{pmatrix}\int t^2K(t)dt\\ \int t^3K(t)dt\end{pmatrix}+o_P(1)\right\}.$$

After some algebra, it can be shown that the conditional bias is

$$E\{\hat{m}(x_0)|\mathcal{F}_n^X\}-m(x_0)=\frac{h^2m''(x_0)}{2}\gamma_*(K)+o_P(h^2).$$

A.2. Proof of Corollary 1

To show that, for any given bandwidth h and kernel K, the asymptotic variance is minimized when the modeled correlation equals the true correlation, it suffices to show that $\sum_{k=1}^{2^J-1} f_k(x_0)\bar{\xi}_k/\ \{\sum_{k=1}^{2^J-1} f_k(x_0)\xi_k\}^2$ is minimized when $\mathbf{V}_1=\Sigma_1$. Recall that $\mathbf{1}=(1,\ldots,1)^T$. Let $\mathbf{b}=\{(\mathbf{I}-\bar{\mathbf{1}}_{10})\Sigma_1(\mathbf{I}-\bar{\mathbf{1}}_{10})\}^{1/2}\{(\mathbf{I}-\bar{\mathbf{1}}_{10})\mathbf{V}_1(\mathbf{I}-\bar{\mathbf{1}}_{10})\}^{-1}\mathbf{1}$. Observe that $\sum_{k=1}^{2^J-1} f_k(x_0)\bar{\xi}_k=\sum_{k=1}^{2^J-1} f_k(x_0)E\{\mathbf{b}^T\mathbf{b}|\mathcal{S}_k(0)\}$ and $\sum_{k=1}^{2^J-1} f_k(x_0)\xi_k=\sum_{k=1}^{2^J-1} f_k(x_0)E\{\mathbf{1}^T\{(\mathbf{I}-\bar{\mathbf{1}}_{10})\Sigma_1(\mathbf{I}-\bar{\mathbf{1}}_{10})\}^{-1/2}\mathbf{b}|\mathcal{S}_k(0)\}$. Then, $(\sum_{k=1}^{2^J-1} f_k(x_0)\bar{\xi}_k)\ /\ [\sum_{k=1}^{2^J-1} f_k(x_0)\xi_k]^2\geq\{\sum_{k=1}^{2^J-1} f_k(x_0)\xi_{k0}\}^{-1}$ by the Cauchy-Schwartz inequality, in which equality holds when $\mathbf{b}=\{(\mathbf{I}-\bar{\mathbf{1}}_{10})\Sigma_1(\mathbf{I}-\bar{\mathbf{1}}_{10})\}^{1/2}\mathbf{1}$. This is certainly implied by $\mathbf{V}_1=\Sigma_1$. This proves that the true variance always leads to the minimum asymptotic variance for any given bandwidth and kernel function.

For any given bandwidth, the variance minimizing kernel is the uniform kernel simply because $\gamma(K)$ is minimized when K is the uniform kernel. This is parallel to the classical result for local polynomial smoothers for nonclustered data; see Fan and Gijbels (1996, p.75)). The next claim about minimization of MSE also follows from the same classical result; see e.g., Fan (1992). We omit the details.

A.3. Proof of Proposition 2.

Part (i) follows from part (ii) by applying Corollary 1. We only prove part (ii), which consists of four steps.

Step 1. Following Corollary 1, it can be shown via a calculation similar to (A.4) that

$$R^*_{0,1}(n, \mathcal{C}_2) \le \frac{3}{4} 15^{-\frac{1}{5}} C^{\frac{2}{5}} \Big[\frac{1}{n\sum_{k=1}^{2^J-1}\xi_{k0} f_k(x_0)}\Big]^{\frac{4}{5}} \{1 + o_P(1)\}. \tag{A.1}$$

Since $R^*_{0,1}(n, C_2) \ge R_{0,\mathcal{L}}(n, \mathcal{C}_2)$, we have

$$R_{0,\mathcal{L}}(n, \mathcal{C}_2) \le \frac{3}{4} 15^{-\frac{1}{5}} C^{\frac{2}{5}} \Big[\frac{1}{n\sum_{k=1}^{2^J-1}\xi_{k0} f_k(x_0)}\Big]^{\frac{4}{5}} \{1 + o_P(1)\}. \tag{A.2}$$

Step 2. Consider linear estimates of the form $\sum_{i=1}^n W_i^T \mathbf{y}_i$, where $W_i = (w_{i1}, \ldots, w_{iJ})^T$ is $\mathcal{F}_n^X$-measurable and

$$\bar{\mathbf{1}}_i W_i = 0. \tag{A.3}$$

Throughout the proof, h is chosen so that it converges to 0 slowly enough, e.g., $h = 1/\log(n)$. Define a restricted linear minimax risk as

$$R_{res}(n, \mathcal{C}_2) = \min_{\substack{\text{linear estimates}\\ \text{satisfying (A.3)}}} \max_{m(\cdot)\in\mathcal{C}_2} E\Big[\Big\{\sum_{i=1}^n W_i^T \mathbf{y}_i - m(x_0)\Big\}^2 |\mathcal{F}_n^X\Big].$$

Notice that (A.3) ensures $W_i^T \Sigma_i W_i = W_i^T (\mathbf{I} - \bar{\mathbf{1}}_i)\Sigma_i(\mathbf{I} - \bar{\mathbf{1}}_i)W_i$. Consequently

$$E\Big[\Big\{\sum_{i=1}^n W_i^T \mathbf{y}_i - m(x_0)\Big\}^2 |\mathcal{F}_n^X\Big] \ge \frac{m(x_0)^2}{1 + \sum_{i=1}^n a_i}, \tag{A.4}$$

where $a_i = m(\mathbf{x}_i)^T \{(\mathbf{I} - \bar{\mathbf{1}}_i)\Sigma_i(\mathbf{I} - \bar{\mathbf{1}}_i)\}^{-1} m(\mathbf{x}_i)$. Then,

$$R_{res}(n, \mathcal{C}_2) \ge \max_{m(\cdot)\in\mathcal{C}_2} \frac{m(x_0)^2}{1 + \sum_{i=1}^n a_i}. \tag{A.5}$$

In (A.4), the equality holds when

$$W_i = \frac{m(x_0)}{1 + a_i} \{(\mathbf{I} - \bar{\mathbf{1}}_i)\Sigma_i(\mathbf{I} - \bar{\mathbf{1}}_i)\}^{-1} m(\mathbf{x}_i). \tag{A.6}$$

It follows from (2.2) that W_i given in (A.6) indeed satisfies (A.3).

Step 3. Set $m(x) = Ch_o^2/2[1 - \{(x - x_0)/h_o\}^2]_+ = 2/3Ch_o^2 K_0\{(x - x_0)/h_o\}$. Then $h_o < h$ for large n. Using the condition of the existence of partial density, we can write

$$E(\sum_{i=1}^{n} a_i) = n\sum_{k=1}^{2^J-1}\int_{x_0-h}^{x_0-h} f_k(x)m^2(x)E[\mathbf{1}^T\{(\mathbf{I}-\bar{\mathbf{1}}_{10})\Sigma_1(\mathbf{I}-\bar{\mathbf{1}}_{10})\}^{-1}\mathbf{1}|\mathcal{S}_k(0)]dx$$
$$\times\{1+o(1)\}$$
$$= n\sum_{k=1}^{2^J-1}\xi_{k0}\int_{x_0-h_o}^{x_0-h_o} f_k(x)m^2(x)dx\{1+o(1)\}$$
$$= n\sum_{k=1}^{2^J-1}\xi_{k0}f_k(x_0)\int_{-1}^{1}\frac{4}{9}C^2h_o^4K_0^2(t)h_o dt\{1+o(1)\}$$
$$= n\frac{4}{15}C^2h_o^5\sum_{k=1}^{2^J-1}\xi_{k0}f_k(x_0)\{1+o(1)\}.$$

It also can be shown that $\sum_{i=1}^n a_i = E(\sum_{i=1}^n a_i)\{1+o_P(1)\}$. By straightforward calculation,

$$\frac{m^2(x_0)}{1+4nC^2h_o^5\sum_{k=1}^{2^J-1}\frac{\xi_{k0}f_k(x_0)}{15}} = \frac{C^2h_o^4}{20}\{1+o(1)\}$$
$$= \frac{3}{4}15^{-\frac{1}{5}}C^{\frac{2}{5}}\Big[\frac{1}{n\sum_{k=1}^{2^J-1}\xi_{k0}f_k(x_0)}\Big]^{\frac{4}{5}}\{1+o(1)\}.$$

It then follows from (A.5) that

$$R_{res}(n,\mathcal{C}_2) \geq \frac{3}{4}15^{-\frac{1}{5}}C^{\frac{2}{5}}\Big[\frac{1}{n\sum_{k=1}^{2^J-1}\xi_{k0}f_k(x_0)}\Big]^{\frac{4}{5}}\{1+o_P(1)\}. \tag{A.7}$$

Step 4. We show that $R_{res}(n,\mathcal{C}_2) = R_{0,\mathcal{L}}(n,\mathcal{C}_2)\{1+o_P(1)\}$. It is clear that $R_{res}(n,\mathcal{C}_2) \geq R_{0,\mathcal{L}}(n,\mathcal{C}_2)$. If a linear estimate $\sum_{i=1}^n W_i^T\mathbf{y}_i = \sum_{i=1}^n\sum_{j=1}^n w_{ij}Y_{ij}$ is linear minimax efficient, then it can be shown that

$$\sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij} = 1,\ \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}(X_{ij}-x_0) = 0 \text{ and } \sup_{1\leq i\leq n}\sup_{1\leq j\leq J}|w_{ij}| = o_P(n^{-\frac{2}{5}}). \tag{A.8}$$

For any $m(\cdot)\in\mathcal{C}_2$, let $r_m(t) = m(t)-m(x_0)-m'(x_0)(t-x_0)$. Consider $m_*(x) = C/2(x-x_0)^2I(|x-x_0|>h)\text{sgn}\{\sum_{i=1}^n\sum_{j=1}^J w_{ij}I(x=X_{ij})\}$ where $\text{sgn}(\cdot)$ is the sign function. Clearly $m_*(\cdot)\in\mathcal{C}_2$ and $\bar{\mathbf{1}}_i m_*(\mathbf{x}_i) = m_*(\mathbf{x}_i)$. Thus, with probability 1,

$$\max_{m(\cdot)\in\mathcal{C}_2} E\Big[\Big\{\sum_{i=1}^n W_i^T\mathbf{y}_i - m(x_0)\Big\}^2|\mathcal{F}_n^X\Big] \geq \Big\{\sum_{i=1}^n W_i^T m_*(\mathbf{x}_i)\Big\}^2$$
$$\geq \frac{C^2h^4}{4}\Big(\sum_{i=1}^n |W_i^T\bar{\mathbf{1}}_i|\mathbf{1}\Big)^2.$$

Therefore, $(\sum_{i=1}^n |W_i^T\bar{\mathbf{1}}_i|\mathbf{1})^2 = O_P(n^{-4/5})$. This and (A.8) ensure that

$$\sum_{i=1}^n W_{i,1}^T\Sigma_i W_{i,2} = \sum_{i=1}^n W_{i,1}^T\Sigma_i\bar{\mathbf{1}}_i W_i \leq o_P(n^{-\frac{2}{5}})\sum_{i=1}^n \mathbf{1}^T|\bar{\mathbf{1}}_i W_i| = o_P(n^{-\frac{4}{5}}),$$

where $W_{i,1} = (\mathbf{I}-\bar{\mathbf{1}}_i)W_i$ and $W_{i,2} = \bar{\mathbf{1}}_i W_i$. For every given $m(\cdot)$, the bias of $\sum_{i=1}^n W_{i,1}^T r_m(\mathbf{x}_i)$ is irrelevant to the value of $r_m(\cdot)$ defined outside the interval $[x_0-h, x_0+h]$, by (2.2). Therefore,

$$\begin{aligned}
&\max_{m(\cdot)\in\mathcal{C}_2} E\Big[\Big\{\sum_{i=1}^n W_i^T\mathbf{y}_i - m(x_0)\Big\}^2|\mathcal{F}_n^X\Big]\\
&= \max_{m(\cdot)\in\mathcal{C}_2}\Big\{\sum_{i=1}^n W_i^T r_m(\mathbf{x}_i)\Big\}^2 + \sum_{i=1}^n W_i^T\Sigma_i W_i\\
&\geq \max_{m(\cdot)\in\mathcal{C}_2}\Big[\Big\{\sum_{i=1}^n W_{i,1}^T r_m(\mathbf{x}_i)\Big\}^2 + 2\sum_{i=1}^n W_{i,1}^T r_m(\mathbf{x}_i)\sum_{i=1}^n W_{i,2}^T r_m(\mathbf{x}_i)\\
&\quad + \Big\{\sum_{i=1}^n W_{i,2}^T r_m(\mathbf{x}_i)\Big\}^2\Big] + \sum_{i=1}^n W_{i,1}^T\Sigma_i W_{i,1} + 2\sum_{i=1}^n W_{i,1}^T\Sigma_i W_{i,2} + \sum_{i=1}^n W_{i,2}^T\Sigma_i W_{i,2}\\
&\geq \max_{m(\cdot)\in\mathcal{C}_2}\Big\{\sum_{i=1}^n W_{i,1}^T r_m(\mathbf{x}_i)\Big\}^2 + \sum_{i=1}^n W_{i,1}^T\Sigma_i W_{i,1} + o_P(n^{-\frac{4}{5}})\\
&\geq R_{res}(n,\mathcal{C}_2) + o_P(n^{-\frac{4}{5}}).
\end{aligned}$$

It then follows from (A.7) that

$$R_{0,\mathcal{L}}(n,\mathcal{C}_2) \geq \frac{3}{4}15^{-\frac{1}{5}}C^{\frac{2}{5}}\Big[\frac{1}{n\sum_{k=1}^{2^J-1}\xi_{k0}f_k(x_0)}\Big]^{\frac{4}{5}}\{1+o_P(1)\}. \tag{A.9}$$

The desired result (3.6) follows from (A.1), (A.2) and (A.9). The proof is complete.

A.4. Proofs of Corollaries 2 and 3

Corollaries 2 and 3 are special cases of Proposition 2. We omit the details of the proof.

Acknowledgement

Chen's research was supported partially by a Hong Kong RGC grant, Fan's research was supported partially by NIH grant R01 GM072611 and NSF FRG grant DMS-0354223, and Jin's research was supported partially by the NSF career award DMS-0134431. We thank an associate editor and a referee for their helpful comments.

References

Chen, K. (2003). Linear minimax efficiency for local polynomial regression smoothers. *J. Nonparametr. Stat.* **15**, 343-353.

Chen, K. and Jin, Z. (2005). Local polynomial regression analysis for clustered data. *Biometrika* **92**, 59-74.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87**, 998-1004.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20**, 2008-2036.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Ann. Statist.* **20**, 2008-2036.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.

Fan, J., Heckman, N. E. and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90**, 141-150.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modelling in longitudinal data analysis. *J. Amer. Statist. Assoc.* **99**, 710–723.

Lin, X. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Amer. Statist. Assoc.* **95**, 520-534.

Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *J. Amer. Statist. Assoc.* **96**, 1045-1056.

Lin, X. and Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *J. Roy. Statist. Soc. Ser. B* **68**, 69-88.

Lin, X., Wang, N., Welsh, A. and Carroll, R. J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data. *Biometrika* **91**, 177-193.

Lin, D. Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data (with discussions). *J. Amer. Statist. Assoc.* **96**, 103-126.

Ross, S. M. (1997). *Simulation*. Second Edition. Academic Press, Inc., San Diego, CA.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Amer. Statist. Assoc.* **92**, 1049-1062.

Ruppert, D. and Wand, M. P. (1994). Multivariate weighted least squares regression. *Ann. Statist.* **22**, 1346-1370.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* **90**, 43-52.

Welsh, A. H., Lin, X. and Carroll, R. J. (2002). Marginal longitudinal nonparametric regression: locality and efficiency of spline and kernel methods. *J. Amer. Statist. Assoc.* **97**, 482-493.

Yao, F., Müller, H. G. and Wang, J. L. (2005a). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100**, 577-590.

Yao, F., Müller, H. G. and Wang, J. L. (2005b). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33**, 2873-2903.

Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.

E-mail: makchen@ust.hk

Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544

E-mail: jqfan@princeton.edu

Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032

E-mail: zjin@biostat.columbia.edu

(Received November 2005; accepted November 2006)