Correlation and Marginal Longitudinal Kernel Nonparametric Regression

Oliver B. Linton
Enno Mammen
Xihong Lin
Raymond J. Carroll

ABSTRACT We consider nonparametric regression in a marginal longitudinal data framework. Previous work ([3]) has shown that the kernel nonparametric regression methods extant in the literature for such correlated data have the discouraging property that they generally do not improve upon methods that ignore the correlation structure entirely. The latter methods are called working independence methods. We construct a two-stage kernel-based estimator that asymptotically uniformly improves upon the working independence estimator. A small simulation study is given in support of the asymptotics.

1 INTRODUCTION

Nonparametric longitudinal regression in the marginal model using kernel methods has been investigated by a number of authors, see [2], [8], [9], [10] and [11], among others. Paper [8] estimates the covariance matrix of the correlated observations and use this in their kernel construction of the nonparametric regression estimate. The other papers effectively ignore the correlation structure entirely and "pretend" that the data are really independent, this being the so–called "working independence" method. Both [3] and [6] provided theoretical evidence in support of the working independence method. In fact, they showed that for many situations and different methods of kernel estimation, the working independence method is most efficient in terms of mean squared error. That is, for the kernel methods proposed in the literature, it is generally better to ignore the correlation structure entirely.

The purpose of this paper is to construct a kernel–type method that can take advantage of the correlations among the data. The method is a simple modification, and generalization to an arbitrary covariance matrix, of a method proposed by [6]. The resulting estimator is asymptotically more efficient than the working independence estimator.

The model for this paper is as follows. Suppose that there are $j = 1, ..., J$ time points, with responses Y_{ij} and covariates X_{ij}. Our basic assumption is that $E(Y_{ij}|X_{ij}) = E(Y_{ij}|X_{i1}, ..., X_{iJ})$, see [5]. Writing $\mathbf{Y}_i = (Y_{i1}, ..., Y_{iJ})^{\mathrm{T}}$ and similarly for $\mathbf{X}_i$ and $\mathbf{m}(\mathbf{X}_i) = m(X_{i1}), \cdots, m(X_{iJ})$, our model is that for an unknown function $m(\cdot)$,

$$\mathbf{Y}_i = \mathbf{m}(\mathbf{X}_i) + \mathbf{\Sigma}^{1/2}\boldsymbol{\epsilon}_i = \mathbf{m}(\mathbf{X}_i) + \mathbf{U}_i, \tag{1.1}$$

where $E(\mathbf{U}_i|\mathbf{X}_i) = 0$, $\mathrm{cov}(\mathbf{U}_i|\mathbf{X}_i) = \mathbf{\Sigma} = \{s_{jk}\}_{j,k}$, $\mathrm{cov}(\boldsymbol{\epsilon}_i|\mathbf{X}_i) = I_J$ and $\mathbf{\Sigma}^{1/2}$ is the symmetric square root of $\mathbf{\Sigma}$. Let $\mathbf{\Omega} = \mathbf{\Sigma}^{-1/2} = \{\omega_{jk}\}_{j,k}$ and let $\mathbf{\Lambda} = \mathrm{diag}(\mathbf{\Omega})$. Finally, let the marginal density of $(X_{ij})_i$ be $f_j(\cdot)$.

Because $\mathbf{\Sigma}$ can be estimated at parametric rates, which are faster than the rates available in nonparametric regression, for purposes of asymptotic theory we may assume without loss of generality that $\mathbf{\Sigma}$ is known. Since $\mathbf{\Sigma}$ is a $J \times J$ matrix, a root–n consistent estimate of it can be obtained by computing the sample covariance matrix of the residuals from a working independence smooth of the data.

The main idea of the two-stage estimator is to construct a linear transformation of $\mathbf{Y}$ and $\mathbf{m}(\mathbf{X})$ that has mean $\mathbf{m}(\mathbf{X})$ and diagonal covariance matrix, and to then apply working independence methods to this transformation. Specifically, for any function $g(\cdot)$, define

$$\boldsymbol{Z}_i(g) = \mathbf{Y}_i + \mathbf{\Lambda}^{-1}(\mathbf{\Omega} - \mathbf{\Lambda})\{\mathbf{Y}_i - \mathbf{g}(\mathbf{X}_i)\}. \tag{1.2}$$

Note that since $\boldsymbol{Z}_i(m) = \mathbf{m}(\mathbf{X}_i) + \mathbf{\Lambda}^{-1}\boldsymbol{\epsilon}_i$, (1.2) is one version of the required transformation.

The method we propose consists of two steps.

1. First assume working independence and fit the function $\widehat{m}_p(\cdot, h_p)$ using a kernel local polynomial method with bandwidth h_p and with the weights s_{jj}^{-1}. Without loss of generality the kernel function $K(\cdot)$ is a density function with variance one.

2. At the second step, run a local linear regression using working independence of $Z_{ij}\{\widehat{m}_p(\cdot, h_p)\}$ on X_{ij} with bandwidth h and with weights ζ_j, calling the result $\widehat{m}(\cdot, h, h_p)$. The weights ζ_j are arbitrary, but we show below that the optimal choice is $\zeta_j = \omega_{jj}^2$.

The method proposed here is essentially the same as that proposed by [6], with two exceptions: (a) our covariance structure is general, while theirs is restricted to the simple 1–way random effects model; and (b) more crucially, we allow the bandwidth at the first step to be different from the bandwidth at the second step. If one forces the two bandwidths to be identical, then as in [6], while the estimator often has a smaller variance than the working independence estimator, it has an extremely complex bias expression, and it need not have smaller mean squared error than the working independence estimator.

The paper is organized as follows. In Section 2, we state the main results. Section 3 gives the results of simulations that demonstrate that our method improves upon working independence in non–asymptotic situations. Section 4 gives concluding remarks. The proofs of the main results are sketched in a technical appendix.

2 Main Results

In this section, we state the main results. As a matter of notation, let $g^{(j)}(\cdot)$ be the jth derivative of a function $g(\cdot)$, and define $K_h(v) = h^{-1}K(v/h)$. All methods are based on working independence *local linear* kernel regression of some response R_{ij} on regressors X_{ij} with some weights W_{jj}. By this we mean that the function estimated at any value x is obtained by a weighted linear regression of the R_{ij} on the X_{ij} with weights $K_h(X_{ij} - x)W_{jj}$.

We state the results without conditions because they are standard, e.g., the regression function and density functions are thrice continuously differentiable, the density functions are positive on their support, etc. The essential condition is that (1.10) in the Appendix (Section 5.1) holds uniformly: conditions which would allow this are of the type used by [4].

By results in [3], to first order, the local linear regression weighted working independence estimator satisfies

$$\begin{aligned} \mathrm{bias}\{\widehat{m}_p(\cdot, h_p)\} &= (h_p^2/2)m^{(2)}(x); & (1.3) \\ \mathrm{var}\{\widehat{m}_p(\cdot, h_p)\} &= (nh_p)^{-1}\int K^2(x)dx \left\{\sum_{j=1}^{J} s_{jj}^{-1} f_j(x)\right\}^{-1}. & (1.4) \end{aligned}$$

Result 1 Assume that $h_p^3 = o(n^{-1/2})$ that $(nh_p)^{-1} = o(n^{-1/2})$, and that $h \propto n^{-1/5}$. Then to first order, the bias and variance of $\widehat{m}(\cdot, h, h_p)$ are given as follows:

$$\begin{aligned} \mathrm{bias} &= (h^2/2)m^{(2)}(x) & (1.5) \\ &\quad -(h_p^2/2)\left\{\sum_{j=1}^{J}\zeta_j f_j(x)\right\}^{-1} \\ &\quad \times \sum_{j=1}^{J}\sum_{k\neq j}^{J}\zeta_j(\omega_{jk}/\omega_{jj})f_j(x)E\left\{m^{(2)}(X_{ik})|X_{ij}=x\right\}; \\ \mathrm{var} &= (nh)^{-1}\int K^2(x)dx\sum_{j=1}^{J}\zeta_j^2 f_j(x)/\omega_{jj}^2\left\{\sum_{j=1}^{J}\zeta_j f_j(x)\right\}^{-2}. & (1.6) \end{aligned}$$

Let $\boldsymbol{e}_J$ be a vector of J–ones. Consider the variance components model $\boldsymbol{\Sigma} = \sigma_1^2\boldsymbol{e}_J\boldsymbol{e}_J^{\mathrm{T}} + \sigma^2 I_J$ studied by [6]. They studied the case that $h_p = h$

and $\zeta_j = \omega_{jj}^2 \equiv \omega$, the last equivalence being a consequence of the form of $\mathbf{\Sigma}$. They derived (1.5)–(1.6) in this special case. under some conditions, they showed that the variance (1.6) of their estimator was smaller than the variance of the working independence estimator (1.4). However, the complex nature of the bias expression (1.5) of their estimator as compared to that of the working independence estimator (1.3) meant that they could not show that the mean squared error of their estimator dominated that of the working independence estimator in any meaningful way.

Our next result shows that by undersmoothing the preliminary estimate, the two-stage estimator has the same simple, well–known bias expression as the working independence estimator, and for any $\mathbf{\Sigma}$ it has smaller variance, thus showing dominance in the mean squared error sense.

Result 2 Suppose that $h_p/h \to 0$ and that (1.6) is minimized by taking $\zeta_j = \omega_{jj}^2$. Then, to first order, the bias and variance of $\widehat{m}(x, h, h_p)$ are given as follows:

$$\text{bias} = (h^2/2)m^{(2)}(x); \tag{1.7}$$

$$\text{var} = (nh)^{-1} \int K^2(x)dx \left\{ \sum_{j=1}^{J} \omega_{jj}^2 f_j(x) \right\}^{-1}. \tag{1.8}$$

Since the bias expressions (1.3) and (1.7) are the same, comparing our method with the working independence estimator reduces to comparing the variance expressions (1.4) and (1.8). We show in the Appendix (Section 5.2) that for any arbitrary covariance structure $\mathbf{\Sigma}$, our method always has smaller variance, i.e., $\text{var}\{\widehat{m}(x; h, h_p)\} \le \text{var}\{\widehat{m}_p(x; h)\}$, and hence smaller mean squared error. Our two-stage estimator hence is uniformly asymptotically dominant in terms of mean squared error compared to the regular kernel estimator.

If we compute the mean squared error at the optimal bandwidth for the working independence estimator using (1.3)–(1.4), and compare it to the mean squared error at the optimal bandwidth for the two–stage estimator using (1.7)–(1.8), we see that the asymptotic mean squared error efficiency of the working independence estimator relative to the two–stage estimator is

$$\left(\sum_{j=1}^{J} \omega_{jj}^2 / \sum_{j=1}^{J} s_{jj}^{-1}\right)^{4/5}. \tag{1.9}$$

3 Simulations

In this section, we present the results of simulations. The situation we consider is that the predictors X_{ij} for $i = 1, ..., n = 50, 100$ and $j = 1, ..., J = 3$

are independent uniform random variables on the interval $[-2, 2]$. Our theory predicts that our method will improve upon working independence. This will be seen to be the case numerically, as we now describe.

The common variance of the Y_{ij} was $\sigma_\epsilon^2 = 1$. We considered three correlation structures: exchangeable with common correlation $\rho = 0.6$, autoregressive with correlation $\rho = 0.6$, and unstructured where the correlation between measurements 1 and 2 and between measurements 2 and 3 is 0.80, and the correlation between units 1 and 3 is $\rho = 0.5$. The unstructured case was chosen because our theory predicts that it is for this case that the greatest gains in efficiency are possible when accounting for the covariance among observations. Specifically, using (1.9), the asymptotic mean squared error efficiency of the working independence estimator to the two-stage GLS estimator is 0.70, 0.67 and 0.38 in these three cases.

For simplicity, in the kernel calculations we assumed that it was known a priori that the variances of Y_{ij} were independent of j, so that no weighting was performed. The Epanechnikov kernel was used. At each stage of the calculations, we computed the bandwidth locally via EBBS ([7]). Undersmoothing in the two–stage estimator was achieved by multiplying the bandwidth by $(nJ)^{-2/15}$.

We estimated $\mathbf{\Sigma}$ by the following simple device: (a) form the residuals r_{ij} from a working independence fit; and (b) compute the covariance matrix of the residual vectors. We could have undersmoothed the working independence fit, but believe that the essential results would not have changes. If $\widehat{\mathbf{\Sigma}}$ is the estimate so formed, the independence covariance matrix is simply $\widehat{\mathbf{\Sigma}}_{indp} = \text{diag}(\widehat{\mathbf{\Sigma}})$.

Let $z = (x+2)/4$. The functions chosen were Case 1 if $m(x) = \sin(2x)$; Case 2 if $m(x) = \sqrt{z(1-z)}\sin\{2\pi(1+2^{-3/5})/(z+2^{-3/5})\}$; Case 3 if $m(x) = \sqrt{z(1-z)}\sin\{2\pi(1+2^{-7/5})/(z+2^{-7/5})\}$; Case 4 if $m(x) = \sin(8z-4) + 2\exp\{-256(z-.5)^2\}$; Case 5 if $m(x) = H(100x) + H\{-100(x-.5)\}$, where $H(x) = 1/\{1+\exp(-x)\}$. These cases are poorly fit by a quadratic polynomial.

The results are displayed in Tables 1.1–1.2, for $n = 50$ and $n = 100$, respectively. Here we compute the simulation mean squared errors of the estimators. The results for the two–stage GLS estimator with bandwidth estimated by EBBS at $n = 100$ are approximately what is predicted by theory for the exchangeable and autoregressive correlation structures, i.e., the MSE efficiency of working independence is approximately 70%. The same efficiencies occur for the unstructured case: generally around 60% efficiency when the theory predicts 38%. While this is somewhat disappointing, it still is clear evidence that the two–stage GLS method outperforms the working independence estimators.

	Case 1	Case 2	Case 3	Case 4	Case 5
Autoregression					
KN(E), Work	8.81	8.94	9.78	10.34	10.69
KN(E), GLSU	6.90	6.96	7.82	7.97	8.45
Exchangeable	Case 1	Case 2	Case 3	Case 4	Case 5
KN(E), Work	9.09	9.60	10.52	10.76	10.91
KN(E), GLSU	6.66	6.64	7.48	7.70	8.00
Unstructured	Case 1	Case 2	Case 3	Case 4	Case 5
KN(E), Work	9.39	9.89	10.83	11.17	11.55
KN(E), GLSU	5.05	5.51	6.17	6.28	6.80

TABLE 1.1. For $n = 50$, $100\times$ MSE for simulations. KN = kernel, GLSU = our GLS method with an undersmoothed preliminary estimate. Work = Working independence. For kernels, the local bandwidth method EBBS was used to estimate the bandwidth. We considered three correlation structures: exchangeable with common correlation $\rho = 0.6$, autoregressive with correlation $\rho = 0.6$, and unstructured where the correlation between measurements 1 and 2 and between measurements 2 and 3 is 0.80, and the correlation between units 1 and 3 is $\rho = 0.5$.

4 Discussion

Our work was motivated by the fact ([3]) that currently existing methods for nonparametric kernel regression with correlated data do not account for the correlations in a sensible way. Indeed, these methods are often worse than simply ignoring the correlation structure entirely, i.e., than the working independence estimate.

Our main result is the construction of a two–stage kernel estimator that we have shown asymptotically uniformly improves upon the working independence estimator.

Interestingly, our proof shows that the *asymptotic* variance of the two-stage estimator can be calculated by applying standard methods to the derived variables $Z_{ij}\{\widehat{m}_p(\cdot, h_p)\}$ in the second stage of the regression. More precisely, having calculated the undersmoothed first–stage working independence estimator and having calculated the derived variables, both the estimator and its asymptotic variance can be computed as if the derived variables were actual independent observations.

We have described the methods for panel data, so that each individual unit has J observations. The methods have anticipated that the covariance matrix has no particular structure. In other longitudinal data problems, the number of observations per individual unit may depend on i, i.e., $J_i \le J$ say. Our methods are easily extended to this case. What is required in this case is an estimate of the covariance matrix $\mathbf{\Sigma}_i$ of the J_i observations in

	Case 1	Case 2	Case 3	Case 4	Case 5
	Autoregression				
	Case 1	Case 2	Case 3	Case 4	Case 5
KN(E), Work	3.92	3.89	5.29	4.86	5.48
KN(E), GLSU	2.91	2.78	3.99	3.53	4.19
	Exchangeable				
	Case 1	Case 2	Case 3	Case 4	Case 5
KN(E), Work	3.91	4.29	5.07	5.11	5.67
KN(E), GLSU	2.73	3.20	3.80	3.81	4.48
	Unstructured				
	Case 1	Case 2	Case 3	Case 4	Case 5
KN(E), Work	3.93	4.04	4.89	5.11	5.49
KN(E), GLSU	2.10	2.17	2.62	2.79	3.26

TABLE 1.2. For $n = 100$, $100\times$ MSE for simulations. KN = kernel, GLSU = our GLS method with an undersmoothed preliminary estimate. Work = Working independence. For kernels, the local bandwidth method EBBS was used to estimate the bandwidth. We considered three correlation structures: exchangeable with common correlation $\rho = 0.6$, autoregressive with correlation $\rho = 0.6$, and unstructured where the correlation between measurements 1 and 2 and between measurements 2 and 3 is 0.80, and the correlation between units 1 and 3 is $\rho = 0.5$.

the ith unit, and then (1.2) can be employed. Generally in such situations, the covariance matrix is estimated in a structured way, as a function of a vector parameter γ, so that the covariance matrix $\mathbf{\Sigma}_i(\gamma)$ is known up to the parameter γ, e.g., as exchangeable, autoregressive, etc. We believe but have not proved that the extension of our two–stage method to this case will still improve upon the working independence estimator.

Finally, we note an important technical point. In (1.2), the choice of $\mathbf{\Lambda}$ is crucial. While it is true that, for *any* diagonal matrix $\mathbf{\Lambda}$, $\boldsymbol{Z}_i(m) = \mathbf{m}(\mathbf{X}_i) + \mathbf{\Lambda}^{-1}\boldsymbol{\epsilon}_i$, our results are only true for our *particular* choice of $\mathbf{\Lambda}$. The last step in the proof in Section 5.1 only holds for our choice.

5 Appendix

5.1 Sketch of Proof of Result 1

From [3], to terms of order $O_p\{h_p^3 + (nh_p)^{-1}\} = o_p(n^{-1/2})$, we have the asymptotic expansion

$$\widehat{m}_p(x, h_p) - m(x) = (h_p^2/2)m^{(2)}(x) + \left\{\sum_{j=1}^{J} s_{jj}^{-1} f_j(x)\right\}^{-1} \quad (1.10)$$

$$\times n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{J} K_{h_p}(X_{ij} - x) U_{ij} s_{jj}^{-1}.$$

Now, $\widehat{m}(x, h, h_p)$ is the intercept when solving the local linear regression estimating equation

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \sum_{j=1}^{J} \begin{bmatrix} 1 \\ (X_{ij} - x)/h \end{bmatrix} \zeta_j K_h(X_{ij} - x) \\ &\qquad \times \left[Z_{ij} \left\{ \widehat{m}_p(X_{ij}, h_p) \right\} - \alpha_0 - \alpha_1 (X_{ij} - x)/h \right]. \end{aligned}$$

Define

$$C_n = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{J} \zeta_j K_h(X_{ij} - x) \left\{1, (X_{ij} - x)/h\right\}^{\mathrm{T}} \left\{1, (X_{ij} - x)/h\right\}.$$

Then by simple algebra, it can be shown that

$$\begin{aligned} \widehat{m}(x, h, h_p) &= (1, 0) C_n^{-1} (B_{1n} + B_{2n}); \\ B_{1n} &= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{J} \zeta_j K_h(X_{ij} - x) \left\{1, (X_{ij} - x)/h\right\}^{\mathrm{T}} Z_{ij}(m); \\ B_{2n} &= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{J} \begin{bmatrix} 1 \\ (X_{ij} - x)/h \end{bmatrix} \zeta_j K_h(X_{ij} - x) \\ &\qquad \times \left[Z_{ij} \left\{ \widehat{m}_p(X_{ij}, h_p) \right\} - Z_{ij}(m) \right]. \end{aligned}$$

Now, to terms of order $O_p(h^3)$,

$$\begin{aligned} Z_{ij}(m) &= m(X_{ij}) + \omega_{jj}^{-1} \epsilon_{ij} \\ &= m(x) + h m^{(1)}(x) \{(X_{ij} - x)/h\} \\ &\quad + (h^2/2) m^{(2)}(x) \{(X_{ij} - x)/h\}^2 + \omega_{jj}^{-1} \epsilon_{ij} \\ &= \left\{m(x), h m^{(1)}(x)\right\} \left\{1, (X_{ij} - x)/h\right\}^{\mathrm{T}} \\ &\quad + (h^2/2) m^{(2)}(x) \{(X_{ij} - x)/h\}^2 + \omega_{jj}^{-1} \epsilon_{ij}. \end{aligned}$$

By standard calculations, it is easily seen that

$$\begin{aligned} (1, 0) C_n^{-1} B_{1n} &= m(x) + (h^2/2) m^{(2)}(x) \\ &\quad + \left\{ \sum_{j=1}^{J} \zeta_j f_j(x) \right\}^{-1} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{J} (\zeta_j \epsilon_{ij} / \omega_{jj}) K_h(X_{ij} - x) \\ &\qquad + o_p\{h^2 + (nh)^{-1/2}\}. \end{aligned}$$

The mean of this expression is $m(x) + (h^2/2) m^{(2)}(x)$ and its variance is (1.6).

Note that

$$
\begin{aligned}
Z_{ij}\left\{\widehat{m}_p(X_{ij},h_p)\right\} - Z_{ij}(m) &= \widehat{m}_p(X_{ij},h_p) - m(X_{ij}) \\
&\quad -\omega_{jj}^{-1}\sum_{k=1}^{J}\omega_{jk}\left\{\widehat{m}_p(X_{ik},h_p) - m(X_{ik})\right\}.
\end{aligned}
$$

Using (1.10), we can write $(1,0)C_n^{-1}B_{2n} = (1,0)C_n^{-1}(B_{2n1} - B_{2n2} + B_{2n3} - B_{2n4})$, where

$$
\begin{aligned}
(1,0)C_n^{-1}B_{2n1} &= (h_p^2/2)\left\{\sum_{j=1}^{J}\zeta_j f_j(x)\right\}^{-1} \\
&\quad \times n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\zeta_j K_h(X_{ij}-x)m^{(2)}(X_{ij}); \\
(1,0)C_n^{-1}B_{2n2} &= (h_p^2/2)\left\{\sum_{j=1}^{J}\zeta_j f_j(x)\right\}^{-1} \\
&\quad \times n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\zeta_j K_h(X_{ij}-x)\sum_{k}\frac{\omega_{jk}}{\omega_{jj}}m^{(2)}(X_{ik}); \\
(1,0)C_n^{-1}B_{2n3} &= \left\{\sum_{j=1}^{J}\zeta_j f_j(x)\right\}^{-1} n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\zeta_j K_h(X_{ij}-x) \\
&\quad \times\left\{\sum_{j=1}^{J}s_{jj}^{-1}f_j(X_{ij})\right\}^{-1} \\
&\quad \times n^{-1}\sum_{\ell=1}^{n}\sum_{r=1}^{J}K_{h_p}(X_{\ell r}-X_{ij})U_{\ell r}s_{rr}^{-1}; \\
(1,0)C_n^{-1}B_{2n4} &= \left\{\sum_{j=1}^{J}\zeta_j f_j(x)\right\}^{-1} n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J}\zeta_j K_h(X_{ij}-x) \\
&\quad \times\omega_{jj}^{-1}\sum_{k=1}^{J}\omega_{jk}\left\{\sum_{j=1}^{J}s_{jj}^{-1}f_j(X_{ik})\right\}^{-1} \\
&\quad \times n^{-1}\sum_{\ell=1}^{n}\sum_{r=1}^{J}K_{h_p}(X_{\ell r}-X_{ik})U_{\ell r}s_{rr}^{-1}.
\end{aligned}
$$

It is easily seen that to order $o_p\{h^2+(nh)^{-1/2}\}$

$$(1,0)C_n^{-1}(B_{2n1}-B_{2n2}) = -(h_p^2/2)\left\{\sum_{j=1}^{J}\zeta_j f_j(x)\right\}^{-1}$$
$$\times \sum_{j=1}^{M}\sum_{k\neq j}^{M}(\zeta_j\omega_{jk}/\omega_{jj})E\left\{m^{(2)}(X_{ik})|X_{ij}=x\right\}.$$

It is tedious but straightforward to show that

$$(1,0)C_n^{-1}(B_{2n3}-B_{2n4}) = o_p\{h^2+(nh)^{-1/2}\},$$

completing the proof.

5.2 *Proof of* $\mathrm{var}\{\widehat{m}(x;h,h_p)\} \le \mathrm{var}\{\widehat{m}_p(x;h)\}$

Denote by $\mathbf{A}$ any $J\times J$ positive definite symmetric matrix. Let $\mathbf{A}=\{a_{ij}\}$, $\boldsymbol{B}=\mathbf{A}^{1/2}=\{b_{ij}\}$ and $\boldsymbol{C}=\mathbf{A}^{-1/2}=\{c_{ij}\}$. We first show that $\sum_{i=1}^J c_{ii}^2 \ge \sum_{i=1}^J (1/a_{ii})$.

Since for any $1\le i\le J$, $a_{ii}=\sum_{j=1}^J b_{ij}^2$, we have $b_{ii}^2\le a_{ii}$. Since $\boldsymbol{B}$ is a positive definite matrix, the standard matrix theory gives $c_{ii}b_{ii}\ge 1$ ([1], page 403). It follows that $c_{ii}\ge 1/b_{ii}\ge 1/\sqrt{a_{ii}}$, i.e., $c_{ii}\ge 1/\sqrt{a_{ii}}$. Hence $\sum_{i=1}^J c_{ii}^2 \ge \sum_{i=1}^J (1/a_{ii})$.

Define $\mathbf{A}=\boldsymbol{f}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{f}^{-1/2}$, where $\boldsymbol{f}=\mathrm{diag}\{f_1(x),\cdots,f_J(x)\}$. Then the diagonal elements of $\mathbf{A}$ are $a_{ii}=s_{ii}/f_j(x)$. Now $\boldsymbol{C}=\mathbf{A}^{-1/2}=\boldsymbol{f}^{1/4}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{f}^{1/4}$. The diagonal elements of $\boldsymbol{C}$ are $c_{ii}=w_{ii}f_i(x)^{1/2}$. Using the above results, we have $\sum_{i=1}^J w_{ii}^2 f_i(x) \ge \sum_{i=1}^J s_{ii}^{-1} f_i(x)$. It follows immediately from equations (1.4) and (1.8) that $\mathrm{var}\{\widehat{m}(x;h,h_p)\}\le \mathrm{var}\{\widehat{m}_p(x;h)\}$. This completes the proof.

Acknowledgments: This research was supported by grants from the National Science Foundation (SBR-9730282), the Deutsche Forschungsgemeinschaft (Project MA-1026/6-2) National Cancer Institute (CA-57030 and CA-76404) and the National Institute of Environmental Health Sciences (P30-ES09106).

6 References

[1] F. A. Graybill. *Matrices with Applications in Statistic.* Wadsworth & Brooks/Cole, 1983.

[2] D. R. Hoover, J. A. Rice, C. O. Wu, and Y. Yang. Nonparametric smoothing estimates of time–varying coefficient models with longitudinal data. *Biometrika*, 85:809–822, 1998.

[3] X. Lin and R. J. Carroll. Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association*, 95:520–534, 2000.

[4] J. S. Marron and W. Härdle. Random approximations to some measures of accuracy in nonparametric curve estimatio. *it Journal of Multivariate Analysis*, 20:91–113, 1986.

[5] M. S. Pepe and D. Couper. Modeling partly conditional means with longitudinal data. *Journal of the American Statistical Association*, 92:991–998, 1997.

[6] A. Ruckstuhl, A. H. Welsh, and R. J. Carroll. Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statistica Sinica*, 10:51–71, 2000.

[7] D. Ruppert. Empirical-bias bandwidths for local polynomial nonparametric regression and density estimatio. *Journal of the American Statistical Association*, 92:1049–1062, 1997.

[8] T. A. Severini and J. G. Staniswalis. Quasilikelihood estimation in semiparametric models. *Journal of the American Statistical Association*, 89:501–511, 1994.

[9] C. J. Wild and T. W. Yee. Additive extensions to generalized estimating equation methods. *J. Royal Statist. Soc. B*, 58:711–725, 1996.

[10] C. O. Wu, C. T. Chiang, and D. R. Hoover. Asymptotic confidence regions for kernel smoothing of a varying coefficient model with longitudinal data. *Journal of the American Statistical Association*, 93:1388–1402, 1998.

[11] S. L. Zeger and P. J. Diggle. Semi-parametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, 50:689–699, 1994.