Biometrika (2005), **92,** 1, *pp.* 59–74

Printed in Great Britain

Local polynomial regression analysis of clustered data

BY KANI CHEN

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

makchen@ust.hk

AND ZHEZHEN JIN

Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10032, U.S.A.

zjin@biostat.columbia.edu

SUMMARY

This paper proposes a classical weighted least squares type of local polynomial smoothing for the analysis of clustered data, with the key idea of using generalised inverses of correlation matrices. The estimator has a simple closed-form expression. Simplicity is achieved also for nonparametric generalised linear models with arbitrary link function via a transformation. Our approach can be characterised by 'local observations with local variances', which yields intuitively correct results in the sense that correct/incorrect specification of within-cluster correlation has respective positive/negative effects. The approach is a natural extension of classical local polynomial smoothing. Consequently, existing theory can be largely carried over and important issues such as bandwidth selection can be tackled in the classical fashion. Moreover, the approach can handle various types of covariate, such as cluster-level, subject-level or partially cluster-level. Numerical studies support the theoretical results. The method is illustrated with a real example on luteinising hormone levels in cows.

Some key words: Asymptotic bias; Bandwidth selection; Generalised estimating equation; Kernel function; Mean squared error; Nonparametric curve estimation.

1. INTRODUCTION

Nonparametric curve estimation, and local polynomial smoothing in particular, is a well-developed methodology; see for example Fan & Gijbels (1992, 1995, 1996), Fan (1993), Ruppert & Wand (1994), Fan et al. (1995) and Ruppert (1997). Analysis of longitudinal or clustered data, especially through the generalised estimating equations of Liang & Zeger (1986), is of considerable current interest (Lin & Carroll, 2000; Wang, 2003). The aim of this paper is to develop a natural local polynomial smoothing method for the analysis of clustered data, which, in both theory and computation, is simpler and more general than the methods of Lin & Carroll (2000) and Wang (2003).

Many articles have addressed the analysis of clustered data via local polynomial smoothing; see Severini & Staniswalis (1994), Wild & Yee (1996), Zeger & Diggle (1994), Ruckstuhl et al. (2000), Lin & Carroll (2000), Verbyla et al. (1999) and Wang (2003). Lin

& Carroll (2000) was the first paper to present a formal extension of parametric generalised estimating equations, but their method yields a counter-intuitive result that correct specification of within-cluster correlation would have an adverse effect on the curve estimation. We may explain this phenomenon as follows. To estimate the curve at a fixed point x_0, all observations with covariates in the neighbourhood of x_0 are called 'local observations' and all observations in the data are called 'global observations'. In Lin & Carroll (2000), each observation within a cluster is localised through a kernel function, but each cluster is weighted with the correlation matrix of the entire cluster, which we call 'global variance', instead of the variance of the local observations, which we call 'local variance'. This mismatch of 'local observations with global variance' is one plausible explanation of why a correct specification of within-cluster correlation would in fact have an adverse effect on the curve estimation. Wang (2003) solves this puzzle and improves the method of Lin & Carroll with 'seemingly unrelated observations'. In her method, all 'global observations' are used in the estimation. The nonlocal observations, after subtraction of their means, can provide information about local observations if their correlation with local observations is correctly modelled; essentially Wang (2003) fixes the mismatch of Lin & Carroll by using the match 'global observations with global variance'. Wang's method does indeed yield smaller asymptotic variance when the within-cluster correlation is correctly specified, but it also has drawbacks. Heuristically, the variance reduction is compromised by a possible increase in bias: the nonlocal observations may bring in a sizeable bias because their means, for subtraction, are unknown. If the within-cluster correlation is modelled incorrectly, the accuracy of curve estimation would deteriorate much more severely than with the method we propose. Moreover, Wang's method has computational and theoretical difficulties. First, it requires an iteration procedure with initial estimate satisfying an asymptotic expansion, which may diverge, and the resulting asymptotic bias term contains the asymptotic bias of the initial estimate; see equations (8) and (11) of Wang (2003). This implies that, even though the initial estimate may be a reasonable choice, its bias will be carried over to the final estimate of Wang (2003). In addition, the asymptotic bias after full iteration does not have a closed-form expression and is difficult to evaluate; see equation (14) in Wang (2003). Furthermore, the theory of both Lin & Carroll (2000) and Wang (2003) deals with local linear estimation rather than general local polynomial estimation, let alone estimation of the higher-order derivatives of the curve.

The method of this paper, in contrast to those of Lin & Carroll (2000) and Wang (2003), can be characterised by the description 'local observations with local variances'. The idea is simple and intuitive. We shall only use the relevant observations, namely the local observations, and weight them by their own variances, namely the local variances, rather than by global variances. The estimator is simple and has an explicit expression similar to that of the weighted least squares estimators; see equation (3). As a result, the computational workload is minimal. Also, the within-cluster correlation is indeed used appropriately to improve the estimation: when the within-cluster correlation is involved in the curve estimation, correct specification has a positive effect; when it is not involved, correct specification does not have an adverse effect. To be specific, if the number of subjects with one covariate in a small neighbourhood of a point is comparable with the number of subjects with two or more covariates in the same small neighbourhood, then correctly specifying the within-cluster correlation indeed results in more accurate estimation of the function at that point. Otherwise, for example, if we assume that

covariates have positive and continuous joint density, correctly or incorrectly specifying the within-cluster correlation leads to the same accuracy of estimation asymptotically. As will be seen, this phenomenon has a theoretical foundation, can be verified empirically and, more importantly, can be justified intuitively; see Remarks 4 and 5. Furthermore, whether the covariate in study is of cluster-level, subject-level or partially cluster-level makes no difference to our estimation method, and the properties of the curve estimator are the same for all cases; see Proposition 2 and Remark 5. In addition, the rich theory of local polynomial smoothing for non-clustered data can be carried over straightforwardly; see for example the calculation of the asymptotic bias and asymptotic variance of the estimators of the function and its higher-order derivatives in Proposition 1. The above advantages hold for all nonparametric generalised linear models with arbitrary link function via a transformation. The key idea is the use of generalised inverses of matrices.

2. The curve estimation procedure

Suppose (x_{ij}, y_{ij}), for $j = 1, \ldots, J$, are the J covariate-response pairs of subject i, which are independent and identically distributed, for $i = 1, \ldots, n$. The covariates x_{ij} are scalar. In the spirit of generalised linear models, we assume that, for $j = 1, \ldots, J$, $E(y_{ij}|x_{ij} = x) = g\{\theta(x)\}$ and $\mathrm{var}(y_{ij}|x_{ij} = x) = \phi_j v[g\{\theta(x)\}]$, where $g(.)$ and $v(.)$ are smooth functions, ϕ_j is associated with dispersion, and $\theta(.)$ is the unknown function to be estimated. The above conditional variance may in general contain a factor w_{ij}, which, without loss of generality, is omitted here for simplicity of presentation.

If $\theta(.)$ is assumed to belong to a parametric family, the parameters can be estimated by the method of parametric generalised estimating equations. The key idea is to use $R_i = R(\alpha, x_{i1}, \ldots, x_{iJ})$ to model the within-cluster correlation matrix R_{i0}, the conditional correlation matrix of $\{y_{ij}, j = 1, \ldots, J\}$, given $\{x_{ij}, j = 1, \ldots, J\}$. For example, one can model the correlations with the autoregressive model or the exchangeable model that assumes all the correlations to be the same. In general, the modelling of R_{i0} by R_i may involve an unknown parameter α to be estimated separately from the data. For simplicity of argument and without loss of generality, we assume throughout the paper that α is fixed, that the eigenvalues of R_i and R_{i0} are uniformly bounded below away from 0, and that the elements of R_i and R_{i0} are continuous functions of the covariates.

In nonparametric regression models, $\theta(.)$ is arbitrary except for certain differentiability properties. Precisely for this reason, the above regression model can be equivalently formulated as

$$y_{ij} = m(x_{ij}) + \sigma(x_{ij})\varepsilon_{ij} \quad (j = 1, \ldots, J, i = 1, \ldots, n), \tag{1}$$

where $m(.) = g\{\theta(.)\}$, $\sigma(.) = [v\{m(.)\}]^{\frac{1}{2}}$ and the error ε_{ij} satisfies

$$E(\varepsilon_{ij}|x_{i1}, \ldots, x_{iJ}) = 0, \quad \mathrm{var}(\varepsilon_{ij}|x_{i1}, \ldots, x_{iJ}) = \mathrm{var}(y_{ij}|x_{i1}, \ldots, x_{iJ})/\sigma^2(x_{ij}) = \phi_j,$$

for $j = 1, \ldots, J$. Let $\Phi = \mathrm{diag}(\phi_1, \ldots, \phi_J)$, $\Sigma_i = \mathrm{diag}\{\sigma^2(x_{i1}), \ldots, \sigma^2(x_{iJ})\}$ and

$$V_{i0} = \mathrm{var}\{(y_{i1}, \ldots, y_{iJ})^{\mathrm{T}}|x_{i1}, \ldots, x_{iJ}\}.$$

Then $R_{i0} = (\Phi\Sigma_i)^{-\frac{1}{2}} V_{i0} (\Sigma_i \Phi)^{-\frac{1}{2}}$. Both V_{i0} and R_{i0} may possibly depend on the covariates of subject i. Unlike in the parametric setting, in the nonparametric setting it might be estimation of $g\{\theta(.)\}$ rather than $\theta(.)$ that is of interest because the former represents the conditional mean of the response given the covariates and the latter is a totally unspecified function which may lack clear interpretation.

The Moore–Penrose generalised inverse of a matrix will play a key role. We define the generalised inverse of any symmetric $J \times J$ matrix A to be a symmetric matrix, denoted still by A^{-1}, such that $AA^{-1}A = A$ and $A^{-1}AA^{-1} = A^{-1}$. To be specific, suppose that $A = \Gamma \operatorname{diag}(\lambda_1, \ldots, \lambda_J)\Gamma^{\mathrm{T}}$, where Γ is an orthonormal matrix. Then, $A^{-1} = \Gamma \operatorname{diag}(1/\lambda_1, \ldots, 1/\lambda_J)\Gamma^{\mathrm{T}}$, where $1/0$ becomes 0.

Throughout the paper, x_0 is an arbitrary but fixed interior point of the domain of x_{ij} and $K(.)$ is a symmetric density function with bounded support assumed, without loss of generality, to be $[-1, 1]$. Define $K_h(t) = K(t/h)/h$, where h is the bandwidth. Typical choices of $K(.)$ are, for example, the Epanechnikov kernel $K(t) = 0{\cdot}75(1 - t^2)I(|t| \leqslant 1)$ and the uniform kernel $K(t) = 0{\cdot}5I(|t| \leqslant 1)$. Here and throughout the paper, $I(.)$ is the indicator function. Let $K_{ih} = \operatorname{diag}\{K_h(x_{i1} - x_0), \ldots, K_h(x_{iJ} - x_0)\}$ and

$$W_i = (K_{ih}^{-\frac{1}{2}}\Phi^{\frac{1}{2}}R_i\Phi^{\frac{1}{2}}K_{ih}^{-\frac{1}{2}})^{-1} = K_{ih}^{\frac{1}{2}}\Phi^{-\frac{1}{2}}(I_iR_iI_i)^{-1}\Phi^{-\frac{1}{2}}K_{ih}^{\frac{1}{2}}, \tag{2}$$

where

$$\begin{aligned} I_i &= \operatorname{diag}[I\{K_h(x_{i1} - x_0) > 0\}, \ldots, I\{K_h(x_{iJ} - x_0) > 0\}] \\ &= \operatorname{diag}\{I(|x_{i1} - x_0| \leqslant h), \ldots, I(|x_{iJ} - x_0| \leqslant h)\}. \end{aligned}$$

Define

$$X_i = \begin{pmatrix} 1 & (x_{i1} - x_0) & \ldots & (x_{i1} - x_0)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (x_{iJ} - x_0) & \ldots & (x_{iJ} - x_0)^p \end{pmatrix}_{J \times (p+1)}, \quad Y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iJ} \end{pmatrix}_{J \times 1}.$$

Minimising $\sum_{i=1}^n \{(Y_i - X_i\beta)^{\mathrm{T}}W_i(Y_i - X_i\beta)\}$ over $\beta = (\beta_0, \ldots, \beta_p)^{\mathrm{T}} \in R^{p+1}$, one obtains

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \left(\sum_{i=1}^n X_i^{\mathrm{T}}W_iX_i \right)^{-1} \left(\sum_{i=1}^n X_i^{\mathrm{T}}W_iY_i \right). \tag{3}$$

Set $\hat{m}_k(x_0) = k!\hat{\beta}_k$ to be the estimator of $m^{(k)}(x_0)$, for $k = 0, \ldots, p$. In particular, $\hat{m}_0(x_0)$ is the estimator of $m(x_0)$. Naturally, one defines $\tilde{\theta}(x_0) = g^{-1}\{\hat{m}(x_0)\} = g^{-1}(\hat{\beta}_0)$ to be the estimator of $\theta(x_0)$. Although it appears to be more natural to use R_i^{-1} rather than $(I_iR_iI_i)^{-1}$ in expression (2), it will become clear later that such a replacement will result in less accurate estimators; see Corollary 2. Proofs are given in the Appendix.

The greatest advantage of the estimator proposed above is its simplicity. Its expression is in closed form and analogous to the classical weighted least squares type of local polynomial estimators for non-clustered data. The related theory and computation are straightforward. The estimators of Lin & Carroll (2000), for instance, are solutions of estimating equations which do not have closed-form expressions, so that numerical solution is necessary. Furthermore, such computation has to be repeated at any point x_0 of interest if one wishes to estimate $\theta(x_0)$ or $m(x_0)$. The computation is more complex in Wang (2003) since it relies on the estimate of Lin & Carroll (2000) as an initial estimate and, moreover, requires another iteration procedure.

The above estimate $\hat{m}_k(x)$ does not use either $g(.)$ or $v(.)$. The function $g(.)$ is only used in deriving the estimate of $\theta(.)$. Therefore, the framework of generalised linear models becomes less essential. This point is most clearly seen from (1), where $\sigma(.)$ plays the role of a nuisance parameter and should not be modelled unless necessary. On the other hand,

if $v(.)$ is known, one may choose to modify the definition of W_i in (2) slightly to incorporate this information and obtain slightly different estimates. However, such a modification incurs more computation and it is unclear if the modified version is better.

3. Asymptotic properties

Let $\mathscr{F}_n^X$ denote the σ-algebra generated by $(x_{i1}, \ldots, x_{iJ})$, for $i = 1, \ldots, n$. Using only the positivity and continuity of $\sigma^2(.)$ at x_0, one can show via a direct calculation that

$$\mathrm{var}\,(\hat{\beta}|\mathscr{F}_n^X) = \sigma^2(x_0) A_n^{-1} B_n A_n^{-1}\{1 + o_P(1)\}, \tag{4}$$

where $A_n = \sum_{i=1}^n X_i^{\mathrm{T}} W_i X_i$ and

$$B_n = \sum_{i=1}^n X_i^{\mathrm{T}} \Phi^{-\frac{1}{2}} K_{ih}^{\frac{1}{2}} (I_i R_i I_i)^{-1} K_{ih}^{\frac{1}{2}} R_{i0} K_{ih}^{\frac{1}{2}} (I_i R_i I_i)^{-1} K_{ih}^{\frac{1}{2}} \Phi^{-\frac{1}{2}} X_i. \tag{5}$$

The asymptotic theory requires some regularity conditions on the local distribution of the covariates. Let Ω_k, for $1 \leqslant k \leqslant 2^J - 1$, be the $2^J - 1$ distinct subsets of $\{1, \ldots, J\}$, except for the empty set, and let $B(x, h)$ denote the interval $[x - h, x + h]$.

Then an important condition is that there exists a $\delta_0 > 0$ such that, for all $x \in B(x_0, \delta_0)$ and all $k = 1, \ldots, 2^J - 1$,

$$\begin{aligned}
&\mathrm{pr}\,\{x_{1j} \in B(x, h) \text{ and are equal for all } j \in \Omega_k, \text{ and } x_{1j_1} \neq x_{1j} \text{ for any } j_1 \notin \Omega_k \text{ and } j \in \Omega_k\} \\
&\quad = \int_{-h}^{h} f_k(x + t)dt \\
&\quad = \mathrm{pr}\,\{x_{1j} \in B(x, h) \text{ for all } j \in \Omega_k, \text{ and } x_{1j} \notin B(x, h) \text{ for all } j \notin \Omega_k\} + o(h),
\end{aligned}$$

for all $0 < h < 2\delta_0$, where $f_k(.)$, for $1 \leqslant k \leqslant 2^J - 1$, 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 1. This condition is referred to as 'the existence of partial density' of the covariates $(x_{11}, \ldots, x_{1J})^{\mathrm{T}}$ at x_0, because, for every $k = 1, \ldots, 2^J - 1$, $f_k(.)$ can be viewed as the partial density of the partial cluster-level covariates $\{x_{1j}, j \in \Omega_k\}$. The condition ensures that, unless they are of partial cluster-level, two covariates take values in a small neighbourhood of x_0 with negligible chance. This condition precisely features the types of covariate of interest: cluster-level covariates, partial cluster-level covariates and covariates with existing joint density. Cluster-level covariates, i.e. such that $x_{i1} = \ldots = x_{iJ}$, typically appear in, for example, the repeated measurements of responses given the same covariates. Partial cluster-level covariates are also rather common, for example when covariates are a random point process observed at different times. Since the paths of a point process are step functions, consecutive observations may well be identical, that is $\mathrm{pr}\,\{x_{i,j} = x_{i,j+1}\} > 0$. Note that the marginal density of x_{1l}, denoted by $f_l^*(.)$, satisfies $f_l^*(.) = \sum_{k=1}^{2^J-1} f_k(.) I(l \in \Omega_k)$. Consider the special case when the joint density of x_{1j} $(1 \leqslant j \leqslant J)$ exists. Suppose, for $k = 1, \ldots, J$, that $\Omega_k = \{k\}$. Then, in this case, $f_k(.)$ is the marginal density of x_{1k}, for $k = 1, \ldots, J$, and the remaining $f_k(.)$ are 0. Consider another special case of cluster-level covariates such that $\mathrm{pr}\,\{x_{11} = \ldots = x_{1J}\} = 1$. Suppose that $\Omega_{2^J-1} = \{1, \ldots, J\}$. Then, in this case, $f_{2^J-1}(.)$ is the common marginal density for every covariate, and all other $f_k(.)$, for $k \neq 2^J - 1$, are 0.

Let $\mathscr{S}_k(0) = \{x_{1j} = x_0 \text{ for all } j \in \Omega_k, \text{ and } x_{1j} \neq x_0 \text{ for all } j \notin \Omega_k\}$, and define

$$\xi_k = E\{1_0^{\mathrm{T}} \Phi^{-\frac{1}{2}} (I_{k0} R_1 I_{k0})^{-1} \Phi^{-\frac{1}{2}} 1_0 | \mathscr{S}_k(0)\},$$

where $I_{k0} = \text{diag}\{I(1 \in \Omega_k), \ldots, I(J \in \Omega_k)\}$, which is a $J \times J$ nonrandom matrix, and 1_0 is a J-vector with all components equal to 1. Define ξ_{k0} as for ξ_k except with the modelled correlation matrix R_1 replaced by the true correlation matrix R_{10}. Define $\bar{\xi}_k$ as for ξ_k except with $(I_{k0}R_1I_{k0})^{-1}$ replaced by $(I_{k0}R_1I_{k0})^{-1}R_{10}(I_{k0}R_1I_{k0})^{-1}$. Note that $\xi_k = \bar{\xi}_k = \xi_{k0}$ if the modelled correlation matrix equals the true correlation matrix, that is $R_1 = R_{10}$.

Some more notation is needed. Set $\mu_j = \int t^j K(t)dt$, $\nu_j = \int t^j K^2(t)dt$, $c_p = (\mu_{p+1}, \ldots, \mu_{2p+1})^{\mathrm{T}}$, $S = (\mu_{i+j})_{0 \leqslant i,j \leqslant p}$ and $\bar{S} = (\nu_{i+j})_{0 \leqslant i,j \leqslant p}$. Note that c_p is a $(p+1)$-vector and that S and $\bar{S}$ are $(p+1) \times (p+1)$ matrices. Let $e_k = (0, \ldots, 0, 1, 0, \ldots, 0)^{\mathrm{T}}$ be a $(p+1)$-vector, where the unique 1 occurs at the kth position, and let e_0 be the $(p+1)$-vector of zeros. Let f_j^* denote the density function of x_{ij}.

PROPOSITION 1. *Suppose that the condition of the existence of partial density holds and $\sigma^2(.)$ and $m^{(p+1)}(.)$ are continuous at x_0 with $\sigma^2(x_0) > 0$. Let $k = 0, \ldots, p$, and assume that $h \to 0$ and $nh \to \infty$. Then the following results hold.*

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

$$\text{var}\{\hat{m}_k(x_0)|\mathscr{F}_n^X\} = \frac{k!^2}{nh^{1+2k}} e_{k+1}^{\mathrm{T}} S^{-1} \bar{S} S^{-1} e_{k+1} \frac{\sigma^2(x_0) \sum_{l=1}^{2^J-1} f_l(x_0) \bar{\xi}_l}{\{\sum_{l=1}^{2^J-1} f_l(x_0) \xi_l\}^2} + o_P\left(\frac{1}{nh^{1+2k}}\right). \quad (6)$$

The conditional bias of $\hat{m}_k(x_0)$ for $p-k$ odd is

$$\text{Bias}\{\hat{m}_k(x_0)|\mathscr{F}_n^X\} = e_{k+1}^{\mathrm{T}} S^{-1} c_p \frac{k!}{(p+1)!} m^{(p+1)}(x_0) h^{p+1-k} + o_P(h^{p+1-k}), \quad (7)$$

where the main term on the right-hand side is free of the distribution of the covariates or the modelled or true correlation matrices. For $p-k$ even, (7) still holds and the first term on the right-hand side is 0.

(ii) *For any given bandwidth and kernel, the asymptotic variance and mean squared error are minimised when the modelled correlation matrix equals the true correlation matrix, that is $R_i = R_{i0}$, in which case $\xi_l = \bar{\xi}_l = \xi_{l0}$, for $1 \leqslant l \leqslant 2^J - 1$.*

(iii) *For any given bandwidth, the asymptotic variance is minimised when the modelled correlation matrix equals the true correlation matrix and the kernel is the uniform kernel.*

(iv) *For a given kernel $K(.)$ and $p-k$ odd, the asymptotic mean squared error of $\hat{m}_k(x_0)$ is minimised when the modelled correlation matrix equals the true correlation matrix and the bandwidth is the optimal local variable bandwidth,*

$$\left[\frac{(2k+1)(p+1)!^2 \int \{K_{k,p}(t)\}^2 dt \sigma^2(x_0)}{n(2p+2-2k)\{m^{(p+1)}(x_0) \int t^{p+1} K_{k,p}(t)dt\}^2 \sum_{l=1}^{2^J-1} f_l(x_0) \xi_{l0}}\right]^{1/(2p+3)}, \quad (8)$$

where $K_{k,p}$ is the equivalent kernel of order (k, p) induced by K. Moreover, for $p-k$ odd, the asymptotic mean squared error is minimised when the modelled correlation matrix is the true correlation matrix, the smooth symmetric nonnegative kernel is the Epanechnikov kernel and the bandwidth is the above optimal bandwidth.

Remark 2. Result (iv) of Proposition 1 provides the optimal theoretical local variable bandwidth based on the mean squared error criterion, which varies with the value of x_0. In practice, however, a global variable bandwidth which is constant might be preferred because of its simplicity. An expression for asymptotic optimal global bandwidth can be obtained by minimising the conditional weighted mean integrated squared error,

$$\text{MISE} = \int ([\text{Bias}\{\hat{m}_k(x)|\mathscr{F}_n^X\}]^2 + \text{var}\{\hat{m}_k(x)|\mathscr{F}_n^X\}) w(x)dx = \int \text{MSE}\{\hat{m}_k(x)|\mathscr{F}_n^X\} w(x)dx,$$

say, where $w(.)$ is a nonnegative weight function. This leads to the global constant optimal bandwidth expression,

$$\left[\frac{(2k+1)(p+1)!^2\int\{K_{k,p}(t)\}^2\,dt\{\int\sigma^2(x)w(x)/\sum_{l=1}^{2^J-1}f_l(x)\xi_{l0}\,dx\}}{n(2p+2-2k)\{\int t^{p+1}K_{k,p}(t)dt\}^2\int\{m^{(p+1)}(x)\}^2w(x)dx}\right]^{1/(2p+3)}.$$

Remark 3. The asymptotic properties of $\hat\theta(x_0)$ can be obtained straightforwardly from Proposition 1. It follows from Taylor expansion that

$$\hat\theta(x_0)-\theta(x_0)=\frac{1}{g'\{\theta(x_0)\}}\{\hat m_0(x_0)-m(x_0)\}-\frac{g''\{\theta(x_0)\}}{2[g'(\theta(x_0)\}]^3}\{\hat m_0(x_0)-m(x_0)\}^2\{1+o_P(1)\}.$$

By a careful examination of the negligible terms, one can show that

$$\begin{aligned}\text{Bias}\,\{\hat\theta(x_0)|\mathscr{F}_n^X\}&=\left[\frac{1}{g'\{\theta(x_0)\}}\text{Bias}\,\{\hat m_0(x_0)|\mathscr{F}_n^X\}-\frac{g''\{\theta(x_0)\}}{2[g'\{\theta(x_0)\}]^3}\,\text{var}\,\{\hat m_0(x_0)|\mathscr{F}_n^X\}\right]\\&\quad\times\{1+o_P(1)\},\\\text{var}\,\{\hat\theta(x_0)|\mathscr{F}_n^X\}&=[g'\{\theta(x_0)\}]^{-2}\,\text{var}\,\{\hat m_0(x_0)|\mathscr{F}_n^X\}\{1+o_P(1)\},\\\text{MSE}\{\hat\theta(x_0)|\mathscr{F}_n^X\}&=[g'\{\theta(x_0)\}]^{-2}\text{MSE}\{\hat m_0(x_0)|\mathscr{F}_n^X\}\{1+o_P(1)\}.\end{aligned}$$

In the special cases when the joint density of the covariates exists or when the covariates are of cluster-level, the expression for the asymptotic variance can be simplified as follows.

COROLLARY 1. *Suppose that the conditions of Proposition* 1 *hold.*

(i) *If, in particular, the joint density of* $(x_{11},\ldots,x_{1J})^{\mathrm T}$ *exists, then the conditional variance of* $\hat m_k(x_0)$ *can be simplified as*

$$\text{var}\,\{\hat m_k(x_0)|\mathscr{F}_n^X\}=\frac{k!^2}{nh^{1+2k}}e_{k+1}^{\mathrm T}S^{-1}\bar SS^{-1}e_{k+1}\frac{\sigma^2(x_0)}{\sum_{l=1}^J f_l^*(x_0)/\phi_l}+o_P\left(\frac{1}{nh^{1+2k}}\right),\qquad(9)$$

where $f_l^*(.)$ *is the marginal density of* x_{1l}. *In this particular case, the modelled correlation matrix, correct or incorrect, does not affect the asymptotic variance of the curve estimation.*

(ii) *If, in particular, the covariates are of cluster level, that is* $\text{pr}\,\{x_{11}=\ldots=x_{1J}\}=1$, *then the conditional variance of* $\hat m_k(x_0)$ *is*

$$\text{var}\,\{\hat m_k(x_0)|\mathscr{F}_n^X\}=\frac{k!^2}{nh^{1+2k}}e_{k+1}^{\mathrm T}S^{-1}\bar SS^{-1}e_{k+1}\frac{\sigma^2(x_0)\bar\xi_*}{f_1^*(x_0)\xi_*^2}+o_P\left(\frac{1}{nh^{1+2k}}\right),\qquad(10)$$

where $f_1^*(.)=\ldots=f_J^*(.)$ *is the common density of* $x_{11},\ldots,x_{1J}$ *and* ξ_* *and* $\bar\xi_*$ *are respectively defined as for* ξ_k *and* $\bar\xi_k$ *except with* $\mathscr{S}_k(0)$ *replaced by* $\{x_{11}=\ldots=x_{1J}=x_0\}$.

Remark 4. When the joint density of covariates exists, the modelled correlation matrix does not appear in the expressions for asymptotic variance and bias and therefore does not affect the accuracy of curve estimation. This phenomenon is in fact not surprising at all. A heuristic but intuitive explanation is as follows. In constructing the estimator $\hat m_k(x_0)$, only those (x_{ij},y_{ij}) with $|x_{ij}-x_0|\leqslant h$ are used. The existence of joint density ensures that the number of subjects with exactly one covariate-response pair used is $2nh\sum_{j=1}^J f_j^*(x_0)\{1+o_P(1)\}$, while the number of subjects with two or more covariate-response pairs used is $O_P(nh^2)$. Apparently, the estimate $\hat m_k(x_0)$ is mostly determined by

subjects with exactly one covariate-response pair used, so that modelling within-cluster correlation becomes redundant and the local polynomial smoothers for clustered data are essentially reduced to those for non-clustered data.

Suppose that, in the definition of β, we replace $(I_i R_i I_i)^{-1}$ by R_i^{-1}. The resulting estimator $\hat{m}_k^*(x_0)$ of $m_k(x_0)$ in the case of $p=1$ is the estimator proposed in Lin & Carroll (2000). Corollary 2 shows that $\hat{m}_k^*(x_0)$ is actually less accurate than $\hat{m}_k(x_0)$.

COROLLARY 2. *Suppose that the conditions of Proposition* 1 *hold. If the joint density of* $(x_{11}, \ldots, x_{1J})^{\mathrm{T}}$ *exists, then the conditional biases of* $\hat{m}_k^*(x_0)$ *are the same as* (7) *and the conditional variance of* $\hat{m}_k^*(x_0)$ *is*

$$\operatorname{var}\{\hat{m}_k^*(x_0)|\mathscr{F}_n^X\} = \frac{k!^2}{nh^{1+2k}} e_{k+1}^{\mathrm{T}} S^{-1} \bar{S} S^{-1} e_{k+1} \frac{\sigma^2(x_0)\{\sum_{l=1}^J f_l^*(x_0) r_l^2/\phi_l\}}{\{\sum_{l=1}^J f_l^*(x_0) r_l/\phi_l\}^2} + o_P\left(\frac{1}{nh^{1+2k}}\right), \tag{11}$$

where $f_l^*(.)$ *is the marginal density of* x_{1l} *and* $r_l = E(r^{il}|x_{il}=x_0)$, *with* r^{il} *being the* l*th diagonal element of* R_i^{-1}. *Moreover,* $\hat{m}_k^*(x_0)$ *has larger asymptotic mean squared error than* $\hat{m}_k(x_0)$ *unless* R_i *is the identity matrix.*

The ease of theoretical analysis is reflected in Proposition 1 and Corollary 1 and the close resemblance of these results to those for non-clustered data. In fact, the asymptotic biases given in (7) are identical to the classical counterpart, and, when $J=1$, the asymptotic variances in (6), (9) and (10) also reduce to their classical counterpart; see for example Fan & Gijbels (1996, p. 62). This can also be seen in the following corollary, which presents asymptotic properties of the widely-used local linear smoothers. As a result of the theoretical simplicity, critical issues such as optimal bandwidth, optimal choice of kernel and minimisation of mean squared error are easily solved.

COROLLARY 3 (*Local linear smoothers*). *Suppose that the condition of the existence of partial density holds and that* $\sigma^2(.)$ *and* $m''(.)$ *are continuous at* x_0 *with* $\sigma^2(x_0)>0$. *Assume that* $h \to 0$ *and* $nh \to \infty$. *Then the following results hold.*

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

$$\operatorname{var}\{\hat{m}(x_0)|\mathscr{F}_n^X\} = \frac{\int K^2(t)dt}{nh} \frac{\sigma^2(x_0)\sum_{k=1}^{2^J-1} f_k(x_0)\bar{\xi}_k}{\{\sum_{k=1}^{2^J-1} f_k(x_0)\xi_k\}^2}\{1+o_P(1)\}, \tag{12}$$

and the conditional bias of $\hat{m}(x_0)$ *is*

$$\operatorname{Bias}\{\hat{m}(x_0)|\mathscr{F}_n^X\} = \frac{1}{2}h^2 m''(x_0)\int t^2 K(t)dt + o_P(h^2). \tag{13}$$

(ii) *Assume that* $m''(x_0) \neq 0$. *The conditional asymptotic mean squared error is minimised when the modelled correlation matrix is the true correlation matrix, the smooth symmetric nonnegative kernel is the Epanechnikov kernel, the bandwidth is*

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

and the minimised asymptotic mean squared error is

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

For a given bandwidth, the uniform kernel with the true correlation matrix achieves the minimum asymptotic variance.

As stated before, the condition of the existence of partial density is broad enough to cover most interesting cases. In fact, for the purpose of variance minimisation, even this mild condition can be dropped. This is reflected in Proposition 2.

PROPOSITION 2. *Suppose that $\sigma^2(.)$ is continuous and positive at x_0, and assume that $h\to 0$ and $nh\to\infty$. Then the asymptotic variance of $\hat{m}_k(x_0)$ is minimised when $R_i=R_{i0}$ and $K(.)$ is the uniform kernel, in which case $A_n=2B_n$. The minimised asymptotic variance of $\hat{\beta}$ is*

$$\sigma^2(x_0)\left\{\sum_{i=1}^{n}X_i^{\mathrm{T}}\Phi^{-\frac{1}{2}}(I_iR_{i0}I_i)^{-1}\Phi^{-\frac{1}{2}}X_i\right\}^{-1}\{1+o_P(1)\}.$$

In particular, if the joint density of $(x_{11},\ldots,x_{1J})^{\mathrm{T}}$ exists, then the asymptotic variance of $\hat{m}_k(x_0)$ is minimised when $K(.)$ is the uniform kernel and R_i is an arbitrary correlation matrix.

Remark 5. Curve estimation is more accurate when the modelled correlation matrix is equal to the true correlation matrix. For illustration, suppose that $g(.)$ is the identity function and $\phi_j=1$, but without assuming within-cluster independence. The conditional variance of the proposed estimator with $R_i=R_{i0}$ and the uniform kernel is

$$\left(\sum_{i=1}^{n}X_i^{\mathrm{T}}W_iX_i\right)^{-1}\left(\sum_{i=1}^{n}X_i^{\mathrm{T}}W_iV_iW_iX_i\right)\left(\sum_{i=1}^{n}X_i^{\mathrm{T}}W_iX_i\right)^{-1}$$
$$\simeq\sigma^2(x_0)\left\{\sum_{i=1}^{n}X_i^{\mathrm{T}}(I_iR_{i0}I_i)^{-1}X_i\right\}^{-1},$$

while the conditional variance of the existing working independence estimator is approximately

$$\sigma^2(x_0)\left(\sum_{i=1}^{n}X_i^{\mathrm{T}}I_iX_i\right)^{-1}\left(\sum_{i=1}^{n}X_i^{\mathrm{T}}I_iR_{i0}I_iX_i\right)\left(\sum_{i=1}^{n}X_i^{\mathrm{T}}I_iX_i\right)^{-1}.$$

It is clear that the former is smaller than or equal to the latter, with equality only when either all $\{R_{i0}\}$ are identity matrices, which implies that there is no within-cluster correlation, or each I_i contains at most one nonzero diagonal element, which implies that every observation has at most one covariate in the interval $[x_0-h, x_0+h]$. Heuristically, if the within-cluster correlation is nonzero and there exist observations with more than one covariate in the interval, then the proposed estimator with correct modelling is better. Otherwise, the proposed estimator is still as good as the existing one. In other words, if the within-cluster correlation matrix is involved in the estimation, the estimator based on a correct specification is more accurate than the existing estimator. With correct specification of within-cluster correlation, the limiting conditional variance of the proposed estimator is always smaller than or equal to that of the existing estimator.

4. NUMERICAL EXAMPLES

Simulation studies are carried out to examine the performance of the proposed method, with data generated from the model

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

where $m(x) = 2 \times \exp\{\sin(10x)\}$ and the errors $(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})$ follow the trivariate normal distribution with mean 0, $\mathrm{var}(\varepsilon_{i1}) = 0{\cdot}25$, $\mathrm{var}(\varepsilon_{i2}) = 0{\cdot}64$, $\mathrm{var}(\varepsilon_{i3}) = 0{\cdot}49$ and $\mathrm{corr}(\varepsilon_{ij}, \varepsilon_{ik}) = 0{\cdot}6$, for $j \neq k$ and $j, k = 1, 2, 3$. The covariates are of partial cluster-level: x_{i1} and x_{i2} are generated independently from the uniform distribution $\mathrm{Un}[-2, 2]$ and $x_{i3} = x_{i1}$. The errors and covariates are independent.

The number of simulations is 1000 and the sample size n is either 50 or 100. The curve estimate $\hat{m}_0(.)$ is computed on the grid points $x_j = -1{\cdot}8 + 0{\cdot}036j$ $(j = 0, \ldots, 100)$, with several choices of bandwidth h, using four different estimation methods: the proposed local linear, $p = 1$, method; the working independence method of Lin & Carroll (2000); the one-step estimation method of Wang (2003); and the estimation method of Wang (2003) with iterations. The Epanechnikov kernel was used in all methods.

For each of the grid points, the bias and variance were computed based on the 1000 simulation runs. Also, the integrated squared error D_i was obtained for the ith simulation, where $D_i = \int_{-1{\cdot}8}^{1{\cdot}8} \{m(x) - \hat{m}_{0i}(x)\}^2 dx$ $(i = 1, \ldots, 1000)$ with the integration replaced by summation over $x_j = -1{\cdot}8 + 0{\cdot}036j$ $(j = 0, \ldots, 100)$. Table 1 summarises the results: 'Bias', the average of the absolute values of biases over the 101 grid points; 'SD', the average of the sample standard deviations over the 101 grid points; and 'MISE', the average of integrated squared errors. Table 1 also reports the relative values of MISE for the three other estimators to that for the proposed estimator: a ratio greater than 1 indicates that the new estimator performs better.

The estimates based on the proposed method have the smallest overall bias among the four methods, for each fixed n and h. Wang's method outperforms that of Lin & Carroll in terms of SD and MISE only when $h = 0{\cdot}2$ for $n = 50$ and when $h = 0{\cdot}1$ for $n = 100$. This indicates that whether or not Wang's method is better than Lin & Carroll's depends

Table 1. *Comparison of methods based on* 1000 *simulations*

	Proposed estimator			Lin–Carroll's estimator			Wang's first-step estimator			Wang's estimator after iterations		
h	Bias	SD	MISE_1	Bias	SD	RMISE	Bias	SD	RMISE	Bias	SD	RMISE
						$n = 50$						
0·1	0·139	1·134	5·329	0·159	2·979	11·76	0·176	2·677	13·10	0·183	2·044	5·81
0·2	0·529	0·411	2·036	0·532	0·471	1·52	0·593	0·438	1·25	0·628	0·448	1·33
0·3	0·951	0·405	4·844	0·969	0·424	1·04	1·043	0·413	1·18	1·065	0·435	1·24
0·4	1·283	0·439	8·261	1·308	0·455	1·04	1·347	0·440	1·09	1·353	0·461	1·11
						$n = 100$						
0·1	0·163	0·280	0·436	0·165	0·442	11·70	0·189	0·335	3·45	0·209	0·339	4·23
0·2	0·554	0·223	1·702	0·563	0·233	1·04	0·640	0·218	1·27	0·668	0·235	1·38
0·3	0·971	0·261	4·657	0·993	0·272	1·05	1·075	0·266	1·21	1·090	0·279	1·25
0·4	1·298	0·298	8·060	1·328	0·307	1·05	1·366	0·299	1·10	1·369	0·314	1·11

Estimators: Lin–Carroll, Lin & Carroll (2000); Wang's first step and Wang's estimator after iterations, Wang (2003). 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 $(i = 1, \ldots, 1000)$ for proposed method; RMISE, MISE as a multiple of MISE_1.

critically on the sample size and the choice of bandwidth. On the other hand, all MISE ratios are greater than 1, indicating that the proposed method outperforms the other three methods. It also shows that the improvement achieved by the proposed estimator can be quite substantial for smaller bandwidths.

We apply the proposed method to a real example. Raz (1989) presented luteinising hormone levels, in ng/ml, in 16 suckled and 16 non-suckled cows at times 1, 2, 3, 4, 5, 5·5, 6, 6·5, 7, 7·5, 8, 8·5, 9, 9·5 and 10 days after their postpartum. Figure 1 shows the profile plot of the raw data. It is of interest to compare the two groups of cows, and visual comparison from Fig. 1 is not easy.

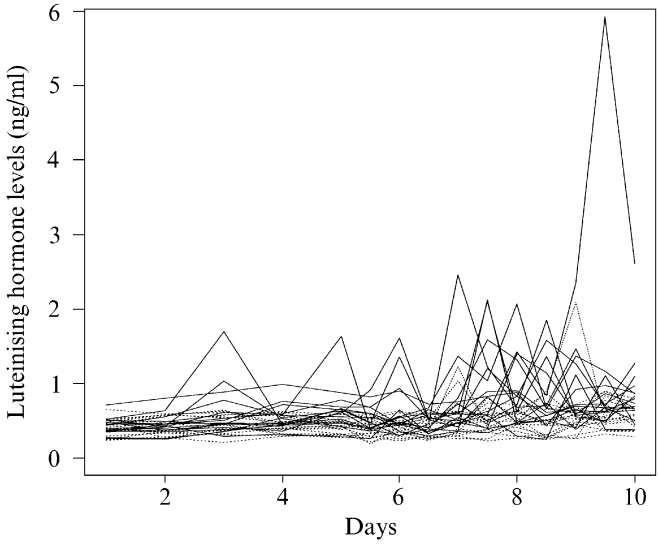

Fig. 1. The profile plot of cow luteinising hormone data; 16 solid lines for nonsuckled cows, 16 dotted lines for suckled cows.

We therefore consider local linear, $p = 1$, and local cubic, $p = 3$, fits separately for each group. Our estimates use two different working within-cluster correlation matrices, the identity matrix I and the estimated correlation matrix $\hat{R}_0$, where $\hat{R}_0$ was based on the residuals from the fit with identity working within-cluster correlation matrix under the assumption that the two groups have identical correlation structures. The global optimal bandwidth in Remark 2 was estimated by mimicking the rule-of-thumb global bandwidth selector in Fan & Gijbels (1996, p. 110) with a constant weight function $w(.) \equiv 1$. With initial global bandwidths ranging from 1 to 6, the approach yielded estimates of the optimal global bandwidth with a range of 1·99 to 4·69 and mean 3·01. Consequently, bandwidth $h = 3$ was used in our computation. We found that the rule-of-thumb global bandwidth selection is easy to apply and reliable. Other competitive approaches, such as crossvalidation and the empirical-bias bandwidth selector (Ruppert, 1997), can also be applied. The two curves were estimated on the grid points $0{\cdot}5 + 0{\cdot}25(j-1)$, for $j = 1, \ldots, 41$, with the Epanechnikov kernel. The results are contained in Fig. 2(a) for the local linear fit and Fig. 2(b) for the local cubic fit, respectively. It is evident from Figs 1 and 2 that nonsuckled cows typically have higher luteinising hormone levels than suckled cows. It is also clear that different working within-cluster correlation matrices lead to different curve estimates.

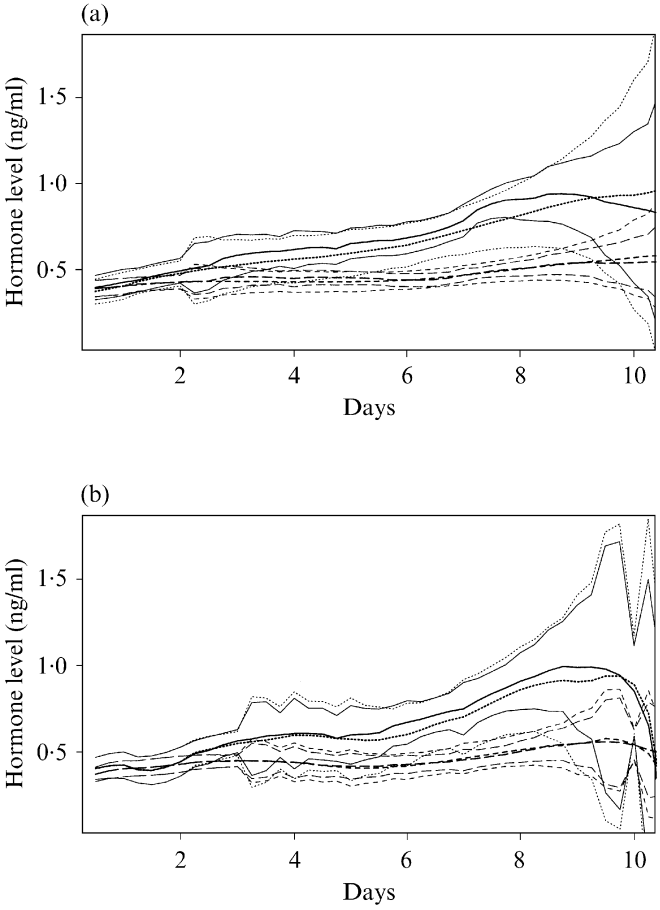

Fig. 2. The estimated curves of cow luteinising hormone levels with ± 2 pointwise standard errors (a) for $p = 1$ and (b) for $p = 3$: for nonsuckled cows using estimated correlation matrix, solid; for nonsuckled cows using identity correlation matrix, dotted; for suckled cows using estimated correlation matrix, long-dashed; for suckled cows using identity correlation matrix, short-dashed. The thicker curves are the estimated curves.

ACKNOWLEDGEMENT

The authors are very grateful to Professors R. Carroll, X. Lin, M. Paik and N. Wang for their constructive comments and suggestions and constant encouragement. This research was supported by a Hong Kong Research Grant Council grant, a U.S. National Science Foundation career award and the New York City Council Speaker's Fund for Public Health Research. The authors thank the editor and the associate editor for their helpful comments.

APPENDIX

Proofs

Proof of Proposition 1. (i) Set $H = \mathrm{diag}(1, h, \ldots, h^p)$ and $c_{ijs} = (x_{ij} - x_0)^s \{K_h(x_{ij} - x_0)/\phi_j\}^{1/2}$, for $i = 1, \ldots, n$, $j = 1, \ldots, J$ and $s = 0, \ldots, p$. For every fixed $v = 1, \ldots, 2^J - 1$, let

$$\mathscr{S}_v(h) = \{x_{1j} \in B(x_0, h) \text{ for all } j \in \Omega_v, \text{ and } x_{1j} \notin B(x_0, h) \text{ for all } j \notin \Omega_v\}.$$

Recall that $\mathscr{S}_v(0) = \{x_{1j} = x_0 \text{ for all } j \in \Omega_v, \text{ and } x_{1j} \neq x_0 \text{ for all } j \notin \Omega_v\}$. The condition of the existence of partial density ensures that $\text{pr}\{x_{1j} \text{ are all equal for all } j \in \Omega_v | \mathscr{S}_v(h)\} = 1 + o(1)$ with their conditional marginal densities being proportional to $f_v(.)\{1 + o(1)\}$ on $B(x_0, h)$ as $h \to 0$. Let $a_{m+1,l+1}$ denote the $(m+1, l+1)$th element of A_n, for $0 \leqslant m$ and $l \leqslant p$. Let $j_v \in \Omega_v$. Then, for $0 \leqslant m$ and $l \leqslant p$,

$$
\begin{aligned}
E(a_{m+1,l+1}) &= \sum_{i=1}^{n} E\left\{(c_{i1m} \ldots c_{iJm})(I_i R_i I_i)^{-1} \begin{pmatrix} c_{i1l} \\ \vdots \\ c_{iJl} \end{pmatrix}\right\} \\
&= n \sum_{v=1}^{2^J-1} E\left\{(c_{11m} \ldots c_{1Jm})(I_1 R_1 I_1)^{-1} \begin{pmatrix} c_{11l} \\ \vdots \\ c_{1Jl} \end{pmatrix} \middle| \mathscr{S}_v(h)\right\} \text{pr}\{\mathscr{S}_v(h)\} \\
&= n \sum_{v=1}^{2^J-1} E[(x_{1j_v} - x_0)^{m+l} K_h(x_{1j_v} - x_0) I\{\mathscr{S}_v(h)\}] \\
&\quad \times E\{1_0^{\mathrm{T}} \Phi^{-1/2} (I_{v0} R_1 I_{v0})^{-1} \Phi^{-1/2} 1_0 | \mathscr{S}_v(0)\}\{1 + o(1)\} \\
&= n \sum_{v=1}^{2^J-1} \int (x - x_0)^{m+l} f_v(x) \frac{1}{h} K\left(\frac{x - x_0}{h}\right) dx \\
&\quad \times E\{1_0^{\mathrm{T}} \Phi^{-1/2} (I_{v0} R_1 I_{v0})^{-1} \Phi^{-1/2} 1_0 | \mathscr{S}_v(0)\}\{1 + o(1)\} \\
&= n \sum_{v=1}^{2^J-1} h^{m+l} \int t^{m+l} K(t) f_v(x_0 + ht) dt \\
&\quad \times E\{1_0^{\mathrm{T}} \Phi^{-1/2} (I_{v0} R_1 I_{v0})^{-1} \Phi^{-1/2} 1_0 | \mathscr{S}_v(0)\}\{1 + o(1)\} \\
&= n h^{m+l} \mu_{m+l} \sum_{v=1}^{2^J-1} f_v(x_0) \xi_v \{1 + o(1)\}.
\end{aligned}
$$

Similarly, we can show that $\{\text{var}(a_{m+1,l+1})\}^{1/2} = o(nh^{m+l})$. Then

$$
a_{m+1,l+1} = E(a_{m+1,l+1}) + O_P[\{\text{var}(a_{m+1,l+1})\}^{1/2}] = n h^{m+l} \mu_{m+l} \sum_{v=1}^{2^J-1} f_v(x_0) \xi_v \{1 + o_P(1)\}.
$$

Therefore,

$$
A_n = n \left\{\sum_{v=1}^{2^J-1} f_v(x_0) \xi_v\right\} HSH\{1 + o_P(1)\}.
$$

By a similar calculation, it follows that

$$
B_n = nh^{-1} \left\{\sum_{v=1}^{2^J-1} f_v(x_0) \bar{\xi}_v\right\} H\bar{S}H\{1 + o_P(1)\}.
$$

Therefore, it follows from (4) that

$$
\begin{aligned}
\text{var}\{\hat{m}_k(x_0) | \mathscr{F}_n^X\} &= k!^2 \sigma^2(x_0) e_{k+1}^{\mathrm{T}} A_n^{-1} B_n A_n^{-1} e_{k+1} \{1 + o_P(1)\} \\
&= \frac{k!^2}{nh^{1+2k}} e_{k+1}^{\mathrm{T}} S^{-1} \bar{S} S^{-1} e_{k+1} \frac{\sigma^2(x_0) \sum_{l=1}^{2^J-1} f_l(x_0) \bar{\xi}_l}{\{\sum_{l=1}^{2^J-1} f_l(x_0) \xi_l\}^2} + o_P\left(\frac{1}{nh^{1+2k}}\right).
\end{aligned}
$$

Thus, (6) holds.

Similarly to the asymptotic expansion of A_n, one can show that

$$\sum_{i=1}^{n} X_i^{\mathrm{T}} W_i \begin{pmatrix} (x_{i1}-x_0)^{p+1} \\ \vdots \\ (x_{iJ}-x_0)^{p+1} \end{pmatrix} = nh^{p+1} \sum_{v=1}^{2^J-1} \{f_v(x_0)\xi_v\} H c_p \{1+o_P(1)\}.$$

By Taylor expansion, the conditional bias of $\hat{\beta}$ is

$$\begin{aligned} E(\hat{\beta}|\mathscr{F}_n^X) - \beta &= \left(\sum_{i=1}^{n} X_i^{\mathrm{T}} W_i X_i\right)^{-1} \sum_{i=1}^{n} X_i^{\mathrm{T}} W_i \left\{ \begin{pmatrix} m(x_{i1}) \\ \vdots \\ m(x_{iJ}) \end{pmatrix} - X_i \beta \right\} \\ &= A_n^{-1} \sum_{i=1}^{n} X_i^{\mathrm{T}} W_i \begin{pmatrix} (x_{il}-x_0)^{p+1} \\ \vdots \\ (x_{iJ}-x_0)^{p+1} \end{pmatrix} \{\beta_{p+1} + o_P(1)\} \\ &= \left[n \left\{ \sum_{v=1}^{2^J-1} f_v(x_0)\xi_v \right\} HSH \right]^{-1} H c_p n h^{p+1} \sum_{v=1}^{2^J-1} \{f_v(x_0)\xi_v\} \{\beta_{p+1} + o_P(1)\} \\ &= H^{-1} S^{-1} c_p h^{p+1} \{\beta_{p+1} + o_P(1)\}. \end{aligned}$$

Recall that $\beta_l = m^{(l)}(x_0)/l!$ for $l \geqslant 0$. Then (7) follows.

If $p-k$ is even, it is well known that $e_{k+1}^{\mathrm{T}} S^{-1} c_p = 0$; see Fan & Gijbels (1996, p. 102). Thus, the leading term in (7) is 0.

(ii) To show that, for any bandwidth and any kernel K, the asymptotic variance is minimised when the modelled correlation matrix equals the true correlation matrix, it suffices to show that $\sum_{l=1}^{2^J-1} f_l(x_0)\bar{\xi}_l / \{\sum_{l=1}^{2^J-1} f_l(x_0)\xi_l\}^2$ is minimised when $R_1 = R_{10}$. Recall that 1_0 is the J-vector with all components equal to 1. Let

$$b_l = (\Phi^{1/2} I_{l0} R_{10} I_{l0} \Phi^{1/2})^{1/2} (\Phi^{1/2} I_{l0} R_1 I_{l0} \Phi^{1/2})^{-1} 1_0 \quad (1 \leqslant l \leqslant 2^J - 1).$$

Then

$$\begin{aligned} \frac{\sum_{l=1}^{2^J-1} f_l(x_0)\bar{\xi}_l}{\{\sum_{l=1}^{2^J-1} f_l(x_0)\xi_l\}^2} &= \frac{\sum_{l=1}^{2^J-1} f_l(x_0) E\{b_l^{\mathrm{T}} b_l | \mathscr{S}_l(0)\}}{[\sum_{l=1}^{2^J-1} f_l(x_0) E\{1_0^{\mathrm{T}} (\Phi^{1/2} I_{l0} R_{10} I_{l0} \Phi^{1/2})^{-1/2} b_l | \mathscr{S}_l(0)\}]^2} \\ &\geqslant \frac{1}{\sum_{l=1}^{2^J-1} f_l(x_0) E\{1_0^{\mathrm{T}} (\Phi^{1/2} I_{l0} R_{10} I_{l0} \Phi^{1/2})^{-1} 1_0 | \mathscr{S}_l(0)\}} \\ &= \left\{ \sum_{l=1}^{2^J-1} f_l(x_0)\xi_{l0} \right\}^{-1}, \end{aligned}$$

where the inequality follows from the Cauchy–Schwartz inequality, and equality holds if and only if $b_l = (\Phi^{1/2} I_{l0} R_{10} I_{l0} \Phi^{1/2}) 1_0$. By the definition of b_l, $R_1 = R_{10}$ implies that $b_l = (\Phi^{1/2} I_{l0} R_{10} I_{l0} \Phi^{1/2}) 1_0$. This proves that, for given bandwidth and kernel, the asymptotic variance is minimised when the modelled correlation matrix equals the true correlation matrix.

(iii)–(iv) It is a classical result that the minimum variance kernel minimising $e_{k+1}^{\mathrm{T}} S^{-1} \bar{S} S^{-1} e_{k+1}$ is the uniform kernel $0{\cdot}5I(|t| \leqslant 1)$; see for example Fan & Gijbels (1996, p. 75). Thus the proof of (ii) implies that, for any given bandwidth, the asymptotic variance is minimised with the true correlation matrix and the uniform kernel. In a similar fashion, the last two statements also follow from the proof of (ii) and the analogous results established for local polynomial regression for non-clustered data; see Fan & Gijbels (1996, Ch. 3). □

Proof of Corollary 1. (i) Without loss of generality, suppose that $\Omega_l = \{l\}$ for $l = 1, \ldots, J$. When the joint density of $(x_{11}, \ldots, x_{1J})^{\mathrm{T}}$ exists, it is seen that $f_l(x_0) = f_l^*(x_0)$, the marginal density of x_{1l} for $1 \leqslant l \leqslant J$, and that $f_l(x_0) = 0$ for $l > J$. Moreover, $\xi_l = \bar{\xi}_l = \phi_l^{-1}$ for $l = 1, \ldots, J$. Therefore, (9) follows from (6). Since the leading term in (9) is free of the modelled correlation matrix or the true correlation matrix, the asymptotic variance remains the same whether the modelled correlation matrix is correct or not.

(ii) Without loss of generality, suppose that $\Omega_{2^J-1} = \{1, \ldots, J\}$. Then $\mathrm{pr}\{x_{11} = \ldots = x_{1J}\} = 1$ implies that $f_l(x_0) = 0$ for all $1 \leqslant l \leqslant 2^J - 2$, and that $f_{2^J-1}(x_0) = f_1^*(x_0) = \ldots = f_J^*(x_0)$. Therefore, $\sum_{l=1}^{2^J-1} f_l(x_0)\xi_1 = f_1^*(x_0)\xi_*$ and $\sum_{l=1}^{2^J-1} f_l(x_0)\bar{\xi}_l = f_1^*(x_0)\bar{\xi}_*$. Then (10) follows from (6). □

Proof of Corollary 2. If we replace $(I_i R_i I_i)^{-1}$ by R_i^{-1}, the conditional asymptotic biases and variance (11) can be derived by the same steps as in the proof of Proposition 1 (i). Thus, we omit the details.

Since $\hat{m}_k(x_0)$ and $\hat{m}_k^*(x_0)$ have the same asymptotic biases, we only need to compare $\mathrm{var}\{\hat{m}_k(x_0)|\mathscr{F}_n^X\}$ with $\mathrm{var}\{\hat{m}_k^*(x_0)|\mathscr{F}_n^X\}$. By the Cauchy–Schwartz inequality, we have

$$\frac{\sum_{l=1}^J f_l^*(x_0) r_l^2/\phi_l}{\{\sum_{l=1}^J f_l^*(x_0) r_l/\phi_l\}^2} \geqslant \frac{1}{\sum_{l=1}^J f_l^*(x_0)/\phi_l},$$

in which equality holds if and only if $r_l = 1$ and $l = 1, \ldots, J$. This implies that

$$\mathrm{var}\{\hat{m}_k^*(x_0)|\mathscr{F}_n^X\} \geqslant \mathrm{var}\{\hat{m}_k(x_0)|\mathscr{F}_n^X\}$$

with equality holding if and only if R_i is the identity matrix. Thus, the last claim holds. □

Proof of Corollary 3. Equations (12) and (13) in part (i) are special cases of (6) and (7) in Proposition 1 with $k = 0$ and $p = 1$, and (ii) is a special case of (iv) of Proposition 1. The calculations are straightforward and we omit the details. □

Proof of Proposition 2. Let Z_i and D_i be any two $J \times (p+1)$ matrices. Observe that

$$\sum_{i=1}^n Z_i^{\mathrm{T}} Z_i - \left(\sum_{i=1}^n Z_i^{\mathrm{T}} D_i\right)\left(\sum_{i=1}^n D_i^{\mathrm{T}} D_i\right)^{-1}\left(\sum_{i=1}^n D_i^{\mathrm{T}} Z_i\right)$$
$$= \sum_{i=1}^n \left[\left\{Z_i - D_i\left(\sum_{j=1}^n D_j^{\mathrm{T}} D_j\right)^{-1}\left(\sum_{j=1}^n D_j^{\mathrm{T}} Z_j\right)\right\}^{\mathrm{T}}\left\{Z_i - D_i\left(\sum_{j=1}^n D_j^{\mathrm{T}} D_j\right)^{-1}\left(\sum_{j=1}^n D_j^{\mathrm{T}} Z_j\right)\right\}\right],$$

which is therefore always nonnegative definite. Moreover, this matrix is the zero matrix if and only if $Z_i = D_i C$, where C is any $(p+1) \times (p+1)$ matrix. Set

$$D_i = (I_i R_{i0} I_i)^{-1/2} \Phi^{-1/2} X_i, \quad Z_i = (I_i R_{i0} I_i)^{1/2} K_{ih}^{1/2} (I_i R_i I_i)^{-1} K_{ih}^{1/2} \Phi^{-1/2} X_i.$$

It is easily checked that

$$\sum_{i=1}^n Z_i^{\mathrm{T}} Z_i = B_n, \quad \sum_{i=1}^n D_i^{\mathrm{T}} Z_i = \sum_{i=1}^n Z_i^{\mathrm{T}} D_i = A_n, \quad \sum_{i=1}^n D_i^{\mathrm{T}} D_i = \sum_{i=1}^n X_i^{\mathrm{T}} \Phi^{-1/2} (I_i R_{i0} I_i)^{-1} \Phi^{-1/2} X_i.$$

Therefore, $B_n - A_n(\sum_{i=1}^n D_i^{\mathrm{T}} D_i)^{-1} A_n$ is always nonnegative definite, implying that

$$A_n^{-1} B_n A_n^{-1} - \left(\sum_{i=1}^n D_i^{\mathrm{T}} D_i\right)^{-1}$$

is always nonnegative definite. Furthermore, $A_n^{-1} B_n A_n^{-1} = (\sum_{i=1}^n D_i^{\mathrm{T}} D_i)^{-1}$ if and only if $Z_i = D_i C$ for any $(p+1) \times (p+1)$ matrix C. Note that $K_{ih} = I_i/2$ if $K(.)$ is the uniform kernel. Then $Z_i = D_i/2$ if $K(.)$ is the uniform kernel and $R_i = R_{i0}$, in which case

$$A_n^{-1} B_n A_n^{-1} = \left\{\sum_{i=1}^n X_i^{\mathrm{T}} \Phi^{-1/2} (I_i R_{i0} I_i)^{-1} \Phi^{-1/2} X_i\right\}^{-1}.$$

Now assume that the joint density of $(x_{11}, \ldots, x_{1J})^{\mathrm{T}}$ exists. Let

$$\mathcal{N} = \{1 \leqslant i \leqslant n : I_i \text{ contains exactly one nonzero element}\}.$$

It is straightforward to show that

$$B_n = \sum_{i \in \mathcal{N}} Z_i^{\mathrm{T}} Z_i \{1 + o_P(1)\}, \quad A_n = \sum_{i \in \mathcal{N}} D_i^{\mathrm{T}} Z_i \{1 + o_P(1)\} = \sum_{i \in \mathcal{N}} Z_i^{\mathrm{T}} D_i \{1 + o_P(1)\}.$$

Evidently, $Z_i = D_i/2$ for all $i \in \mathcal{N}$, if K is the uniform kernel. Then the claimed minimisation follows from an argument analogous to the above. The proof is complete. □

REFERENCES

FAN, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196–216.

FAN, J. & GIJBELS, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20**, 2008–36.

FAN, J. & GIJBELS, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. R. Statist. Soc.* B **57**, 371–94.

FAN, J. & GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications.* London: Chapman & Hall.

FAN, J., HECKMAN, N. E. & WAND, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Am. Statist. Assoc.* **90**, 141–50.

LIANG, K. Y. & ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

LIN, X. & CARROLL, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Am. Statist. Assoc.* **95**, 520–34.

RAZ, J. (1989). Analysis of repeated measurements using nonparametric smoothers and randomization tests. *Biometrics* **45**, 851–71.

RUCKSTUHL, A. F., WELSH, A. H. & CARROLL, R. J. (2000). Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statist. Sinica* **10**, 51–71.

RUPPERT, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Am. Statist. Assoc.* **92**, 1049–62.

RUPPERT, D. & WAND, M. P. (1994). Multivariate weighted least squares regression. *Ann. Statist.* **22**, 1346–70.

SEVERINI, T. A. & STANISWALIS, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *J. Am. Statist. Assoc.* **89**, 501–11.

VERBYLA, A. P., CULLIS, B. R., KENWARD, M. G. & WELHAM, S. J. (1999). The analysis of designed experiments and longitudinal data using smoothing splines (with Discussion). *Appl. Statist.* **48**, 269–311.

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

WILD, C. J. & YEE, T. W. (1996). Additive extensions to generalized estimating equations methods. *J. R. Statist. Soc.* B **58**, 711–25.

ZEGER, S. L. & DIGGLE, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689–99.

[*Received July* 2003. *Revised April* 2004]