

Nonparametric additive regression for repeatedly measured data

BY RAYMOND J. CARROLL

Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
carroll@stat.tamu.edu

ARNAB MAITY

Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.
amaity@hsph.harvard.edu

ENNO MAMMEN AND KYUSANG YU

Department of Economics, University of Mannheim, L 7, 3-5, 68131 Mannheim, Germany
emammen@rumms.uni-mannheim.de kysangu@yahoo.co.kr

SUMMARY

We develop an easily computed smooth backfitting algorithm for additive model fitting in repeated measures problems. Our methodology easily copes with various settings, such as when some covariates are the same over repeated response measurements. We allow for a working covariance matrix for the regression errors, showing that our method is most efficient when the correct covariance matrix is used. The component functions achieve the known asymptotic variance lower bound for the scalar argument case. Smooth backfitting also leads directly to design-independent biases in the local linear case. Simulations show our estimator has smaller variance than the usual kernel estimator. This is also illustrated by an example from nutritional epidemiology.

Some key words: Additive model; Generalized least square; Nonparametric regression; Repeated measure; Smooth backfitting.

1. INTRODUCTION

We consider efficient estimation of an additive nonparametric regression model from repeated measures data when the covariates are multivariate. To date, while there is a considerable literature in the scalar covariate case, see below, the problem has not been addressed in the multivariate additive model case. Ours represents a first contribution in this direction.

There has been much interest in the simplest version of this problem. Suppose that there are $i = 1, \dots, n$ individuals, and $j = 1, \dots, J$ observations per individual. The responses are Y_{ij} and the scalar predictors are X_{ij} . A simple model says that given (X_{i1}, \dots, X_{iJ}) ,

$$Y_{ij} = m_{\text{true}}(X_{ij}) + \epsilon_{ij}, \quad \text{cov}(\tilde{\epsilon}_i) = \text{cov}(\epsilon_{i1}, \dots, \epsilon_{iJ})^T = \Sigma_{\text{true}}, \quad (1)$$

where $\tilde{\epsilon}$ has mean zero.

Much of the literature for estimating $m_{\text{true}}(\cdot)$ in model (1) has used the kernel regression framework for theoretical convenience. The majority of this literature has been based upon the idea of ignoring the covariance matrix Σ_{true} and fitting that model as if there were no correlation, fixing up the standard errors later. This is the so-called working independence or pooled data method, which has been described in different variants by Hoover et al. (1998), Lin & Carroll (2000, 2001), Lin & Ying (2001) and Chen & Jin (2005), among many others. Early on it was recognized that naive methods of accounting for the correlation could have problems (Lin & Carroll, 2000) and lead to losses of efficiency in comparison to working independence. Ruckstuhl et al. (2000) and Linton et al. (2004) developed two-step methods that they showed improved upon working independence.

Efficient estimation of the regression function $m_{\text{true}}(\cdot)$ in model (1) was first solved in the kernel context by Wang (2003); see also Lin et al. (2004) who showed that Wang's iterative method had an exact solution, and Huggins (2006) for an alternative and simpler exact solution. The method has been extended to general likelihood models, these generalizations providing efficient inference in semiparametric problems (Lin & Carroll, 2006). A disadvantage of the method is that unlike for ordinary local linear kernel regression, the asymptotic bias is not design-independent (Fan & Gijbels, 1996), i.e. it depends on the density function of the predictors.

In this paper, we consider repeated measures models such as (1) in the case that the argument of $m_{\text{true}}(\cdot)$ is multivariate, rather than as a scalar. As is usual, to avoid the curse of dimensionality we take an additive modelling approach. For $i = 1, \dots, n$, we observe a random sample (Y_i, X_i) , where $Y_i = (Y_{i1}, \dots, Y_{iJ})^\top$ and $X_i = (X_{i11}, \dots, X_{i1J}, \dots, X_{iD1}, \dots, X_{iDJ})^\top$ with

$$Y_{ij} = m_{0,\text{true}} + \sum_{d=1}^D m_{d,\text{true}}(X_{idj}) + \epsilon_{ij}, \quad \tilde{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{iJ})^\top \sim (0, \Sigma_{\text{true}}), \quad (2)$$

where Σ_{true} has the elements $\sigma_{\ell k, \text{true}}$ and is positive definite. We have J repeated measurements and D regressors from n individuals. Assume X has the density $p(\cdot)$. We denote the density of $(X_{11\ell}, \dots, X_{1D\ell})^\top$ as $p_{\ell\ell}(\cdot)$, the density of $(X_{1sk}, X_{1b\ell})^\top$ as $p_{k\ell}^{sb}$ and the density of X_{1dk} as p_k^d .

To fit model (2), we generalize the idea of smooth backfitting for independent data and time series developed by Nielsen & Linton (1998) and Mammen et al. (1999); see also Mammen & Park (2005, 2006), Mammen & Nielsen (2003), Nielsen & Sperlich (2005) and Yu et al. (2008). Linton & Mammen (2008) apply smooth backfitting for dependent data but with dependence coming from autoregressive errors. We will show that our repeated measures smooth backfitting algorithm achieves the same efficiency component-wise as achieved by Wang (2003) in the single function problem, and is in addition automatically design independent.

2. ESTIMATOR

2.1. Notation and Definitions

The basic idea of our approach is to linearly transform the data such that the errors are uncorrelated and to use localization and smoothing on the transformed data.

Write the symmetric matrix $\Sigma^{-1/2}$ to have (j, k) element $a_{jk, \Sigma}$. Then consider the transformed data $\mathcal{Y}_i^\Sigma = (\mathcal{Y}_{i1}^\Sigma, \dots, \mathcal{Y}_{iJ}^\Sigma)^\top = \Sigma^{-1/2} Y_i$, so that

$$\mathcal{Y}_{ij}^\Sigma = \sum_{k=1}^J a_{jk, \Sigma} \left\{ m_{0,\text{true}} + \sum_{d=1}^D m_{d,\text{true}}(X_{idk}) \right\} + U_{ij}. \quad (3)$$

For $\Sigma = \Sigma_{\text{true}}$, it follows that $\tilde{U}_i = (U_{i1}, \dots, U_{iJ})^\top$ has mean zero and identity covariance matrix.

As seen in (3), the transformed data preserve or produce additivity in the regression model when $D > 1$ or $D = 1$, respectively. Based on this observation, we apply the smooth backfitting framework to fit the additive models. We will use two approaches: local constant and local linear smoothing; see § 2.2 and Appendix A.2. We use boundary-corrected kernels. Let K^0 be a base kernel function and $K_h^0(u) = h^{-1}K^0(u/h)$. Define a boundary-corrected kernel function by $K_h(u, v) = I(u, v \in S)K_h^0(u - v)/\int_S K_h^0(w - v)dw$, where S is the support.

We allow for estimation of the covariance matrix Σ_{true} . Let the (j, k) elements of $\Sigma_{\text{true}}^{-1}$ be $b_{jk, \text{true}}$, and for an arbitrary covariance matrix Σ , let the elements of Σ^{-1} be $b_{jk, \Sigma}$, so that $\sum_{\ell} a_{j\ell, \Sigma} a_{k\ell, \Sigma} = b_{jk, \Sigma}$. One choice of Σ is the covariance estimator based on the residuals of the pooled data, but more complex generalizations are easily handled. The component functions in the additive regression are identifiable only up to a constant and thus we identify the component functions with the condition

$$\int \sum_{\ell=1}^J \sum_{k=1}^J b_{k\ell, \Sigma} m_{d, \text{true}}(x_{d\ell}) p_{\ell}^d(x_{d\ell}) dx_{d\ell} = 0.$$

2.2. Local constant estimator

Here we discuss the construction of local constant estimators, while local linear estimators are derived in Appendix A.2. Let e be a vector of ones and define $\hat{m}_0^{\Sigma} = n^{-1} \sum_{i=1}^n e^T \Sigma^{-1} Y_i / e^T \Sigma^{-1} e$. For a local constant fit, the smoothed sum of squares is defined by

$$\begin{aligned} S^{\Sigma}(m_1, \dots, m_D) &= \int n^{-1} \sum_{i=1}^n \sum_{j,k=1}^J b_{jk, \Sigma} \left\{ Y_{ij} - \hat{m}_0^{\Sigma} - \sum_{d=1}^D m_d(x_{dj}) \right\} \\ &\quad \times \left\{ Y_{ik} - \hat{m}_0^{\Sigma} - \sum_{d=1}^D m_d(x_{dk}) \right\} \\ &\quad \times \prod_{k=1}^J K_{h_1}(x_{1k}, X_{i1k}) \cdots K_{h_D}(x_{Dk}, X_{iDk}) dx. \end{aligned} \tag{4}$$

Here, $x = (x_{11}, \dots, x_{1J}, \dots, x_{D1}, \dots, x_{DJ})^T$. The goal of smooth backfitting is to minimize (4) in the functions (m_1, \dots, m_D) .

For $j, k = 1, \dots, J$ and $a, b = 1, \dots, D$, define estimators $\hat{p}_k^a(x) = n^{-1} \sum_{i=1}^n K_{h_a}(x, X_{iak})$ and $\hat{p}_{jk}^{ab}(x, y) = n^{-1} \sum_{i=1}^n K_{h_a}(x, X_{iaj}) K_{h_b}(y, X_{ibk})$. In Appendix A.1, we show that the minimizer $(\hat{m}_1, \dots, \hat{m}_D)$ of (4) exists and is the solution to the integral equations

$$\begin{aligned} m_d(x_d) &= \tilde{m}_d^{\Sigma}(x_d) - \int m_d(t) \frac{\sum_{j=1}^J \sum_{k \neq j}^J b_{jk, \Sigma} \hat{p}_{jk}^{dd}(x_d, t)}{\sum_{j=1}^J b_{jj, \Sigma} \hat{p}_j^d(x_d)} dt \\ &\quad - \sum_{s=1, \neq d}^D \int m_s(t) \frac{\sum_{j=1}^J \sum_{k=1}^J b_{jk, \Sigma} \hat{p}_{jk}^{ds}(x_d, t)}{\sum_{j=1}^J b_{jj, \Sigma} \hat{p}_j^d(x_d)} dt, \end{aligned} \tag{5}$$

where $\tilde{m}_d^{\Sigma}(x) = n^{-1} \sum_{i=1}^n \sum_{j,k=1}^J b_{jk, \Sigma} (Y_{ik} - \hat{m}_0^{\Sigma}) K_{h_d}(x, X_{idj}) / \sum_{j=1}^J b_{jj, \Sigma} \hat{p}_j^d(x)$, and, as stated previously, Σ^{-1} has (j, k) element $b_{jk, \Sigma}$. Here and later on we use the convention that a summation over an empty index set is equal to zero. We impose the identification condition

$$\int \sum_{\ell=1}^J \sum_{k=1}^J b_{k\ell, \Sigma} m_d(x_{d\ell}) \hat{p}_{\ell}^d(x_{d\ell}) dx_{d\ell} = 0. \tag{6}$$

It is easy to see that the solution of equation (5) satisfies the identification condition (6). An algorithm for the implementation of (5) is discussed below in § 2.3.

2.3. *Implementation*

There is no need to compute possibly high-dimensional integrals such as (4). Here we derive a straightforward iterative algorithm that includes only one-dimensional direct inversions. Define the operators

$$\begin{aligned}
 (\hat{\mathcal{G}}_{dd}^\Sigma f)(x) &= \int f(t) \frac{\sum_{j=1}^J \sum_{k \neq j, =1}^J b_{jk, \Sigma} \hat{p}_{jk}^{dd}(x, t)}{\sum_{j=1}^J b_{jj, \Sigma} \hat{p}_j^d(x)} dt, \\
 (\hat{\mathcal{G}}_{ds}^\Sigma f)(x) &= \int f(t) \frac{\sum_{j=1}^J \sum_{k=1}^J b_{jk, \Sigma} \hat{p}_{jk}^{ds}(x, t)}{\sum_{j=1}^J b_{jj, \Sigma} \hat{p}_j^d(x)} dt.
 \end{aligned}$$

Equation (5) can be rewritten as $\{(I + \hat{\mathcal{G}}_{dd}^\Sigma)m\}(x_d) = \tilde{m}_d^\Sigma(x_d) - \sum_{s=1, \neq d}^D (\hat{\mathcal{G}}_{ds}^\Sigma \hat{m}_s)(x_d)$, where I is the identity mapping. Based upon this observation, we build up a backfitting-type algorithm, as follows.

Step 1. Set initial values $m_d^{[0]}$ for $d = 1, \dots, D$ that satisfy the identification condition.

Step 2. Let $m_d^{[k]}$ be the estimates of the d th function in the k th iteration. The updating equation is defined as

$$\{(I + \hat{\mathcal{G}}_{dd}^\Sigma)m^{[k]}\}(x_d) = \tilde{m}_d^\Sigma(x_d) - \sum_{s=1}^{d-1} (\hat{\mathcal{G}}_{ds}^\Sigma m_s^{[k]})(x_d) - \sum_{s=d+1}^D (\hat{\mathcal{G}}_{ds}^\Sigma m_s^{[k-1]})(x_d).$$

Step 3. Iterate until convergence; see Appendix A.4 for a proof of convergence.

A similar algorithm is used in the local linear case; see Appendices A.2 and A.3.

3. ASYMPTOTIC PROPERTIES

3.1. *Main results*

One advantage of the additive model is that it makes it possible to achieve the one-dimensional convergence rate. Here, we choose the bandwidths h_d so that $n^{1/5}h_d$ converges to constants $c_d > 0$ for $d = 1, \dots, D$ as n goes to infinity. This choice of bandwidth is known to be optimal when the regression function is twice continuously differentiable. We give the formal assumptions in Appendix A.5 and sketched proofs in Appendix A.6.

Let Σ be any positive definite matrix. Let $P^d(x_d) = \text{diag}\{p_1^d(x_d), \dots, p_J^d(x_d)\}$. We say that $\hat{\Sigma}$ converges to Σ in probability if $\sup_{x: |x|_2=1} x^T(\hat{\Sigma} - \Sigma)x = o_p(1)$. Define, for $d = 1, \dots, D$,

$$\begin{aligned}
 v_d^\Sigma(x_d) &= c_d^{-1} \mu_0 \{(K^0)^2\} \text{trace}\{\Sigma^{-1} \Sigma_{\text{true}} \Sigma^{-1} P^d(x_d)\} \left\{ \sum_{j=1}^J b_{jj, \Sigma} p_j^d(x_d) \right\}^{-2}, \\
 \begin{Bmatrix} \beta_1^\Sigma(x_1) \\ \vdots \\ \beta_D^\Sigma(x_D) \end{Bmatrix} &= \frac{1}{2} \mu_2(K^0) \begin{Bmatrix} h_1^2 m''_{1, \text{true}}(x_1) \\ \vdots \\ h_D^2 m''_{D, \text{true}}(x_D) \end{Bmatrix} + \mu_2(K^0) (I - \mathcal{G}^\Sigma)^{-1} \begin{Bmatrix} \gamma_1^\Sigma(x_1) \\ \vdots \\ \gamma_D^\Sigma(x_D) \end{Bmatrix},
 \end{aligned}$$

where $\mu_r(K) = \int t^r K(t)dt$ for a function K and $\gamma_d^\Sigma(x_d)$ is of order $O(h_d^2)$ and is given in (A15). The operator \mathcal{G}^Σ is defined in Appendix A.3.

THEOREM 1 (Asymptotic distributions of the local constant estimator). *Under conditions (A1)–(A5) in Appendix A.1, if the positive definite estimator $\hat{\Sigma}$ converges to Σ in probability, then for any $x_1, \dots, x_d \in (0, 1)$, we have the following convergences in distribution*

$$n^{2/5} \begin{Bmatrix} \hat{m}_1^{\hat{\Sigma}}(x_1) - m_{1,\text{true}}(x_1) - \beta_1^\Sigma(x_1) \\ \vdots \\ \hat{m}_D^{\hat{\Sigma}}(x_D) - m_{D,\text{true}}(x_D) - \beta_D^\Sigma(x_D) \end{Bmatrix} \rightarrow N \left[0, \text{diag}\{v_1^\Sigma(x_1), \dots, v_D^\Sigma(x_D)\} \right],$$

$$n^{1/2}(\hat{m}_0^{\hat{\Sigma}} - m_{0,\text{true}}) \rightarrow N \left\{ 0, \frac{e^\top \Sigma^{-1} \Sigma_{\text{true}} \Sigma^{-1} e}{(e^\top \Sigma^{-1} e)^2} \right\}.$$

If the covariance estimator $\hat{\Sigma}$ converges to Σ_{true} in probability, for $d = 1, \dots, D$, we have $\beta_{d,\text{true}}(x_d)$ defined with Σ_{true} instead of Σ and

$$v_{d,\text{true}}(x_d) = c_d^{-1} \mu_0\{(K^0)^2\} \left\{ \sum_{\ell=1}^J b_{\ell\ell,\text{true}} p_\ell^d(x_d) \right\}^{-1},$$

$$n^{1/2}(\hat{m}_0^{\hat{\Sigma}} - m_{0,\text{true}}) \rightarrow N\{0, (e^\top \Sigma_{\text{true}}^{-1} e)^{-1}\}.$$

As seen above, for local constant estimation the asymptotic bias of the estimator of $m_{d,\text{true}}$ depends on the choice of Σ and also on the other functions $m_{s,\text{true}}, s \neq d$. For local linear estimation, however, the asymptotic biases depend neither on the choice of Σ nor on the other additive functions. This is the result of the next theorem.

THEOREM 2 (Asymptotic distributions of local linear estimator). *Under the same conditions as in Theorem 1, the local linear estimator has the same asymptotic distribution as the local constant estimator but with design-independent bias terms $\beta_d^{LL}(x_d) = (1/2)\mu_2(K^0)h_d^2 m_{d,\text{true}}''(x_d)$. The local linear estimator achieves the same asymptotic biases and variances as the infeasible single-dimensional estimators applied to the models with known functions $m_s, s \neq d$.*

The next theorem shows that one achieves the minimum asymptotic variance if one uses the correct covariance matrix.

THEOREM 3 (Optimal asymptotic variance). *Under the same conditions as in Theorem 1, we have, for $d = 1, \dots, D$, $v_d^\Sigma(x_d) \geq v_{d,\text{true}}(x_d)$ for any positive definite Σ . Equality holds when $\Sigma = c\Sigma_{\text{true}}$ for some $c > 0$.*

Remark 1. The results in Theorems 1 and 2 do not require that the covariance estimator converge to the true covariance.

Remark 2. Our method is new even with the choice of $\hat{\Sigma} = I$, also known as working independence with common variances. In that case, the standard analysis ignores the repeated measures entirely. Our method, in contrast, differs because we account for the possibly different densities across time within the subject.

Remark 3. In the one-dimensional case, the asymptotic distributions coincide with those of Wang (2003), if we choose an undersmoothed starting bandwidth in Wang’s method.

Remark 4. An expansion similar to that of Lin & Carroll (2006) to order $o_p(n^{-1/2})$ is also possible; see Mammen & Park (2005) for the nonrepeated measures case. This expansion can be used in the discussion of data-adaptive bandwidth choice; see Mammen & Park (2005).

Remark 5. In some problems, it might be the case that the covariate terms X_{idj} are repeated across replicates, the extreme case being baseline covariates such that $X_{idj} \equiv X_{id}$ for $d = 1, \dots, D$ does not vary across replications. Efficient methods for these cases can also be constructed: we do not do so here for the reason of brevity, but some details are available in a long version of the paper at <http://mammen.vwl.uni-mannheim.de>.

3.2. Bandwidth selection

In this section, we describe simple bandwidth selection strategies. The methods are facilitated by noting that the bias and the variance of the d th estimated function depend only upon the d th bandwidth, and not on any other bandwidth.

To emphasize the dependence upon bandwidths, we denote the asymptotic biases in Theorems 1 and 2 as $\beta_d^\Sigma(h, x_d)$ and $\beta_d^{LL}(h_d, x_d)$ instead of $\beta_d^\Sigma(x_d)$ and $\beta_d^{LL}(x_d)$ ($d = 1, \dots, D$). The weighted asymptotic mean integrated squared errors are given by

$$\mathcal{T}_{\text{MSE}}(h_1, \dots, h_D, \hat{m}^{LC}) = \int \left[\left\{ \sum_{d=1}^D \beta_d^\Sigma(h, x_d) \right\}^2 + \sum_{d=1}^D \frac{c_d}{nh_d} v_d^\Sigma(x_d) \right] w(x) dx, \quad (7)$$

$$\mathcal{T}_{\text{MSE}}(h_1, \dots, h_D, \hat{m}^{LL}) = \int \left[\left\{ \sum_{d=1}^D \beta_d^{LL}(h_d, x_d) \right\}^2 + \sum_{d=1}^D \frac{c_d}{nh_d} v_d^\Sigma(x_d) \right] w(x) dx, \quad (8)$$

where \hat{m}^{LC} and \hat{m}^{LL} are the local constant and local linear estimator of the additive function, respectively, and w is a weight function. The optimal choices of bandwidths for the local constant and local linear estimators minimize (7) and (8), respectively.

A plug-in method minimizes an estimated mean integrated squared error obtained by plugging in the estimates of unknown quantities in \mathcal{T}_{MSE} . Then $\mathcal{T}_{\text{MSE}}(h_1, \dots, h_D, \hat{m}^{LC})$ involves first and second derivatives of component functions, one- and two-dimensional marginal densities and their derivatives and the true covariance, while $\mathcal{T}_{\text{MSE}}(h_1, \dots, h_D, \hat{m}^{LL})$ involves only the second derivatives of component functions, one-dimensional marginal densities and the true covariance. One- and two-dimensional marginal densities and their derivatives and the true covariance matrix can be estimated via standard kernel smoothing methods. Estimates of the first and second derivatives of component functions can be obtained from the estimates of the component functions by numerical differentiation or by local quadratic approximation, as suggested by Mammen & Park (2005).

The terms $\beta_d^{LL}(h_d, x_d)$ and $v_d^\Sigma(x_d)$ involve only h_d but do not depend on any of the other bandwidths. Thus we can consider componentwise optimal bandwidths. Define the componentwise weighted asymptotic mean integrated squared errors as

$$\mathcal{T}_{d, \text{MSE}}(h_d, \hat{m}_d^{LL}) = \int \left[\left\{ \beta_d^{LL}(h_d, x_d) \right\}^2 + \frac{c_d}{nh_d} v_d^\Sigma(x_d) \right] w_d(x_d) dx_d,$$

where \hat{m}_d^{LL} ($d = 1, \dots, D$) are the local linear estimators of component functions and w_d are weight functions. We also define $h_{d,\text{MSE}}$ as the minimizer of $\mathcal{T}_{d,\text{MSE}}(h_d, \hat{m}_d^{LL})$ for $d = 1, \dots, D$. Define $A_d(w_d) = \{\mu_2(K^0)\}^2 \int \{m''_{d,\text{true}}(x_d)\}^2 w_d(x_d) dx_d$ and define

$$B_d(\Sigma, w_d) = \mu_0\{(K^0)^2\} \int \text{trace}\{\Sigma^{-1} \Sigma_{\text{true}} \Sigma^{-1} P^d(x_d)\} \left\{ \sum_{j=1}^J b_{jj,\Sigma} p_j^d(x_d) \right\}^{-2} w_d(x_d) dx_d.$$

Then, we have the explicit formula for $h_{d,\text{MSE}}$ given as

$$h_{d,\text{MSE}} = n^{-1/5} \left\{ \frac{B_d(\Sigma, w_d)}{A_d(w_d)} \right\}^{1/5}.$$

Thus, in this case we can obtain the data-driven bandwidths by plugging the estimates of $m''_{d,\text{true}}(x_d)$, $p_j^d(x_d)$, Σ and Σ_{true} into $A_d(w_d)$ and $B_d(\Sigma, w_d)$.

Only $B_d(\Sigma, w_d)$ depends on Σ . This implies if we have the optimal bandwidths for the choice $\Sigma = \Sigma_1$, we can obtain the optimal bandwidths for another choice $\Sigma = \Sigma_2$ simply by multiplying the factors $\{B_d(\Sigma_2, w_d)/B_d(\Sigma_1, w_d)\}^{1/5}$ by the optimal bandwidths for the choice $\Sigma = \Sigma_1$. The most interesting case might be when one compares $\Sigma = \Sigma_{\text{true}}$ and $\Sigma = I$. In this case, we have $B_d(\Sigma_{\text{true}}, w_d) = \mu_0\{(K^0)^2\} \int \{\sum_{j=1}^J b_{jj,\Sigma_{\text{true}}} p_j^d(x_d)\}^{-1} w_d(x_d) dx_d$ and $B_d(I, w_d) = \mu_0\{(K^0)^2\} \int \{\sum_{j=1}^J b_{jj,\Sigma_{\text{true}}} p_j^d(x_d)\} \{\sum_{j=1}^J p_j^d(x_d)\}^{-2} w_d(x_d) dx_d$.

4. SIMULATIONS

In this section, we discuss finite sample properties of the proposed estimators via simulation studies. We will compare our method with the ordinary kernel estimator using pooled data, i.e. working independence ignoring the correlation structure.

The sample size was 200, the bandwidth was 0.1, there were three repeated measures and the grid size for integration was 0.01. For each scenario we generated 500 datasets. Estimation of the covariance matrix used the residuals from the pooled data estimator. We generated $\tilde{\epsilon}_i$ from $\text{Normal}(0, \Sigma_E)$ where Σ_E has elements σ_{ij} . We investigated seven cases. For Cases 1, 2 and 3, we used the exchangeable covariance matrix $\Sigma_E = (1 - \rho_a)I + \rho_a ee^T$ with $\rho_1 = 0.9$, $\rho_2 = 0.5$ and $\rho_3 = 0.1$, respectively. For Case 4, we used common variances $\sigma_{11} = \sigma_{22} = \sigma_{33} = 1$ with $\sigma_{12} = 0.9$, $\sigma_{13} = 0.5$ and $\sigma_{23} = 0.4$. For Case 5, we used common variances $\sigma_{11} = \sigma_{22} = \sigma_{33} = 1$ with AR(1) structure having coefficient -0.9 . Finally, for Cases 6 and 7 we allowed heteroscedasticity with $\sigma_{11} = 9$, $\sigma_{22} = 4$ and $\sigma_{33} = 1$, and with common correlation 0.9 and 0.1, respectively.

We also allowed for one function and an additive model with two functions. In the single function case, $m_1(x) = \sin\{2\pi(x - 0.5)\}$, where the three repeated measures X_i were generated from $\text{Normal}\{0.5e, \Sigma_X^b\}$ but truncated to the unit cube, with $\Sigma_X = \{(1 - \rho)I + \rho ee^T\}/4$ with $\rho = 0.8$ and $\rho = 0.1$. In the two-dimensional case, $m_1(x) = \sin\{2\pi(x - 0.5)\}$ and $m_2(x) = x - 0.5 + \sin\{2\pi(x - 0.5)\}$. The six-dimensional vector X_i was again generated as a truncated normal on the six-dimensional cube with common correlation 0.125.

In Table 1, we report the finite sample performance of the estimators, working with the true covariance matrix and an estimated covariance matrix. The results show that working with the true covariance matrix is slightly better than working with an estimated one but the differences are quite small. Table 2 summarizes the results of the one-dimensional models. As the theory suggests, the proposed estimator outperforms the conventional kernel method with pooled data in finite samples, sometimes dramatically. Table 3 shows similar efficiency gains for the

Table 1. *Finite sample performances of the estimators. Estimators for the function $m_1(x) = \sin\{2\pi(x - 0.5)\}$ in the single function case. The seven covariance types are explained in the text*

Covariance type		One-dimensional				Two-dimensional			
		LC		LL		LC		LL	
		T	E	T	E	T	E	T	E
1	ISB	29	30	9	9	28	30	9	9
	IV	34	36	39	39	37	37	40	40
2	ISB	28	30	8	8	31	28	7	7
	IV	94	94	125	126	97	98	129	131
3	ISB	27	28	7	7	28	31	7	7
	IV	119	127	160	162	130	131	178	179
4	ISB	30	29	10	11	31	31	10	10
	IV	45	49	54	55	49	50	57	57
5	ISB	29	29	7	7	31	32	7	8
	IV	34	37	37	38	35	35	38	38
6	ISB	28	30	8	9	31	31	8	9
	IV	59	58	73	74	60	63	76	76
7	ISB	32	31	9	9	26	33	9	9
	IV	256	258	371	376	264	274	395	398

ISB, integrated squared biases $\times 10^4$; IV, integrated variances $\times 10^4$; T, results with the true covariance matrix; E, results for an estimated covariance matrix; LC, local constant estimator; LL, local linear estimator.

Table 2. *Results of one-dimensional models. The table lists integrated squared biases $\times 10^4$, ISB, and integrated variances $\times 10^4$, IV, of our proposed estimator. The description of the seven covariance types is given in the text. The symbols Σ_X^A and Σ_X^B refer to two covariance structures for the distribution of the covariates*

Covariance type		Σ_X^A				Σ_X^B			
		LC		LL		LC		LL	
		WD	PD	WD	PD	WD	PD	WD	PD
1	ISB	30	28	9	5	25	24	7	7
	IV	36	141	39	185	27	107	26	133
2	ISB	30	29	8	7	24	24	7	7
	IV	94	132	126	182	76	106	90	133
3	ISB	28	28	7	7	24	23	8	7
	IV	127	128	162	163	108	109	132	133
4	ISB	29	30	11	8	25	24	7	6
	IV	49	132	38	168	38	108	40	130
5	ISB	29	27	7	8	27	26	8	8
	IV	37	116	38	168	31	104	27	136
6	ISB	30	30	9	6	24	25	7	6
	IV	58	587	74	818	44	465	54	595
7	ISB	31	30	9	12	24	23	8	9
	IV	258	528	376	717	224	462	302	615

LC, local constant estimator; LL, local linear estimator; WD, our proposed estimator that accounts for correlation; PD, pooled data or working independence estimator that ignores correlation.

Table 3. Results of two-dimensional models with functions $m_1(\cdot)$ and $m_2(\cdot)$. For details see Tables 1 and 2

Covariance type		m_1				m_2			
		LC		LL		LC		LL	
		WD	PD	WD	PD	WD	PD	WD	PD
1	ISB	30	32	9	10	27	22	8	7
	IV	37	142	40	194	45	154	42	182
2	ISB	28	27	7	7	26	24	11	10
	IV	98	136	131	189	101	145	124	174
3	ISB	31	23	7	7	23	22	8	7
	IV	131	131	179	183	132	133	169	173
4	ISB	31	27	10	10	29	25	9	8
	IV	50	140	57	192	56	146	57	180
5	ISB	32	29	8	8	26	27	7	8
	IV	35	115	38	173	45	131	42	163
6	ISB	31	24	9	10	27	28	8	8
	IV	63	595	76	849	69	599	78	764
7	ISB	33	25	9	9	28	28	7	11
	IV	274	540	398	822	270	535	376	758

two-dimensional models. Finally, our theory says that for estimating the function $m_1(x)$ that is common to the one function and two function cases, there should be little penalty from the increased dimension, which is seen to hold by comparing the results in Table 2 with those in Table 3.

5. EMPIRICAL EXAMPLE

To illustrate the repeated measures smooth backfitting algorithm, we use data from the OPEN study (Kipnis et al., 2003). The study is the first large biomarker study in nutritional epidemiology that attempts to understand how well people can report their actual dietary intakes. Background on nutritional epidemiology may be found in Willett (1990).

We use a dataset of 294 men and women measured at two visits who reported their short-term intake of protein Y_{ij} as measured by the biomarker urinary nitrogen. Here the protein biomarker data were log-transformed. To predict protein intake we used two variables, body mass index X_{i1j} and log-protein intake X_{i2j} as measured by a 24-hour recall instrument. Preliminary analysis using generalized least squares and a quadratic parametric fit suggested statistically significant nonlinearity in these two predictors. The residuals from an additive regression fit with the pooled data suggested an estimated covariance matrix with variances 0.065 and 0.074 and with a correlation of 0.506. The bandwidths were selected by the plug-in method with constant weight and are given as 3.02 for body mass index X_{i1j} and 0.374 for log-protein via the 24-hour recall X_{i2j} .

Figure 1 gives the results of the function fits, indicating the curvature that is found in the quadratic fits. We also computed the fits and associated 95% confidence intervals for the fit with pooled data, i.e. working independence. The theory suggests that our method, which accounts for correlation, should have smaller variance than the working independence estimator, which is seen in the fact that our method has confidence intervals that were approximately 20% shorter throughout the range of the predictors.

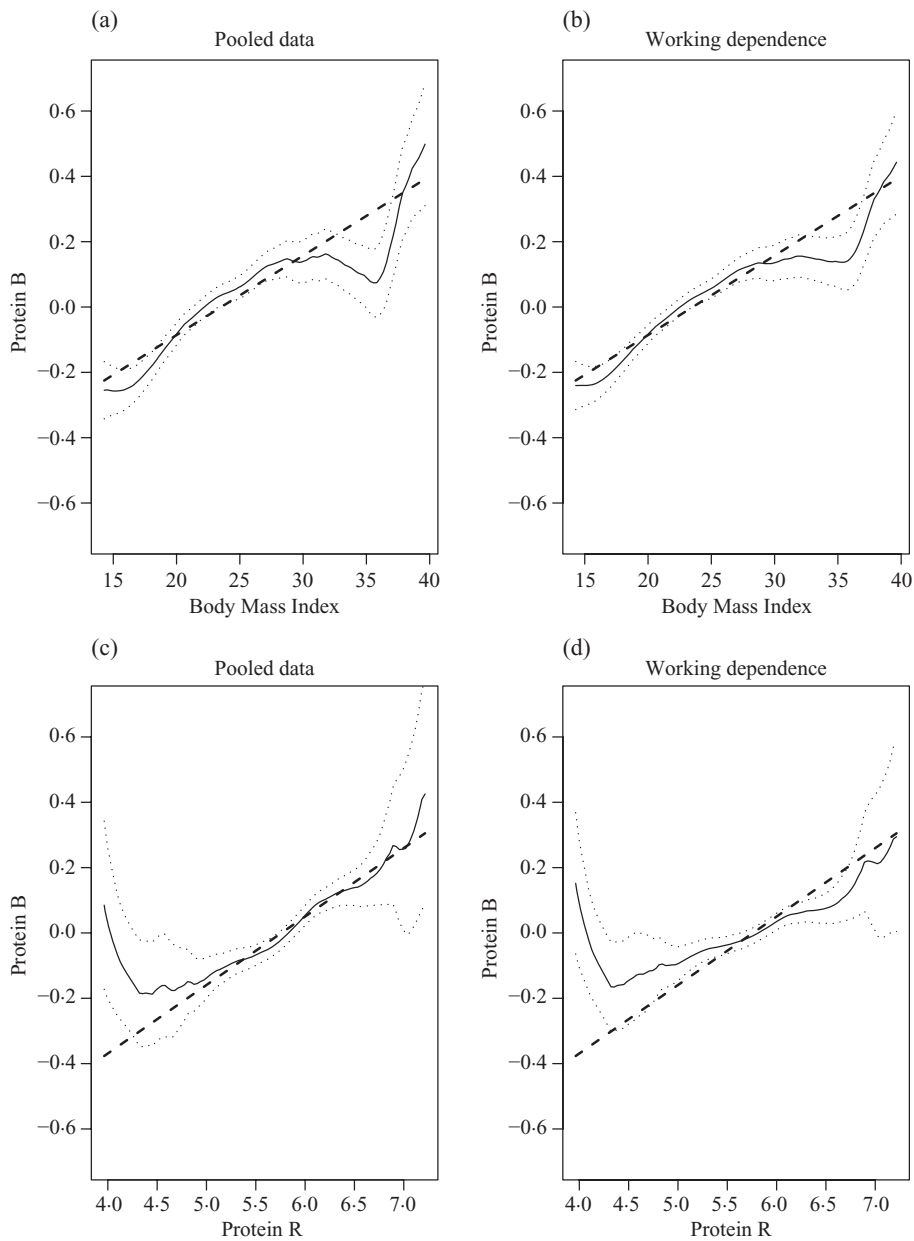


Fig. 1. OPEN study data: nonparametric and linear function fits for body mass index, (a) and (b), and the protein recall, (c) and (d), as predictors of the protein biomarker, 'Protein B', along with the 95% pointwise confidence interval for the nonparametric fit. The dashed line is the linear fit, the solid line the nonparametric fit and dotted lines represent the confidence intervals.

ACKNOWLEDGEMENT

The authors are grateful to the editor, associate editor and two referees for their invaluable comments and suggestions. Yu and Mammen's research was supported by the Deutsche Forschungsgemeinschaft. Carroll and Maity's research was supported by grants from the National Cancer Institute. Part of Carroll's work was supported by an award made by the King Abdullah University of Science and Technology.

APPENDIX

A.1. Derivation of the local constant estimator

Showing (5) is facilitated by the following device. We can rewrite (4) as a quadratic form,

$$S^\Sigma(m_1, \dots, m_D) = \int n^{-1} \sum_{i=1}^n \sum_{j=1}^J \left[\mathcal{Y}_{ij}^\Sigma - \sum_{k=1}^J a_{jk,\Sigma} \left\{ \hat{m}_0^\Sigma + \sum_{d=1}^D m_d(x_{dk}) \right\} \right]^2 \times \prod_{k=1}^J K_{h_1}(x_{1k}, X_{i1k}) \cdots K_{h_D}(x_{Dk}, X_{iDk}) dx. \tag{A1}$$

Suppose that (m_1, \dots, m_D) is the minimizer of (A1). Then we have the inequality

$$S^\Sigma(m_1, \dots, m_D) \leq S^\Sigma(m_1 + \eta_1, \dots, m_D + \eta_D), \tag{A2}$$

for any tuple of functions (η_1, \dots, η_D) for which $S^\Sigma(m_1 + \eta_1, \dots, m_D + \eta_D)$ exists. Define a linear functional $dS_{(m_1, \dots, m_D)}^\Sigma$ as

$$dS_{(m_1, \dots, m_D)}^\Sigma(\eta_1, \dots, \eta_D) = -2 \int n^{-1} \sum_{i=1}^n \sum_{j=1}^J \left[\mathcal{Y}_{ij}^\Sigma - \sum_{k=1}^J a_{jk,\Sigma} \left\{ \hat{m}_0^\Sigma + \sum_{d=1}^D m_d(x_{dk}) \right\} \right] \times \sum_{k=1}^J a_{jk,\Sigma} \left\{ \sum_{d=1}^D \eta_d(x_{dk}) \right\} \prod_{k=1}^J K_{h_1}(x_{1k}, X_{i1k}) \cdots K_{h_D}(x_{Dk}, X_{iDk}) dx,$$

for any measurable (η_1, \dots, η_D) for which this integral exists. The linear functional $dS_{(m_1, \dots, m_D)}^\Sigma$ is the differential of S at (m_1, \dots, m_D) . We obtain

$$S^\Sigma(m_1 + \eta_1, \dots, m_D + \eta_D) = S^\Sigma(m_1, \dots, m_D) + dS_{(m_1, \dots, m_D)}^\Sigma(\eta_1, \dots, \eta_D) + \int \sum_{j=1}^J \left\{ \sum_{k=1}^J a_{jk,\Sigma} \sum_{d=1}^D \eta_d(x_{dk}) \right\}^2 \hat{p}(x) dx, \tag{A3}$$

where $\hat{p}(x) = n^{-1} \sum_{i=1}^n \prod_{k=1}^J K_{h_1}(x_{1k}, X_{i1k}) \cdots K_{h_D}(x_{Dk}, X_{iDk})$. The last term in (A3) is nonnegative for any (η_1, \dots, η_D) . Thus it is clear from (A2) that the minimizer of S^Σ , (m_1, \dots, m_D) , satisfies $dS_{(m_1, \dots, m_D)}^\Sigma(\eta_1, \dots, \eta_D) = 0$ for any (η_1, \dots, η_D) . This is equivalent to (5).

A.2. The local linear estimator

For a local linear fit, consider the smoothed sum of squares given by

$$S(m_1, \dots, m_D, m^1, \dots, m^D) = \int n^{-1} \sum_{i=1}^n \sum_{j,k=1}^J b_{jk,\Sigma} \left\{ Y_{ij} - \hat{m}_0^\Sigma - \sum_{d=1}^D m_d(x_{dj}) - \sum_{d=1}^D \frac{X_{idj} - x_{dj}}{h_d} m^d(x_{dj}) \right\} \times \left\{ Y_{ik} - \hat{m}_0^\Sigma - \sum_{d=1}^D m_d(x_{dk}) - \sum_{d=1}^D \frac{X_{idk} - x_{dk}}{h_d} m^d(x_{dk}) \right\} \times \prod_{k=1}^J K_{h_1}(x_{1k}, X_{i1k}) \cdots K_{h_D}(x_{Dk}, X_{iDk}) dx. \tag{A4}$$

For $q = 1, \dots, D$, $r = 0, 1, 2$, define

$$\hat{u}_q^r(x) = \sum_{j=1}^J b_{jj,\Sigma} n^{-1} \sum_{i=1}^n \left(\frac{X_{iqj} - x}{h_q} \right)^r K_{h_q}(x, X_{iqj}),$$

and for $q, p = 1, \dots, D, r, s = 0, 1$, define

$$\hat{v}_q^{rs}(x, y) = \sum_{j=1}^J \sum_{k=1, \neq j}^J b_{jk, \Sigma} n^{-1} \sum_{i=1}^n \left(\frac{X_{iqj} - x}{h_q} \right)^r \left(\frac{X_{iqk} - y}{h_q} \right)^s K_{h_q}(x, X_{iqj}) K_{h_q}(y, X_{iqk}),$$

$$\hat{v}_{qp}^{rs}(x, y) = \sum_{j=1}^J \sum_{k=1}^J b_{jk, \Sigma} n^{-1} \sum_{i=1}^n \left(\frac{X_{iqj} - x}{h_q} \right)^r \left(\frac{X_{ipk} - y}{h_p} \right)^s K_{h_q}(x, X_{iqj}) K_{h_p}(y, X_{ipk}).$$

Then, in the same way as for the local constant case, we obtain a system of fitting integral equations:

$$\begin{aligned} \begin{Bmatrix} m_d(x_d) \\ m^d(x_d) \end{Bmatrix} &= \begin{Bmatrix} \tilde{m}_d^\Sigma(x_d) \\ \tilde{m}^{d, \Sigma}(x_d) \end{Bmatrix} - \hat{M}_d^{-1}(x_d) \int \hat{V}_d(x_d, t) \begin{Bmatrix} m_d(t) \\ m^d(t) \end{Bmatrix} dt \\ &\quad - \sum_{s=1, \neq d}^D \hat{M}_d^{-1}(x_d) \int \hat{V}_{ds}(x_d, t) \begin{Bmatrix} m_s(t) \\ m^s(t) \end{Bmatrix} dt, \end{aligned} \tag{A5}$$

where

$$\begin{aligned} \hat{M}_d(x_d) &= \begin{Bmatrix} \hat{u}_d^0(x_d) & \hat{u}_d^1(x_d) \\ \hat{u}_d^1(x_d) & \hat{u}_d^2(x_d) \end{Bmatrix}, \\ \begin{Bmatrix} \tilde{m}_d^\Sigma(x_d) \\ \tilde{m}^{d, \Sigma}(x_d) \end{Bmatrix} &= \hat{M}_d^{-1}(x_d) \begin{Bmatrix} n^{-1} \sum_{i=1}^n \sum_{j,k=1}^J b_{jk, \Sigma} (Y_{ik} - \hat{m}_0^\Sigma) K_{h_d}(x_d, X_{idj}) \\ n^{-1} \sum_{i=1}^n \sum_{j,k=1}^J b_{jk, \Sigma} (Y_{ik} - \hat{m}_0^\Sigma) \frac{X_{idj} - x_d}{h_d} K_{h_d}(x_d, X_{idj}) \end{Bmatrix}, \\ \hat{V}_d(x_d, t) &= \begin{Bmatrix} \hat{v}_d^{0,0}(x_d, t) & \hat{v}_d^{0,1}(x_d, t) \\ \hat{v}_d^{1,0}(x_d, t) & \hat{v}_d^{1,1}(x_d, t) \end{Bmatrix}, \quad \hat{V}_{ds}(x_d, t) = \begin{Bmatrix} \hat{v}_{ds}^{0,0}(x_d, t) & \hat{v}_{ds}^{0,1}(x_d, t) \\ \hat{v}_{ds}^{1,0}(x_d, t) & \hat{v}_{ds}^{1,1}(x_d, t) \end{Bmatrix}, \end{aligned}$$

for $d, s = 1, \dots, D$. We impose the new identification condition

$$\int \sum_{k=1}^J \sum_{j=1}^J b_{jk, \Sigma} \left\{ m_d(x) \hat{p}_j^d(x) + m^d(x) n^{-1} \sum_{i=1}^n \left(\frac{X_{idj} - x}{h_d} \right) K_{h_d}(x, X_{idj}) \right\} dx = 0. \tag{A6}$$

It is easy to check that the solution of equation (A5) satisfies the identification condition (A6).

A.3. The existence of the estimator

Let $m(x_1, \dots, x_D) = \{m_1(x_1), \dots, m_D(x_D)\}^T$, $\tilde{m}^\Sigma(x_1, \dots, x_D) = \{\tilde{m}_1^\Sigma(x_1), \dots, \tilde{m}_D^\Sigma(x_D)\}^T$. Then equation (5) can be written as

$$m(x_1, \dots, x_D) = \tilde{m}^\Sigma(x_1, \dots, x_D) + (\hat{G}^\Sigma m)(x_1, \dots, x_D),$$

where the d th element of $(\hat{G}^\Sigma m)(x_1, \dots, x_D)$ is given by

$$\int m_d(t) \frac{\sum_{k \neq j, =1}^J \sum_{j=1}^J b_{jk, \Sigma} \hat{p}_{jk}^{dd}(x_d, t)}{\sum_{j=1}^J b_{jj, \Sigma} \hat{p}_j^d(x_d)} dt + \sum_{s=1, \neq d}^D \int m_s(t) \frac{\sum_{j=1}^J \sum_{k=1}^J b_{jk, \Sigma} \hat{p}_{jk}^{ds}(x_d, t)}{\sum_{j=1}^J b_{jj, \Sigma} \hat{p}_j^d(x_d)} dt. \tag{A7}$$

Since the operator $I - \hat{G}^\Sigma$ is invertible under the constraint (6), the existence of the estimator is guaranteed and we can define an estimator $\hat{m}^\Sigma(x_1, \dots, x_D) = \{\hat{m}_1^\Sigma(x_1), \dots, \hat{m}_D^\Sigma(x_D)\}^T$ as

$$\hat{m}^\Sigma(x_1, \dots, x_D) = (I - \hat{G}^\Sigma)^{-1} \tilde{m}^\Sigma(x_1, \dots, x_D).$$

We now discuss the invertibility of the operator. We have that $S^\Sigma(m_1, \dots, m_D)$ defined in (A1) is an empirical version of $\sum_{j=1}^J E[\mathcal{Y}_j^\Sigma - \sum_{k=1}^J a_{jk, \Sigma} \{\sum_{d=1}^D m_d(X_{dj})\}]^2$ with centred responses. We define

$(\mathcal{G}^\Sigma m)(\cdot)$ as $(\hat{\mathcal{G}}^\Sigma m)(\cdot)$ in (A7) by replacing each density estimator with the corresponding true density. It is easy to check that

$$\begin{aligned} & \sum_{j=1}^J E \left[\sum_{k=1}^J a_{jk,\Sigma} \left\{ \sum_{d=1}^D m_d(X_{dj}) \right\} \right]^2 \\ &= \int m(x_1, \dots, x_D)^\top (I - \mathcal{G}^\Sigma) m(x_1, \dots, x_D) \sum_{j=1}^J \sum_{\ell=1}^J a_{j\ell,\Sigma}^2 p_{\ell\ell}(x_1, \dots, x_D) dx. \end{aligned} \tag{A8}$$

Let $m^\lambda = (m_1^\lambda, \dots, m_D^\lambda)^\top$ be an eigenfunction corresponding to an eigenvalue λ of the operator \mathcal{G} . Then, from (A8), we have

$$\begin{aligned} & \sum_{j=1}^J E \left[\sum_{k=1}^J a_{jk,\Sigma} \left\{ \sum_{d=1}^D m_d^\lambda(X_{dk}) \right\} \right]^2 \\ &= (1 - \lambda) \int m^\lambda(x_1, \dots, x_D)^\top m^\lambda(x_1, \dots, x_D) \sum_{j=1}^J \sum_{\ell=1}^J a_{j\ell,\Sigma}^2 p_{\ell\ell}(x_1, \dots, x_D) dx. \end{aligned} \tag{A9}$$

Since Σ is positive definite, the left side of equation (A9) cannot be zero unless $m_1^\lambda(X_{1k}) + \dots + m_D^\lambda(X_{Dk}) = 0$ with probability 1 for all $k = 1, \dots, J$. This implies that all eigenvalues of \mathcal{G}^Σ should be less than 1 and thus $I - \mathcal{G}^\Sigma$ is invertible. Assumption A1 below is a sufficient condition for the operator \mathcal{G}^Σ to be compact; see Linton & Mammen (2005) for details. This holds uniformly over Σ positive definite and $c \leq |\Sigma| \leq C$ for given constants $0 < c < C$, where $|\Sigma| = \sup_{a \in \mathbb{R}^J, |a|_2=1} a^\top \Sigma a$. This argument remains valid for the operator $I - \hat{\mathcal{G}}^\Sigma$ that uses estimated densities. Similar arguments can be applied to the local linear case.

A.4. The algorithm in § 2.3 converges

Let $\mathcal{F} = \{f = (f_1, \dots, f_J)^\top : f_j \text{ are real valued functions on } [0, 1]^D\}$ and $\mathcal{H} = \{f = (f, \dots, f)^\top : f(x) = \sum_{d=1}^D f_d(x_d) \text{ is additive functions on } [0, 1]^D\}$. Here \mathcal{H} is a subspace of \mathcal{F} and has one common function, f . We equip \mathcal{F} with a Hilbert seminorm, $|f|_\Sigma = \int f^\top(x) \Sigma^{-1} f(x) \hat{p}(x) dx$. Let also $\mathcal{H}^0 = \{f \in \mathcal{H} : \langle e, f \rangle_\Sigma = 0\}$, where $e = (1, \dots, 1)^\top \in \mathcal{F}$ and $\langle \cdot, \cdot \rangle_\Sigma$ is the inner product inducing the norm $|\cdot|_\Sigma$. Define, for $d = 1, \dots, D$, $\mathcal{H}_d = \{f \in \mathcal{H}^0 : f(x) = f_d(x_d) \text{ for some real } f_d \text{ on } [0, 1]\}$. Then it is clear that $\mathcal{H}^0 = \mathcal{H}_1 + \dots + \mathcal{H}_D$. Let $\Pi_d(\cdot)$ be the projection operator onto \mathcal{H}_d and $P_d(\cdot|\mathcal{H}_a)$ be the restriction of Π_d to \mathcal{H}_a . It is easy to see that $P_d(\cdot|\mathcal{H}_a) = (I + \hat{\mathcal{G}}_{da}^\Sigma)^{-1} \hat{\mathcal{G}}_{da}^\Sigma$. Hence this algorithm is an alternating projection algorithm and thus it converges; see, e.g. Appendix 4 in Bickel et al. (1993) for details.

We can also define the same kind of algorithm for local linear fitting by using, in Step 2, the operators

$$\begin{aligned} \{\hat{\mathcal{G}}_{da}^\Sigma(f, g)^\top\}(x) &= \hat{M}_d^{-1}(x) \int \hat{V}_d(x, t) \{f(t), g(t)\}^\top dt, \\ \{\hat{\mathcal{G}}_{da}^\Sigma(f, g)^\top\}(x) &= \hat{M}_d^{-1}(x) \int \hat{V}_{da}(x, t) \{f(t), g(t)\}^\top dt. \end{aligned}$$

The discussion of local linear fitting follows the same lines as the local constant case.

A.5. Assumptions

Here we collect the assumptions for the theoretical results.

Assumption A1. For $j, k = 1, \dots, J$ and $d, s = 1, \dots, D$, p_{jk}^{ds} is bounded away from zero and infinity on its support, $[0, 1]^2$, and has continuous partial derivatives.

Assumption A2. $E|Y_{1j}|^{r_0} < \infty$ for $j = 1, \dots, J$ and some $r_0 > 5/2$.

Assumption A3. The true component functions $m_{d,\text{true}}(\cdot)$ are twice continuously differentiable.

Assumption A4. The base kernel function K^0 is a symmetric density function with compact support, $[-1, 1]$, say, and is Lipschitz continuous.

Assumption A5. $n^{1/5}h_d$ converge to constants $c_d > 0$ for $d = 1, \dots, D$ as n goes to infinity.

Here, we present only a sketch proof of Theorem 1. Details and other proofs are available in a long version of the paper given at <http://mammen.vwl.uni-mannheim.de>.

A.6. *Sketch of proof for Theorem 1*

For a sequence $\delta_n \rightarrow 0$ let $\mathcal{S}(\Sigma) = \{V : \text{positive definite and } |V - \Sigma| \leq \delta_n\}$. Here $\hat{\Sigma} \in \mathcal{S}(\Sigma)$ with probability tending to one if δ_n converges to zero slowly enough. We will now show an expansion of \hat{m}^V that holds uniformly for $V \in \mathcal{S}(\Sigma)$. We decompose \hat{m}^V into a mean part $\hat{m}^{V,M}$, and an error part $\hat{m}^{V,E}$, where for $s = M, E$ we define $\hat{m}^{V,s} = (I - \hat{\mathcal{G}}^V)^{-1} \tilde{m}^{V,s}$ for $\tilde{m}^{V,s} = (\tilde{m}_1^{V,s}, \dots, \tilde{m}_D^{V,s})^\top$ with elements

$$\tilde{m}_d^{V,M}(x_d) = n^{-1} \sum_{i=1}^n \sum_{j,k=1}^J b_{jk,V} \sum_{s=1}^D m_{s,\text{true}}(X_{isk}) K_{h_d}(x_d, X_{idj}) \bigg/ \sum_{j=1}^J b_{jj,V} \hat{p}_j^d(x_d),$$

$$\tilde{m}_d^{V,E}(x_d) = n^{-1} \sum_{i=1}^n \sum_{j,k=1}^J b_{jk,V} \epsilon_{ik} K_{h_d}(x_d, X_{idj}) \bigg/ \sum_{j=1}^J b_{jj,V} \hat{p}_j^d(x_d),$$

for $d = 1, \dots, D$. Then $\hat{m}^V = \hat{m}^{V,M} + \hat{m}^{V,E}$ holds since the operator $(I - \hat{\mathcal{G}}^V)^{-1}$ is linear and $\tilde{m}_d^V(x_d) = \tilde{m}_d^{V,M}(x_d) + \tilde{m}_d^{V,E}(x_d)$. Let $I_d = [2h_d, 1 - 2h_d]$, $m_{\text{true}} = (m_{1,\text{true}}, \dots, m_{D,\text{true}})^\top$ and $|m|_* = \max_{1 \leq d \leq D} \sup_{x_d \in I_d} |m_d(x_d)|$. We will argue that

$$\sup_{V \in \mathcal{S}(\Sigma)} |(I - \hat{\mathcal{G}}^V)^{-1} \tilde{m}^{V,M} - m_{\text{true}} - (\beta_1^\Sigma, \dots, \beta_D^\Sigma)^\top|_* = o_p(n^{-2/5}), \tag{A10}$$

$$\sup_{V \in \mathcal{S}(\Sigma)} |(I - \hat{\mathcal{G}}^V)^{-1} \tilde{m}^{V,E} - \tilde{m}^{\Sigma,E}|_* = o_p(n^{-2/5}), \tag{A11}$$

and that

$$n^{2/5} \{ \tilde{m}_1^{\Sigma,E}(x_1), \dots, \tilde{m}_D^{\Sigma,E}(x_D) \}^\top \rightarrow N[0, \text{diag}\{v_1^\Sigma(x_1), \dots, v_D^\Sigma(x_D)\}], \tag{A12}$$

where $v_d^\Sigma(x_d) = c_d^{-1} \mu_0 \{ (K^0)^2 \} \text{trace}\{ \Sigma^{-1} \Sigma_{\text{true}} \Sigma^{-1} P^d(x_d) \} \{ \sum_{j=1}^J b_{jj}^\Sigma p_j^d(x_d) \}^{-2}$. These claims immediately imply the statement of the theorem.

We now outline the proofs of (A10)–(A12). First, the convergence in (A12) follows directly from standard kernel smoothing theory. For the proof of (A10) and (A11), we will use the convergence of the operator

$$\sup_{V \in \mathcal{S}(\Sigma), |m|_* \leq 1} |(I - \hat{\mathcal{G}}^V)^{-1} - (I - \mathcal{G}^V)^{-1}| m|_* = o_p(n^{-3/10+\xi}) \tag{A13}$$

for $\xi > 0$. For a proof of claim (A13), see Appendix B in Linton & Mammen (2005) for a detailed treatment of a similar expansion. The basic argument is the continuity of the map $V \rightarrow (I - \mathcal{G}^V)^{-1}$.

For a proof of (A11), one uses the decomposition $(I - \hat{\mathcal{G}}^V)^{-1} \tilde{m}^{V,E} - \tilde{m}^{\Sigma,E} = \{(I - \hat{\mathcal{G}}^V)^{-1} - (I - \mathcal{G}^V)^{-1}\} \tilde{m}^{V,E} + (I - \mathcal{G}^V)^{-1} (\tilde{m}^{V,E} - \tilde{m}^{\Sigma,E}) + (I - \mathcal{G}^V)^{-1} \mathcal{G}^V \tilde{m}^{\Sigma,E}$. The first two terms in the decomposition are asymptotically negligible because of (A13) and $\sup_{V \in \mathcal{S}(\Sigma)} |\tilde{m}^{V,E} - \tilde{m}^{\Sigma,E}|_* = o_p(n^{-2/5})$. For the third term, we have $\sup_{V \in \mathcal{S}(\Sigma)} |\mathcal{G}^V \tilde{m}^{\Sigma,E}|_* = o_p(n^{-2/5})$ because the integration in the operator \mathcal{G}^V changes $\tilde{m}^{\Sigma,E}$ from a local average to a global average.

The proof of (A10) is based on lengthy bias calculations based on the decomposition

$$\begin{aligned} \hat{m}^{V,M} - m_{\text{true}} &= (I - \mathcal{G}^V)^{-1} \{ \tilde{m}^{V,M} - (I - \hat{\mathcal{G}}^V) m_{\text{true}} \} \\ &\quad + \{ (I - \hat{\mathcal{G}}^V)^{-1} - (I - \mathcal{G}^V)^{-1} \} \{ \tilde{m}^{V,M} - (I - \hat{\mathcal{G}}^V) m_{\text{true}} \}. \end{aligned} \tag{A14}$$

Using similar calculations as in Mammen et al. (1999), we have

$$\begin{aligned} \{\tilde{m}^{V,M} - (I - \hat{G}^V)m_{\text{true}}\}(x) &= (I - \mathcal{G}^V) \left[\frac{1}{2} \mu_2(K^0) \begin{Bmatrix} h_1^2 m'_{1,\text{true}}(x_1) \\ \vdots \\ h_D^2 m'_{D,\text{true}}(x_D) \end{Bmatrix} \right. \\ &\quad \left. + \begin{Bmatrix} m'_{1,\text{true}}(x_1) \int K_{h_1}(x_1, u)(u - x_1) du \\ \vdots \\ m'_{D,\text{true}}(x_D) \int K_{h_D}(x_D, u)(u - x_D) du \end{Bmatrix} \right] (x) \\ &\quad + \mu_2(K^0) (\gamma_1^V \cdots \gamma_D^V)^\top(x) + R_n^V(x), \end{aligned}$$

where $\sup_{V \in \mathcal{S}(\Sigma), x_d \in [0,1]} |(R_n^V)_d(x_d)| = o_p(n^{-2/5})$. Here, for $d = 1, \dots, D$,

$$\begin{aligned} \gamma_d^V(x_d) &= h_d^2 \sum_{j=1}^J b_{jj,V} m'_{d,\text{true}}(x_d) p_j^{d'}(x_d) + \left\{ \sum_{j=1}^J b_{jj,V} p_j^d(x_d) \right\}^{-1} \\ &\quad \times \left\{ h_d^2 \int \sum_{j=1}^J \sum_{k \neq j, k=1}^J b_{jk,V} m'_{d,\text{true}}(t) \frac{dp_{jk}^{dd}(x_d, t)}{dt} dt \right. \\ &\quad \left. + \sum_{s=1, \neq d}^D h_s^2 \int \sum_{j=1}^J \sum_{k=1}^J b_{jk,V} m'_{d,\text{true}}(t) \frac{dp_{jk}^{ds}(x_d, t)}{dt} dt \right\}, \end{aligned} \tag{A15}$$

with $V^{-1} = (b_{jk,V})_{j,k=1,\dots,J}$. Claim (A10) follows from (A13) and (A14).

REFERENCES

BICKEL, P., KLAASSEN, A., RITOV, Y. & WELLNER, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: Johns Hopkins University Press.

CHEN, K. & JIN, Z. (2005). Local polynomial regression analysis of clustered data. *Biometrika* **92**, 59–74.

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

HOOVER, D. R., RICE, J. A., WU, C. O. & YANG, Y. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–22.

HUGGINS, R. (2006). Understanding nonparametric estimation for clustered data. *Biometrika* **93**, 486–89.

KIPNIS, V., SUBAR, A. F., MIDTHUNE, D., FREDMAN, L. S., BALLARD-BARBASH, R., TROIANO, R., BINGHAM, S., SCHOELLER, D. A., SCHATZKIN, A. & CARROLL, R. J. (2003). The structure of dietary measurement error: results of the OPEN biomarker study. *Am. J. Epidemiol.* **158**, 14–21.

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.

LIN, X. & CARROLL, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *J. Am. Statist. Assoc.* **96**, 1045–56.

LIN, X. & CARROLL, R. J. (2006). Semiparametric estimation in general repeated measures problems. *J. R. Statist. Soc. B* **68**, 68–88.

LIN, D. Y. & YING, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *J. Am. Statist. Assoc.* **96**, 103–26.

LIN, X., WANG, N., WELSH, A. H. & CARROLL, R. J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for longitudinal/clustered data. *Biometrika* **91**, 177–94.

LINTON, O. & MAMMEN, E. (2005). Estimating semiparametric ARCH(∞) models by kernel smoothing methods. *Econometrica* **73**, 771–836.

LINTON, O. & MAMMEN, E. (2008). Nonparametric transformation to white noise. *J. Economet.* **142**, 241–64.

LINTON, O. B., MAMMEN, E., LIN, X. & CARROLL, R. J. (2004). Correlation and marginal longitudinal kernel nonparametric regression. In *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*, Ed. D. Y. Lin and P. J. Heagerty, pp. 23–33. New York: Springer.

NIELSEN, J. & LINTON, O. (1998). An optimization interpretation of integration and back-fitting estimators for separable nonparametric models. *J. R. Statist. Soc. B* **60**, 217–22.

MAMMEN, E., LINTON, O. & NIELSEN, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27**, 1443–90.

- MAMMEN, E. & NIELSEN, J. P. (2003). Generalised structured models. *Biometrika* **90**, 551–66.
- MAMMEN, E. & PARK, B. U. (2005). Bandwidth selection for smooth backfitting in additive models. *Ann. Statist.* **33**, 1260–94.
- MAMMEN, E. & PARK, B. U. (2006). A simple smooth backfitting method for additive models. *Ann. Statist.* **34**, 2252–71.
- NIELSEN, J. & SPERLICH, S. (2005). Smooth backfitting in practice. *J. R. Statist. Soc. B* **67**, 43–61.
- RUCKSTUHL, A., WELSH, A. H. & CARROLL, R. J. (2000). Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statist. Sinica* **10**, 51–71.
- WANG, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* **90**, 43–52.
- WILLETT, W. C. (1990). *Nutritional Epidemiology*. New York: Oxford University Press.
- YU, K., MAMMEN, E. & PARK, B. (2008). Smooth backfitting in generalized additive models. *Ann. Statist.* **36**, 228–60.

[Received October 2007. Revised October 2008]