Histospline Method in Nonparametric Regression Models with Application to Clustered/Longitudinal Data

Raymond J. Carroll[*]   Peter Hall[†]

Tatiyana V. Apanasovich[‡]   Xihong Lin[**]

[*]Texas A&M University, rcarroll@tamu.edu

[†]Australian National University, halpstat@fac.anu.edu.au

[‡]Texas A&M University, tanya@stat.tamu.edu

[**]University of Michigan, xlin@umich.edu

Histospline Method in Nonparametric Regression Models with Application to Clustered/Longitudinal Data

Raymond J. Carroll, Peter Hall, Tatiyana V. Apanasovich, and Xihong Lin

Abstract

Kernel and smoothing methods for nonparametric function and curve estimation have been particularly successful in "standard" settings, where function values are observed subject to independent errors. However, when aspects of the function are known parametrically, or where the sampling scheme has significant structure, it can be quite difficult to adapt standard methods in such a way that they retain good statistical performance and continue to enjoy easy computability and good numerical properties. In particular, when using local linear modeling it is often awkward to both respect the sampling scheme and produce an estimator with good variance properties, without resorting to iterative methods: a good case in point is longitudinal and clustered data. In this paper we suggest a simple approach to overcoming these problems. Using a histospline technique we convert a problem in the continuum to one that is governed by only a finite number of parameters, and which is often explicitly solvable. The simple expedient of running a local linear smoother through the histospline produces a function estimator which achieves optimal nonparametric properties, and the "raw" histospline-based estimator of the semiparametric component itself attains optimal semiparametric performance. The function estimator can be used in its own right or as the starting value for an iterative scheme based on a different approach to inference.

Histospline Method in Nonparametric Regression Models with Application to Clustered/Longitudinal Data

Raymond J. Carroll [1], Peter Hall [2], Tatiyana V. Apanasovich [3], and Xihong Lin [4]

Abstract

Kernel and smoothing methods for nonparametric function and curve estimation have been particularly successful in "standard" settings, where function values are observed subject to independent errors. However, when aspects of the function are known parametrically, or where the sampling scheme has significant structure, it can be quite difficult to adapt standard methods in such a way that they retain good statistical performance and continue to enjoy easy computability and good numerical properties. In particular, when using local linear modeling it is often awkward to both respect the sampling scheme and produce an estimator with good variance properties, without resorting to iterative methods: a good case in point is longitudinal and clustered data. In this paper we suggest a simple approach to overcoming these problems. Using a histospline technique we convert a problem in the continuum to one that is governed by only a finite number of parameters, and which is often explicitly solvable. The simple expedient of running a local linear smoother through the histospline produces a function estimator which achieves optimal nonparametric properties, and the "raw" histospline-based estimator of the semiparametric component itself attains optimal semiparametric performance. The function estimator can be used in its own right or as the starting value for an iterative scheme based on a different approach to inference.

Keywords: Bandwidth, Binwidth, Clustered data, Efficient estimation, Financial data, Histogram, Interpolation, Kernel methods, Least squares, Local linear, Local polynomial, Longitudinal data, Optimality, Smoothing.

Short Title: Histospline Method in Nonparametric Regression Models

[1](E-mail: carroll@stat.tamu.edu), Department of Statistics, TAMU 3143, Texas A&M University, College Station TX 77843–3143, USA.

[2](E-mail: halpstat@fac.anu.edu.au), Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia.

[3](E-mail: tanya@stat.tamu.edu), Department of Statistics, TAMU 3143, Texas A&M University, College Station TX 77843–3143, USA.

[4](E-mail: xlin@umich.edu), Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.

1 INTRODUCTION

In conventional nonparametric regression problems, data Y_i are observed and have respective conditional means $\theta(X_i)$ and observation errors ϵ_i, where the latter are independent. In this context there is a wealth of literature on ways of estimating the function θ under smoothness assumptions alone using kernel and spline methods. It includes, for example, for kernel methods, Fan and Gijbels (1995) and Wand and Jones (1995); for smoothing splines, Green and Silverman (1994), Simonoff (1996) and Wahba (1991); for regression splines, Stone (1991) and Huang (2003). Local polynomial estimators, with very good statistical and numerical performance (see e.g. Fan, 1993), can be written down explicitly and are readily computed.

However, under even small changes to the model these attractive features of local polynomial estimators can quickly disappear. A particularly important example, having significant practical implications and where a great deal of work has been done, arises in clustered and longitudinal data applications. There, one observes data (X_{ij}, Y_{ij}) generated by the model

$$Y_{ij} = \theta(X_{ij}) + \epsilon_{ij}, \quad 1 \le i \le n, \quad 1 \le j \le m, \tag{1.1}$$

where, conditional on the m-vectors $X_i = (X_{i1}, \ldots, X_{im})^{\mathrm{T}}$ and defining $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{im})^{\mathrm{T}}$, the m-vectors $\Sigma^{-1/2}\epsilon_i$, for $1 \le i \le n$, are independent and identically distributed with a known m-variate distribution having zero mean, for example the Normal $\mathrm{N}(0, I_m)$ distribution, and $\Sigma = \Sigma(\alpha)$ is an $m \times m$ covariance matrix. The Normal assumption is usually a surrogate, doing no more than motivating an estimator that is consistent quite generally and efficient under Normality. Typically, m would be small and n relatively large.

Kernel methods have been proposed for the model (1.1), see for example, Zeger and Diggle (1994), Wild and Yee (1996), Hoover et al. (1998), Wu et al. (1998), Lin and Carroll (2000), Ruckstuhl et al. (2000), Wang (2003) and Linton et al. (2003). However, with one exception, the local linear modeling literature contains no explicitly-representable, well-performing nonparametric methods for estimating θ under the model at (1.1) and its many generalizations. Iterative techniques, derived from local linear ideas but implemented through backfitting, can be employed, but for effective implementation they require a good first approximation to θ. The iterative kernel method of Wang (2003), unlike the other kernel methods referenced, is efficient for the model at (1.1) by effectively accounting for the correlation, and

1

has a closed–form solution (Lin, et al., 2003). However the solution requires the inversion of an $nm \times nm$ matrix and the order of computation is $O\{(nm)^3\}$, and thus is not typically available in practice.

Smoothing spline methods have also been proposed for the model (1.1), for example see Brumback and Rice (1998), Wang (1998), Zhang, et al. (1998), Verbyla, et al. (1999), and Lin, et al., (2003). An attractive feature of the smoothing spline method is that it can be fit through a mixed effects model. Lin, et al., (2003) further show that the smoothing spline estimator is asymptotically equivalent to the iterative kernel method of Wang (2003) and is efficient by effectively accounting for the correlation. However, calculations of the smoothing spline estimator are computationally intensive and require inverting an $nm \times nm$ matrix with the order of computation being $O\{(nm)^3\}$, and the classical efficient Reinsch algorithm is not applicable.

Another important "perturbation" of the standard nonparametric regression model, one which has been addressed by (for example) Zeger and Diggle (1994), Zhang, et al. (1998), Lin and Carroll (2001a,b), Lin and Ying (2001) and Wang et al. (2003), is a partially linear model:

$$Y_{ij} = \beta^{\mathrm{T}} Z_{ij} + \theta(X_{ij}) + \epsilon_{ij}, \quad 1 \le i \le n, \quad 1 \le j \le m, \tag{1.2}$$

where $\beta = (\beta_1, \ldots, \beta_m)^{\mathrm{T}}$ and, conditional on X_i and $Z_{ij} = (Z_{ij1}, \ldots, Z_{ijm})^{\mathrm{T}}$, the vectors $\Sigma^{-1/2} \epsilon_i$ are independent and identically distributed with zero mean and identity covariance. Here, $\Sigma = \Sigma(\gamma)$ denotes an $m \times m$ covariance matrix.

A third example, arising in finance (Hafner, 1998; Carroll et al., 2002), is

$$Y_i = \sum_{j=1}^{m} \beta^{j-1} \theta(X_{ij}) + \epsilon_i, \quad 1 \le i \le n, \tag{1.3}$$

where β is a scalar and, conditional on the X_i's, the ϵ_i's are independent and identically distributed with zero mean and variance γ. Of course, β is identifiable only if $\theta(\bullet) \not\equiv 0$. In all three examples we wish to estimate θ under smoothness assumptions alone, and to estimate too the parameter α, representing the concatenation of β and γ.

Our aim in this paper is to suggest a straightforward yet general methodology which gives readily computed solutions to these problems and to a wide range of others. The solutions

2

have attractive statistical and numerical properties. Our approach is based on fitting a histospline to θ using regularly spaced histogram bins. One can take the histospline to be the final estimator of θ, or smooth it a little using a version of the average shifted histogram. Alternatively, one can pass a local polynomial smoother through the histogram estimator, obtaining a function estimator that is optimal in the context of Fan (1993). If one is set on computing an estimator that is based on iteration, then our approach provides an explicit, easily computed starting value for the algorithm.

The attraction of the histospline approach is that it explicitly converts a problem which was originally in the continuum, to one that is determined by only a finite number of parameters which enter the estimating equations in an elementary way. Indeed, the estimating equations take such a simple form that in the cases of the models at (1.1) and (1.2), least-squares estimators of all the unknowns are given explicitly in terms of the data, and Computation requires nothing more complex than matrix inversion. In the case of (1.3) the estimator of θ is obtained in the same way, for each fixed value of β; and then β is estimated by an elementary, univariate search routine. This offers substantially greater ease of computation than existing methods.

In work that will be reported elsewhere, we have shown that the estimators of discrete parameters β and γ, arising in this way, achieve semiparametric efficiency bounds and are root-n consistent. This level of performance is attained using the same smoothing parameters that given optimal estimation of nonparametric components, in particular of the functions θ in (1.1)–(1.3). Indeed, optimal semiparametric estimation is achieved by the basic histogram estimator, without the need for either undersmoothing or two-step methods.

General methodology is described in Section 2, and specialized there to the model at (1.1). Numerical properties are described in Section 3. Theory is discussed in Section 4, with concluding remarks given in Section 5. Theoretical development is given in an Appendix. Formulae for computing estimators in the cases of models (1.2) and (1.3) are also given in the Appendix.

3

2 METHODOLOGY

2.1 Estimators Under General Models

To introduce our technique, consider a general model which encompasses those at (1.1)–(1.3). Assume the data for an individual i are expressed in several vectors,

$$X_i^{\text{vec}} = (X_{i1}, \ldots, X_{im_x})^{\text{T}}, \quad Y_i^{\text{vec}} = (Y_{i1}, \ldots, Y_{im_y})^{\text{T}}, \quad Z_i^{\text{vec}} = (Z_{i1}, \ldots, Z_{im_z})^{\text{T}},$$

and the negative loglikelihood function for an individual i is, say,

$$\mathcal{L}\{Y_i^{\text{vec}}, Z_i^{\text{vec}}, \theta(X_{i1}), \ldots, \theta(X_{im_x}), \alpha\}, \tag{2.1}$$

with θ denoting a function that occurs multiple times but with different arguments, and α a vector of parameters. For simplicity we take θ to be a real-valued function of a real variable, but versions of the methodology we shall suggest are readily developed in multivariate settings. In the case of the model at (1.1), $m_x = m_y = m$ and Z_i plays no role. For (1.2), $m_x = m_y = m_z = m$ and $\alpha = (\beta^{\text{T}}, \gamma^{\text{T}})^{\text{T}}$, and for (1.3), $m_x = m$, $m_y = 1$, β is a scalar, Z_i is degenerate and $\alpha = (\beta, \gamma)^{\text{T}}$.

Our histospline estimator is constructed as follows. Assume we wish to calculate an estimator of "order" $2(p+1) \geq 2$, suitable for estimating functions with $2(p+1)$ bounded derivatives. The case of even order is more common in practice, although odd-order cases may be treated similarly. Let $\overline{\theta}$ denote the polynomial interpolant of degree $2p$ or $2p+1$ on bins of width h, fitted to a histogram of height c_ℓ on the bin $\mathcal{B}_\ell = (x^0 + (\ell-1)h, x^0 + \ell h]$, where x^0 is arbitrary. Then $\overline{\theta}$ is completely determined by the column vector $c = (c_\ell)$, which is generally of finite length since the data X_{ij} take values only on a compact interval. Replacing θ at (2.1) by $\overline{\theta}$ we obtain a likelihood under a finite-parameter model, in which the number of parameters is determined by h. Parameter estimators may be obtained in the conventional way, by solving the estimating equations, giving estimators $\widehat{c}_\ell$ of the respective c_ℓ's.

The most important special case is $p = 0$, and there the vector of estimators is often very easy to compute, with all or most of its components being given as the solutions of linear equations and therefore defined explicitly in terms of matrices. Details of algorithms for the

4

model at (1.1) are given in Section 2.2. Estimating equations for the model at (1.2), and elementary methods for their solution, are discussed in Appendix A.4. There, as in the case of the model at (1.1), estimators are defined explicitly in terms of simple matrix formulae. Analogous estimators for the model at (1.3) are described in Appendix A.5.

Next we describe calculation of $\overline{\theta}$ when $p = 0$, the simple histogram case, under the general model (2.1). We conduct inference conditional on the X_i's, which, although taken to be constant, are usually assumed to have been generated as independent and identically distributed variables, each with a non-degenerate m-variate distribution. We have, from (2.1), that the negative loglikelihood of the data is

$$\sum_{i=1}^{n} \mathcal{L}\{Y_i^{\text{vec}}, Z_i^{\text{vec}}, \theta(X_{i1}), \ldots, \theta(X_{im_x}), \alpha\}. \tag{2.2}$$

Incorporating the discrete approximation with bins $\mathcal{B}_\ell$, we obtain

$$\sum_{i=1}^{n} \sum_{\ell_1} \cdots \sum_{\ell_{m_x}} \mathcal{L}(Y_i^{\text{vec}}, Z_i^{\text{vec}}, c_{\ell_1}, \ldots, c_{\ell_{m_x}}, \alpha) \, I(X_{i1} \in \mathcal{B}_{\ell_1}, \ldots, X_{im_x} \in \mathcal{B}_{\ell_{m_x}}). \tag{2.3}$$

Assuming there are L histogram bins and that α is q-variate, and writing $\mathcal{L}_\alpha$ and $\mathcal{L}_{c_\ell}$ for the derivative of $\mathcal{L}$ with respect to α and c_ℓ, respectively, for $1 \leq \ell \leq L$, we deduce that the estimators $\widehat{c}_1, \ldots, \widehat{c}_L$ and $\widehat{\alpha}$ which we seek are defined by the $L + q$ equations

$$\sum_{i=1}^{n} \sum_{\ell_1} \cdots \sum_{\ell_{m_x}} \mathcal{L}_\psi \left(Y_i^{\text{vec}}, Z_i^{\text{vec}}, c_{\ell_1}, \ldots, c_{\ell_{m_x}}, \alpha\right) \times I(X_{i1} \in \mathcal{B}_{\ell_1}, \ldots, X_{im_x} \in \mathcal{B}_{\ell_{m_x}}) = 0, \tag{2.4}$$

where $\psi = c_\ell$ for $1 \leq \ell \leq L$ or $\psi = \alpha$.

When $p = 0$ the estimator $\overline{\theta} = \widetilde{\theta}$ is visibly unsmooth, but this difficulty may be alleviated by computing $\widetilde{\theta}$ for a range of values of x^0 in the definition of $\mathcal{B}_j$, and averaging the result, reflecting Scott's (1985) averaged shifted histogram (or ASH) approach. Another simple solution is to pass a polynomial interpolant through the bin centers. If the interpolant is of degree $2q + 1$, if θ has $2q + 2$ bounded derivatives, and if h is taken to be asymptotic to a constant multiple of $n^{-1/(4q+5)}$, then the resulting interpolated histogram estimator of θ converges to θ at the optimal rate, $n^{-2(q+1)/(4q+5)}$. When $q = 0$ the interpolant is linear, and we denote it by $\widetilde{\theta}_{\text{lin}}$.

Specifically, suppose the histogram has L bins, with respective bin centers $x_\ell = x^0 + (\ell - \frac{1}{2})h$ and each of width $h = L^{-1}$, in the interval $[0, 1]$. Then

$$\widetilde{\theta}_{\text{lin}}(x) = h^{-1}\{(x_{\ell+1} - x)\widehat{c}_\ell + (x - x_\ell)\widehat{c}_{\ell+1}\}, \tag{2.5}$$

5

provided $x_\ell \leq x \leq x_{\ell+1}$ and $1 \leq \ell \leq L$, and by

$$\widetilde{\theta}_{\text{lin}}(x) = \begin{cases} h^{-1}\left\{(x_2 - x)\,\widehat{c}_1 + (x - x_1)\,\widehat{c}_2\right\} & \text{if } 0 \leq x \leq x_1, \\ h^{-1}\left\{(x_L - x)\,\widehat{c}_{L-1} + (x - x_{L-1})\,\widehat{c}_L\right\} & \text{if } x_L \leq x \leq 1. \end{cases} \quad (2.6)$$

An alternative to interpolation is smoothing. In particular, we may pass a local polynomial smoother through the sequence of points $(x_\ell, \widehat{c}_\ell)$, representing bin centers and bin heights respectively. Here we would use a new bandwidth h_1, say; see Section 2.2 for details. Let $\widehat{\theta}$ denote the resulting estimator of θ. If h_1 is of conventional size, and h is smaller, then $\widehat{\theta}$ has virtually the same bias and variance properties as the difficult-to-compute iterative polynomial smoother $\widehat{\theta}_{\text{pol}}$ based on bandwidth h_1. Moreover, $\widehat{\theta}_{\text{pol}}$ can be defined only by iterative algorithms, and needs a good starting point, such as our estimator $\widetilde{\theta}$.

The histogram bins, $\mathcal{B}_j = (x^0 + (j-1)h, x^0 + jh]$, have respective centers $x_j = x^0 + (j - \frac{1}{2})h$. Passing a local linear smoother, using a kernel K and bandwidth h_1 say, through the data pairs $(x_j, \widehat{c}_j)$, we obtain a smooth estimator $\widehat{\theta}$ of θ. Specifically, let K be a kernel and h_1 a new bandwidth, let a_0 and a_1 be the intercept and slope in a local linear regression problem, choose $(\widehat{a}_0, \widehat{a}_1) = (a_0, a_1)$ to minimize

$$\sum_{\ell=1}^{L} \left[\widehat{c}_\ell - \{a_0 + a_1(x - x_\ell)\}\right]^2 K\left(\frac{x - x_\ell}{h_1}\right),$$

and put $\widehat{\theta}(x) = \widehat{a}_0$. More explicitly, $\widehat{\theta}(x) = \sum_\ell \widehat{c}_\ell w_\ell(x)$, where

$$\begin{aligned} w_l(x) &= (nh_1)^{-1}\left\{s_0(x)\,s_2(x) - s_1(x)^2\right\}^{-1}\left\{s_2(x) - s_1(x)\,\frac{x - x_\ell}{h}\right\} K\left(\frac{x - x_\ell}{h}\right), \\ s_j(x) &= \frac{1}{nh_1}\sum_{\ell=1}^{L}\left(\frac{x - x_\ell}{h}\right)^j K\left(\frac{x - x_\ell}{h}\right) \end{aligned} \quad (2.7)$$

Properties of $\widehat{\theta}$ are discussed in Section 4.2.

2.2 The Model at (1.1)

Here we specialize our method to the case of the model at (1.1), showing that it gives elementary and explicit estimators there. Assuming Normal errors ϵ_{ij}, and using a simple (i.e. $p = 0$) histogram approximation to θ with height c_ℓ on the ℓth bin, we see that in the case of (1.1), (2.4) reduces to

$$-\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{\ell_1}\sum_{\ell_2}(Y_{ij} - c_{\ell_1})\,\sigma^{jk}\,(Y_{ik} - c_{\ell_2})\,I(X_{ij} \in \mathcal{B}_{\ell_1}, X_{ik} \in \mathcal{B}_{\ell_2}) - \tfrac{1}{2}\log|\Sigma|, \quad (2.8)$$

6

where $\Sigma^{-1} = (\sigma^{jk})$, $|\Sigma|$ is the determinant of Σ, and I denotes the indicator function. Of course, the assumption of Normality serves only to define least-squares estimators, which are valid in much more general settings.

We may estimate Σ in a variety of ways. One is to estimate all unknowns, both $c = (c_\ell)$ and Σ, simultaneously. Another is to use a "guess," V say, at Σ; then estimate c under the temporary assumption $\Sigma = V$, producing a histogram estimator $\breve{\theta}$, say, of θ; and subsequently estimate Σ from the residuals, $\hat{\epsilon}_{ij} = Y_{ij} - \breve{\theta}(X_{ij})$. This estimator, $\hat{\Sigma}$ say, is robust against errors in the guess, V. Given a general, symmetric matrix estimator, $\hat{\Sigma} = (\hat{\sigma}_{jk})$, of Σ, and hence an estimator $\hat{\Sigma}^{-1} = (\hat{\sigma}^{jk})$ of Σ^{-1}, we obtain from (2.8) the approximate negative loglikelihood:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{\ell_1} \sum_{\ell_2} (Y_{ij} - c_{\ell_1}) \, \hat{\sigma}^{jk} \, (Y_{ik} - c_{\ell_2}) \, I(X_{ij} \in \mathcal{B}_{\ell_1}, X_{ik} \in \mathcal{B}_{\ell_2}),$$

the minimum of which gives $\hat{c} = (\hat{c}_\ell)$ very simply as the solution of the equation

$$\hat{c}^{\mathrm{T}} \hat{A} = 1^{\mathrm{T}} \hat{S}, \tag{2.9}$$

where $1^{\mathrm{T}} = (1, \dots, 1)$, $\hat{A} = (\hat{a}_{\ell_1 \ell_2})$, $\hat{S} = (\hat{s}_{\ell_1 \ell_2})$ and

$$\left. \begin{aligned} \hat{a}_{\ell_1 \ell_2} \\ \hat{s}_{\ell_1 \ell_2} \end{aligned} \right\} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} \sigma^{jk} I(X_{ij} \in \mathcal{B}_{\ell_1}, X_{ik} \in \mathcal{B}_{\ell_2}) \times \left\{ \begin{aligned} & 1 \\ & Y_{ij} \end{aligned} \right., \tag{2.10}$$

respectively.

The histospline estimator can be viewed as an extension of the classical regression spline estimator to model (1.1) with step base functions. Huang, et al. (2002) consider estimation using general basis expansions in model (1.1) by ignoring the within-cluster correlation and assuming working independence. We here account for the within-cluster correlation in our estimation. We show in section 4.1 that our estimator effectively accounts for the within-cluster correlation, and is most efficient when the true covariance matrix is used and is more efficient than the working independence histospline estimator.

Of course, when we apply the argument leading to (2.10) in the case of the models at (1.2) or (1.3), we obtain a semiparametric estimator of α as well as our nonparametric estimator of θ. As we argue in Section 4, the former is, under mild regularity conditions, root-n

consistent (and asymptotically efficient) for β, and the latter converges at the same rate as does the estimator of θ in the case of the model at (1.1). Moreover, these results pertain to one-step estimators, where the binwidth, h, is chosen so that it gives optimal performance for the estimator of θ. It is not necessary to use different binwidths for estimating θ and α.

3 NUMERICAL EXAMPLES

To understand the performance of the methods in the model at (1.1), we ran a small simulation with $n = 1,000$ clusters and $m = 3$ observations per cluster. The correlation structure was either exchangeable with correlation $\rho = 0.6$, autoregressive with correlation $\rho = 0.6$, or nearly singular with the correlations between the first and second observations and the second and third being 0.80, while the correlation between the first and third observations was 0.50. There were 200 simulations. In all cases, referring to (1.1), $\text{var}(\epsilon_{ij}) = 1$. Six functions were evaluated: (a) $\sin(4x - 2)$; (b) $\exp(4x - 2)$; (c) $\sin(2(4x - 2))$; (d) $\sqrt{x(1 - x)}\sin\{2\pi(1 + 2^{-3/5})/(x + 2^{-3/5})\}$; (e) $\sqrt{x(1 - x)}\sin\{2\pi(1 + 2^{-7/5})/(x + 2^{-7/5})\}$; and (f) $\sin(8x - 4) + 2\exp\{-256(x - .5)^2\}$.

Five estimators were considered:

- A working independence ordinary local linear kernel smoother using Epanechnikov kernel with support $[-1, 1]$. The bandwidth was chosen using the DPI bandwidth selection method of Ruppert, et al. (1995).

- The linear interpolant method (2.5)–(2.6) under working independence.

- The linear interpolant method (2.5)–(2.6) with estimated covariance matrix.

- The two–stage histospline–kernel method under working independence, see definition above (2.7)

- The two–stage histospline–kernel method with estimated covariance matrix.

The covariance matrix was estimated by first running the local linear smoother under working independence, forming the residuals, and then forming the covariance matrix of the residuals. See also the discussion after (4.3). For the linear interpolant methods, the number of bins was selected by leaving one cluster out cross-validation from among $L = 4, 8, ..., 32$.

For the two–stage histospline–kernel methods, we used a local linear kernel smoother with the Epanechnikov kernel with support $[-1, 1]$. The number of bins at the first stage was

selected from among $L = 5, 10, ..., 45$. Whatever the number of bins, the bandwidth at the second stage was chosen again using the DPI bandwidth selection method of Ruppert, et al. (1995) applied to the "pseudo-data" consisting of the bin centers and first stage histospline estimates The number of bins at the first stage was selected by leaving one cluster out cross-validation. This particular use of the DPI methods, namely with pseudo-data, is not optimal, because the bandwidth at the second stage should be determined by the number of observations, not the number of bins used. To check this, we considered two additional methods. In the first, we used the bandwidth selected by the working independence ordinary local linear kernel smoother. In the second, we used cross-validation for both the number of bins and the bandwidth. Neither alternative improved upon the first method overall, and indeed the latter was generally worse, see below.

The results are given in Table 1 for the 3 correlation structures. As seen in Table 1, in all cases the weighted methods using the estimated covariance outperform their working independence counterparts. Thus, for example, the weighted linear interpolant method by accounting for the correlation always has mean squared error efficiency higher than that of the linear interpolant method without accounting for the correlation. The weighted linear interpolant method by accounting for the correlation has either modest losses or modest gains compared to the working independence kernel method except in the nearly unstructured case. The two–stage histospline-kernel method generally outperforms both the working independence kernel and the linear interpolant method. As expected, the gains in efficiency are greatest for the nearly singular covariance matrix case.

We also ran but do not report here the same simulations when the data were independent. As expected, the weighted and unweighted linear interpolant methods were essentially equivalent, as were the weighted and unweighted two–stage methods, thus indicating little if any loss of efficiency for estimating the covariance matrix. The working independence local linear kernel and two–stage methods were also essentially equivalent, with the latter being on average 10% less efficient. The linear interpolant methods were on average approximately 25% less efficient than the local linear kernel methods under independence.

Finally, we comment briefly on an alternative approach. As indicated in Section 2.2, histosplines are simply regression splines of order 0, the basis functions being the indicators

9

of belonging to specific intervals. An alternative is to replace this set of basis functions by smoother bases, e.g., the truncated power series basis functions or the Bspline basis functions. For the former, we placed knots $(\kappa_1, ..., \kappa_L)$ at the sample quantiles of the X's, and the basis functions were $\{1, x, x^2, x^3, (x - \kappa_1)^3_+, ..., (x - \kappa_L)^3_+\}$, where the subscript $+$ indicates truncation at zero. For the latter, we placed the knots as equally spaced. For both sets of basis functions, we selected the number of knots by cross-validation. We found that the truncated power series basis led to estimates that were generally inferior to our two–stage methods due to numerical instability associated with such basis functions, which are usually handled via penalization, see Ruppert, et al. (2003). Use of the cubic Bspline basis functions led to estimates that were roughly comparable to our two–stage methods, being inferior in the exchangeable case, nearly equivalent in the autoregressive case and somewhat superior in the near–singular case. We have, unfortunately, no intuition as to why this happens. See Chen and Jin (2003) for more details about this approach for the model at (1.2).

4 THEORETICAL PROPERTIES

4.1 Properties Of Histogram Estimator Under Model (1.1)

In this section and the next we address least-squares procedures, and therefore construct the likelihood $\mathcal{L}$ under the assumption of Normal errors, even if the errors do not have that distribution. Many other settings are possible, but they generally require different regularity conditions, and for brevity we do not follow that route.

Let $\widehat{\Sigma}$ denote a symmetric $m \times m$ matrix, perhaps determined by the data and representing an approximation to, or guess at, the covariance matrix Σ of the errors ϵ_i. Assume $\widehat{\Sigma}$ converges in probability to a strictly positive definite matrix V, write $V^{-1} = (v^{jk})$, suppose the design points are distributed on the interval $[0, 1]$, let f_j and f_{jk} be the respective densities of X_{ij} and (X_{ij}, X_{ik}), let $f_{jk}^{(1,0)}$ denote the first derivative of f_{jk} with respect to its first component, and define the function ξ by

$$12\, \xi(x) = \sum_{j=1}^{m} v^{jj} \{\theta'(x)\, f_j'(x) + \theta''(x)\, f_j(x)\}$$

10

$$+ \sum\sum_{1 \le j \ne k \le m} v^{jk} \int_0^1 \left\{ \theta'(u)\, f_{jk}^{(1,0)}(u,x) + \theta''(u)\, f_{jk}(u,x) \right\} du\,.$$

Put $\eta = \sum_j v^{jj} f_j$ and $\zeta = \sum\sum_{j \ne k} f_{jk}$, denoting univariate and bivariate functions, respectively, and let b represent the unique solution of the linear integral equation

$$\xi(x) = b(x)\,\eta(x) + \int_0^1 b(u)\,\zeta(u,x)\,du\,. \tag{4.1}$$

Define the matrices $\widehat{A} = (\widehat{a}_{\ell_1 \ell_2})$ and $\widehat{S} = (\widehat{s}_{\ell_1 \ell_2})$ by (2.10), and then let $\widehat{c} = (\widehat{c}_\ell)$ be given by (2.9).

Our first result gives uniform stochastic approximations to the histospline estimator $\widehat{c}_\ell$ at the bin center x_l. Regularity conditions (4.9)–(4.14) are given in Section 4.3.

Theorem 4.1. *Assume conditions* (4.9)–(4.13) *below, and that the data are generated by the model at* (1.1). *Then,*

$$\widehat{c}_\ell = \theta(x_\ell) + h^2\, b(x_\ell) + \frac{1 + o_p(1)}{nh\,\eta(x_\ell)} \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^m v^{jk}\, I(X_{ik} \in \mathcal{B}_\ell)\, \epsilon_{ij} + o_p(h^2)\,, \tag{4.2}$$

where the $o_p(\cdot)$ terms are of the stated orders uniformly in $1 \le \ell \le L$. Furthermore,

$$\max_{1 \le \ell \le L} |\widehat{c}_\ell - \theta(x_\ell)| = O_p\{h^2 + (nh)^{-1/2}\,(\log n)^{1/2}\}\,. \tag{4.3}$$

Next we discuss aspects and implications of the theorem, beginning with the construction of $\widehat{\Sigma}$. If $\widehat{\Sigma}$ is computed from the data then it is likely to be based on a relatively coarse approximation, $\breve{\theta}$ say, to θ, such as the linear histospline calculated from the histogram using $V = I_m$ (the $m \times m$ identity matrix) and a bandwidth of size $n^{-1/5}$. Such a version of $\breve{\theta}$ converges to θ at rate $n^{-2/5}$, and the naive estimator, $\widehat{\Sigma}$ say, of $\Sigma = \mathrm{var}(\epsilon_i)$, computed from the residuals $\widehat{\epsilon}_{ij} = Y_{ij} - \breve{\theta}(X_{ij})$, satisfies $\widehat{\Sigma} = \Sigma + O_p(n^{-2/5})$. This rate can be improved to $O_p(n^{-1/2})$ by undersmoothing $\breve{\theta}$, although this slight enhancement does not affect first-order properties of our estimators of θ.

Note that this particular choice of $\widehat{\Sigma}$ satisfies the regularity condition (4.12) below, and so, provided Σ is nonsingular, our $\widehat{c}_\ell$ satisfy (4.2) with $v^{jk} = \sigma^{jk}$, where $\Sigma^{-1} = (\sigma^{jk})$. The resulting estimators achieve a minimum variance bound in the sense of Wang (2003), e.g.,

11

smaller variance than any working independence estimator and the smallest variance among our class of estimators based on a working covariance matrix.

For a general choice of V it is readily proved that the random variable

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk} I(X_{ik} \in \mathcal{B}_\ell) \, \epsilon_{ij} \,,$$

appearing at (4.2), is asymptotically Normally distributed with zero mean and variance $nh \sum_j (V^{-1}\Sigma V^{-1})^{jj} f_j(x_\ell)$. Therefore, a central limit theorem for $\hat{c}_\ell$ is directly obtainable from (4.10). More to the point, properties of linear histospline estimators, constructed by interpolating the histogram with bin centers and heights x_ℓ and $\hat{c}_\ell$ respectively, are easily obtainable from the theorem.

Indeed, our next result, essentially a corollary of Theorem 4.1, describes properties of $\widetilde{\theta}_{\mathrm{lin}}$, defined at (2.5) and (2.6). Put $\tau_0(x)^2 = \eta(x)^{-2} \sum_j (V^{-1}\Sigma V^{-1})^{jj} f_j(x)$ for $0 \le x \le 1$, and, for $0 < x < 1$,

$$\tau(x)^2 = h^{-2} \left\{ (x_\ell - x)^2 + (x_{\ell+1} - x)^2 \right\} \tau_0(x)^2 \in \left[\tfrac{1}{2}\tau_0(x)^2, \tau_0(x)^2 \right],$$

with ℓ chosen so that $x_\ell \le x \le x_{\ell+1}$ and, for $x = 0$, $\tau(0)^2 = \tfrac{5}{2}\tau_0(0)^2$.

Theorem 4.2. *Assume conditions (4.9)–(4.13) below, and that the data are generated by the model at (1.1). Define $\widetilde{\theta}_{\mathrm{lin}}$ as above. Then for each $x \in (0,1)$,*

$$\widetilde{\theta}_{\mathrm{lin}}(x) = \theta(x) + \tfrac{1}{2}(x - x_\ell)(x_{\ell+1} - x)\,\theta''(x) + b(x)\,h^2 + \tau(x)\,(nh)^{-1/2}\,Z_n(x) + o_p(h^2)\,, \quad (4.4)$$

where the random variable $Z_n(x)$ is asymptotically Normal $N(0,1)$. When $x = 0$,

$$\widetilde{\theta}_{\mathrm{lin}}(0) = \theta(0) - \theta''(0)\,h^2 + b(0)\,h^2 + \tau(0)\,(nh)^{-1/2}\,Z_n(0) + o_p(h^2)\,, \quad (4.5)$$

where $Z_n(0)$ is asymptotically Normal $N(0,1)$. The analogous result holds when $x = 1$. Furthermore,

$$\sup_{0 \le x \le 1} |\widetilde{\theta}_{\mathrm{lin}}(x) - \theta(x)| = O_p\{ h^2 + (nh)^{-1/2}(\log n)^{1/2} \}\,. \quad (4.6)$$

12

4.2 Properties of Kernel–Smoothed Histogram Estimator Under Model (1.1)

Here we discuss properties of the estimator $\widehat{\theta}$ obtained by passing a linear smoother, using the bandwidth h_1, through the data $(x_\ell, \widehat{c}_\ell)$ for $1 \le \ell \le L$, the latter coming from a histogram with binwidth h. A definition of $\widehat{\theta}$ is given in the formulae leading to (2.7). If h and h_1 are taken to be of approximately the same size, meaning that h/h_1 is bounded away from zero and infinity as $n \to \infty$, then it may be shown that $\widehat{\theta}(x)$, like $\widetilde{\theta}_{\mathrm{lin}}(x)$, has bias and variance of orders h_1^2 and $(nh_1)^{-1}$, respectively, as $n \to \infty$.

Of more interest is the case where the histogram is constructed by undersmoothing, i.e. $h/h_1 \to 0$ as $n \to \infty$. There the asymptotic bias formula for $\widehat{\theta}(x)$ is the traditional one associated with a local linear estimator, and the asymptotic variance formula also has the traditional local-linear form, except that the error variance, which is no longer appropriate, since the errors are correlated, is replaced by the minimum variance bound, $\tau_0(x)^2$, for grouped data.

Details are given in the theorem below. Note particularly that, unlike the case for the estimator $\widetilde{\theta}_{\mathrm{lin}}$, the bias of $\widehat{\theta}$ does not depend on the design density f. As discussed by Fan (1993), this is the key to strong theoretical performance. The estimator $\widehat{\theta}$ is 100% efficient, in the class of linear estimators computed under the assumption that Σ is known, and in the sense of Fan (1993).

Define $\kappa = \int K^2$ and $\kappa_2 = \int u^2 K(u)\, du$, and let $b(c)$ and $v(c)$ denote the standard asymptotic bias and variance constants for local linear estimators at a point that is distant ch from a boundary; for definitions, see p. 74 of Fan and Gijbels (1995). If K vanishes outside an interval $(-s, s)$ then $b(c) = \kappa_2$ and $v(c) = \kappa$ for $c \ge s$.

Theorem 4.3. *Assume conditions (4.9)–(4.14), that the data are generated by the model at (1.1), and that $h_1 = h_1(n) \to 0$ and $h/h_1 \to 0$. Then for each $x \in (0,1)$,*

$$\widehat{\theta}(x) = \theta(x) + \tfrac{1}{2}\,\kappa_2\,\theta''(x)\,h_1^2 + (nh_1)^{-1/2}\,\kappa^{1/2}\,\tau_0(x)\,Z_n(x) + o_p(h_1^2)\,, \tag{4.7}$$

where the random variable $Z_n(x)$ is asymptotically Normal $N(0,1)$. When $x = ch$ where

$0 \leq c < \infty$,

$$\widehat{\theta}(0) = \theta(0) + \tfrac{1}{2} b(c) \, \theta''(0) \, h_1^2 + (nh_1)^{-1/2} \, v(c)^{1/2} \, \tau_0(0) \, Z_n(0) + o_p(h_1^2), \qquad (4.8)$$

where $Z_n(0)$ is asymptotically Normal $N(0, 1)$. The analogous result holds when $x = 1$. Furthermore,

$$\sup_{0 \leq x \leq 1} |\widehat{\theta}(x) - \theta(x)| = O_p\{h_1^2 + (nh)^{-1/2} \, (\log n)^{1/2}\}.$$

Results (4.7) and (4.8) imply that, as in the case of conventional second-order curve estimators, the asymptotically optimal choice of h_1 equals a constant multiple of $n^{-1/5}$. For this choice of h_1, condition (4.13) below permits a relatively wide range of selections of h which allow the undersmoothing (i.e. $h/h_1 \to 0$ as $n \to \infty$) that is necessary for the theorem.

4.3 Assumptions and Discussion

First we state regularity conditions for Theorems 4.1–4.3. Of course, the function $\theta(x)$ is assumed to satisfy the usual smoothness conditions of being twice continuously differentiable.

the distribution of $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{im})^{\mathrm{T}}$ has zero mean, finite third moments and covariance matrix Σ, $\qquad (4.9)$

the support of the density f of X_i equals the unit cube $[0, 1]^m$, f has a continuous derivative there and is bounded away from zero and infinity, and θ has two continuous derivatives on $[0, 1]$, $\qquad (4.10)$

the vectors X_i and ϵ_i, $1 \leq i \leq n$, are completely independent, $\qquad (4.11)$

the symmetric random matrix $\widehat{\Sigma} = (\widehat{\sigma}_{jk})$ satisfies $\max_{j,k} |\widehat{\sigma}_{jk} - v_{jk}|$ $= O_p(h^{1/2})$, where $V = (v_{jk})$ is strictly positive definite, $\qquad (4.12)$

$h = L^{-1}$, where $L > 1$ is an integer; as $n \to \infty$, $h = h(n) \to \infty$ and for some $\delta > 0$, $n^{(1/4)-\delta} h \to \infty$; m is fixed; and x^0, in the definition of the bins $\mathcal{B}_\ell = (x^0 + (\ell-1)h, x^0 + \ell h]$, equals 0, $\qquad (4.13)$

K is a bounded, symmetric, Hölder continuous, compactly supported probability density. $\qquad (4.14)$

The assumption at (4.13) that h is the inverse of an integer, and $x^0 = 0$, serves only to ensure there is a whole number of bins $\mathcal{B}_\ell$ in $[0, 1]$. In particular, there are no bin fragments at the ends of the interval. Minor modifications of our arguments allow the case of bin fragments to be treated. Note that we do not specify how $\widehat{\Sigma}$ is constructed. It might be calculated directly from data, or it could be a deterministic guess at the true Σ.

14

4.4 Theory Under Other Models, Especially (1.2) and (1.3)

Here we make a few remarks about other models. For simplicity we confine attention to the simple histogram case (i.e. $p = 0$), where general methodology for estimating the bin heights $\widehat{c}_\ell$ and discrete parameters $\widehat{a}$ was described in Section 2; see (2.2)–(2.4). Limit theory for estimation of θ is very similar under the models (1.2) and (1.3) to what it is under (1.1). Slightly altered versions of Theorems 4.1–4.3 hold, having the same convergence rates but different bias and variance formulae.

In particular, assuming (1.2) or (1.3), the histospline estimator $\widetilde{\theta}_{\text{lin}}$ (i.e. the case $p = 0$) has asymptotic bias and variance of orders h^2 and $(nh)^{-1}$, respectively, and uniform convergence rate $O_p\{h^2 + (nh)^{-1/2}(\log n)^{1/2}\}$. Under (1.2) and (1.3) the bias of $\widehat{c}_\ell$ admits a Taylor expansion, $E(\widehat{c}_\ell) = \theta + h^2\, b(x) + o(h^2)$, where the bias function b satisfies an integral equation analogous to (4.1). In the context of (1.2) the bias functions and variance formulae in Theorems 4.2–4.3 are in fact identical in the cases of models (1.1) and (1.2); see below for further discussion.

Under (1.2) and (1.3), in work to be reported elsewhere, we have shown that the semiparametric estimator of β achieves a semiparametric minimum variance bound and has bias of smaller order than $n^{-1/2}$. The efficiency bound itself for (1.2) was exhibited by Lin and Carroll (2001a), while a semiparametric efficient estimate using interactive kernel methods was developed by Wang, et al. (2003).

The fact that $\widehat{\beta}$ converges at rate $n^{-1/2}$ suggests that the properties of $\widetilde{\theta}_{\text{lin}}$ and $\widehat{\theta}$ are equivalent, to first order, to those that would be obtained if $\widetilde{\theta}_{\text{lin}}$ and $\widehat{\theta}$ were computed with β replaced by its true value, without attempting estimation of β. This is readily proved to be the case for models (1.2) and (1.3), and also in more general cases. The method of proof is straightforward; see Hall, Reiman and Rice (1999) for a recent example.

5   CONCLUDING REMARKS

The models (1.1)–(1.3) and its more general version (2.1) have significant structure that makes it difficult to adopt standard local linear methods to them. For example, in model

<center>15</center>

(1.1), longitudinal/clustered data with a marginal mean structure, traditional kernel approaches are known to have failed to successfully construct methods that can account for correlation structure and produce an estimator with good variance properties. In model (1.3), the financed–inspired model, even constructing a computationally feasible local linear method has proved difficult, much less one with good variance properties.

Our histospline technique converts the problem from one in the continuum to one that is governed by only a finite number of parameters. As we have shown, this can greatly lessen the computational burden, largely by avoiding some of the iteration necessary in other methods. As a general tool, at the very least the histospline approach can be used to produce starting values for iterative methods. This ease of computation is not, however, accompanied by any loss of asymptotic efficiency for both the nonparametric model (1.1) and the semiparametric models ((1.2)–(1.3). A referee has pointed out that the histospline method could be viewed as a member of the class of estimation methods based on finite-dimensional linear estimating spaces, for which Huang (1998, 2003) has developed general theory not in our problem but in the independent data case.

We have pointed out two major advantages of our approach, namely that its computation can be very simple and yet the method can lead to recognizable asymptotic bias and variance calculations in the nonparametric case, and semiparametric efficient estimates in the semiparametric case. We have also pointed out that our estimates can be used as starting values for other approaches that hope to capture more small sample efficiency. A few remarks then about small sample efficiency comparisons may then be in order.

Consider the model at (1.1). Here it is known that Wang's (2003) computationally more complex iterative kernel methods dominate working independence kernel methods. She also showed via simulation that her iterative kernel methods are as efficient as generalized least squares (GLS) penalized regression and smoothing splines, both of which are straightforward to implement in the model at (1.1). In simulations not reported here, we compared our simple histospline techniques with GLS penalized regression splines, and hence from Wang's simulations also to her method. For the exchangeable and autocorrelated cases in Table 1, our two-stage kernel methods were on average approximately 10% less efficient in mean squared error than the spline method, while for the nearly singular case the loss of efficiency

16

was approximately 15%–20%. When compared to Wang's iterative kernel methods, this slight loss of efficiency may be an acceptable price to pay for computational convenience. When compared to penalized regression splines and smoothing splines, we again point out that our methods are rather more straightforward to implement and computationally have closed form solution, but have the same asymptotic efficiency as smoothing splines.

ACKNOWLEDGMENTS AND SOFTWARE

REFERENCES

Brumback, B. and Rice, J. A. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves (with discussions). *Journal of the American Statistical Association*, 93, 961-1006.

Carroll, R. J., Härdle, W. and Mammen, E. (2002). Estimation in an additive model when components are linked parametrically. *Journal of Econometrics*, 18, 886–912.

Chen, K. and Jin, Z. (2003). Partial linear regression models for clustered data. Preprint.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics*, 21, 196–216.

Fan, J. and Gijbels, I. (1995). *Local Polynomial Modeling and its Applications.* Chapman and Hall, London.

Hall, P., Reimann, J. and Rice, J. (2000). Nonparametric estimation of a periodic function. *Biometrika*, 87, 545–557.

Hafner, C. M. (1998). *Nonlinear Time Series Analysis with Applications to Foreign Exchange Rate Volatility.* Physica, Heidelberg and New York.

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

Huang, J., Wu, C. and Zhou, L. (2002). Varying-coefficient models and basis function approximation for the analysis of repeated measures. *Biometrika*, 89, 111-128.

Huang, J. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *Annals of Statistics*, 26, 242-272.

Huang, J. (2001). Concave extended linear modeling: A theoretical synthesis. *Statistica Sinica*, 11, 173-197.

Huang, J. (2003). Local asymptotics for polynomial spline regression. *Annals of Statistics*, in press.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman and Hall, London.

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

Lin, X. and Carroll, R. J. (2001a). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, 96, 1045–1056.

Lin, X. and Carroll, R. J. (2001b). Semiparametric regression for clustered Data. *Biometrika*, 88, 1179–1865.

Lin, X., Wang, N., Welsh, A. H. and Carroll, R. J. (2003). Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data. Preprint.

Lin, D. Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 96, 103–126.

Linton, O. B., Mammen, E., Lin, X. and Carroll, R. J. (2003). Accounting for correlation in marginal longitudinal nonparametric regression. *Second Seattle Symposium on Biostatistics*, editor D. Lin, to appear.

Petrov, V. V. (1975). *Sums of Independent Random Variables.* Springer, Berlin.

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

Ruppert, D., Sheather, S. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression (Corr: 96V91 p1380). *Journal of the American Statistical Association*, 90, 1257–1270.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

Scott, D. W. (1985). Average shifted histograms: effective nonparametric density estimators in several dimensions. *Annals of Statistics*, 13, 1024–1040.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer Verlag, New York.

Stone, C. J. (1991). Asymptotics for doubly flexible logspline response models. *Annals of Statistics*, 19 , 1832-1854.

Verbyla, A. P., Cullis, B. R., Kenward, M. G. and Welham, S. J. (1999). The Analysis of Designed Experiments and Longitudinal Data Using Smoothing Splines (with discussion). *Applied Statistics*, 48, 269–311.

Wahba, G. (1991). *Spline Models for Observational Data*, SIAM, Philadelphia.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, to appear.

Wang, N., Carroll, R. J. and Lin, X. (2003). Efficient marginal estimation for longitudinal/clustered data. Manuscript.

Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B*, 60, 159-174.

Wild, C. J. and Yee, T. W. (1996). Additive extensions to generalized estimating equation methods. *Journal of the Royal Statistical Society, Series B*, 58, 711–725.

Wu, C. O., Chiang, C. T. and Hoover, D. R. (1998). Asymptotic Confidence Regions for Kernel Smoothing of a Varying Coefficient Model with Longitudinal Data. *Journal of the American Statistical Association*, 93, 1388–1402.

Zeger, S. L. and Diggle, P. J. (1994). Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, 50, 689-699.

Zhang D., Lin X., Raz J., and Sowers M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, 93, 710-719.

Appendix: Technical Arguments

A.1 Proof of Theorem 4.1

Observe that $\widehat{S} = \widehat{S}_1 + \widehat{S}_2$ where, for $r = 1, 2$, we define

$$
\left.\begin{array}{r}
(\widehat{S}_1)_{\ell_1\ell_2} \\[2mm]
(\widehat{S}_2)_{\ell_1\ell_2} \\[2mm]
(\widehat{T}_r)_{\ell_1\ell_2}
\end{array}\right\} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} \widehat{\sigma}^{jk}\, I(X_{ij} \in \mathcal{B}_{\ell_1}, X_{ik} \in \mathcal{B}_{\ell_2}) \times \left\{\begin{array}{c}
\theta(X_{ij}) \\[2mm]
\epsilon_{ij} \\[2mm]
(X_{ij} - x_{\ell_1})^r
\end{array}\right. ,
$$

respectively. Note too that

$$
\theta(X_{ij}) = \theta(x_\ell) + (X_{ij} - x_\ell)\, \theta'(x_\ell) + \tfrac{1}{2}\,(X_{ij} - x_\ell)^2\, \theta''(x_\ell) + o_p(h^2)\,,
$$

uniformly in i, j, ℓ such that $X_{ij} \in \mathcal{B}_\ell$. Therefore,

$$
(\widehat{S}_1)_{\ell_1\ell_2} = \theta(x_{\ell_1})\,\widehat{a}_{\ell_1\ell_2} + \sum_{r=1}^{2} \frac{1}{r}\, \theta^{(r)}(x_{\ell_1})\, (\widehat{T}_r)_{\ell_1\ell_2} + o_p(h^2|\widehat{a}_{\ell_1\ell_2}|)\,, \tag{A.1}
$$

uniformly in $1 \le \ell_1, \ell_2 \le L$.

For $r = 0, 1, 2$, define

$$
U_{\ell_1\ell_2}(j, k, r) = \frac{1}{n} \sum_{i=1}^{n} I(X_{ij} \in \mathcal{B}_{\ell_1}, X_{ik} \in \mathcal{B}_{\ell_2})\, (X_{ij} - x_{\ell_1})^r\,.
$$

If $j \ne k$ then the variance of $U_{\ell_1\ell_2}(j, k, r)$ is asymptotic to $n^{-1}h^{2r+2}g_{jk}(x_{\ell_1}, x_{\ell_2})$, where the function g_{jk} is bounded away from zero and infinity; if $j = k$ and $\ell_1 = \ell_2$, the asymptote is $n^{-1}h^{2r+1}g_j(x_{\ell_1})$, where g_j is bounded away from zero and infinity; and $U_{\ell_1\ell_2}(j, k, r) = 0$ if $j = k$ and $\ell_1 \ne \ell_2$. Therefore, using Bernstein's inequality we may prove that for each $C, \delta > 0$, and for $r = 0, 1, 2$,

$$
\max_{j \ne k, \ell_1, \ell_2} P\big\{|U_{\ell_1\ell_2}(j, k, r) - E\, U_{\ell_1\ell_2}(j, k, r)| > n^{\delta - (1/2)}\, h^{r+1}\big\} = O(n^{-C})\,,
$$

$$
\max_{j, l} P\big\{|U_{\ell\ell}(j, j, r) - E\, U_{\ell\ell}(j, j, r)| > n^{\delta - (1/2)}\, h^{r+(1/2)}\big\} = O(n^{-C})\,.
$$

Hence, for all $C, \delta > 0$,

$$
P\Big\{\max_{j \ne k, \ell_1, \ell_2} |U_{\ell_1\ell_2}(j, k, r) - E\, U_{\ell_1\ell_2}(j, k, r)| > n^{\delta - (1/2)}\, h^{r+1}\Big\} = O(n^{-C})\,, \tag{A.2}
$$

$$
P\Big\{\max_{j, l} |U_{\ell\ell}(j, j, r) - E\, U_{\ell\ell}(j, j, r)| > n^{\delta - (1/2)}\, h^{r+(1/2)}\Big\} = O(n^{-C})\,. \tag{A.3}
$$

Therefore, by (A.1),

$$
(\widehat{S})_{\ell_1\ell_2} = \theta(x_{\ell_1})\,\widehat{a}_{\ell_1\ell_2} + \sum_{r=1}^{2} \frac{1}{r}\, \theta^{(r)}(x_{\ell_1})\, (\widehat{U}_r)_{\ell_1\ell_2}
$$

$$
+ O_p\{n^{\delta - (1/2)}\, (h^2 + \delta_{\ell_1\ell_2}\, h^{3/2})\} + o_p(h^2|\widehat{a}_{\ell_1\ell_2}|)\,, \tag{A.4}
$$

20

uniformly in $1 \leq \ell_1, \ell_2 \leq L$, where

$$(\widehat{U}_r)_{\ell_1 \ell_2} = \sum_{j=1}^{m} \sum_{k=1}^{m} \widehat{\sigma}_{jk} \, E\{U_{\ell_1 \ell_2}(j,k,r)\} \,.$$

Let f_{jk} denote the joint density of (X_{ij}, X_{ik}). If $\ell_1 \neq \ell_2$ and $j \neq k$ then

$$E\{U_{\ell_1 \ell_2}(j,k,r)\} = h^{r+2} \int_{-\frac{1}{2}}^{\frac{1}{2}} u^r \, du \int_{-\frac{1}{2}}^{\frac{1}{2}} f_{jk} \left(x_{\ell_1} + hu, x_{\ell_2} + hv \right) dv \,.$$

When $r = 2$ the right-hand side equals $\frac{1}{12} h^4 f_{jk}(x_{\ell_1}, x_{\ell_2}) + o(h^4)$, uniformly in $j \neq k$ and $\ell_1 \neq \ell_2$. When $r = 1$ it equals

$$h^4 \int_{-\frac{1}{2}}^{\frac{1}{2}} u^2 \, du \int_{-\frac{1}{2}}^{\frac{1}{2}} f_{jk}^{(1,0)} \left(x_{\ell_1}, x_{\ell_2} \right) dv + o(h^4) = \frac{1}{12} h^4 \, f_{jk}^{(1,0)} \left(x_{\ell_1}, x_{\ell_2} \right) + o(h^4) \,,$$

uniformly in the same j, k, ℓ_1 and ℓ_2, where $f_{jk}^{(1,0)}(u,v) = (\partial/\partial u) f_{jk}(u,v)$. Furthermore, $E\{U_{\ell\ell}(j,k,r)\} = O(h^4)$ uniformly in $j \neq k$, $1 \leq \ell \leq L$ and $r = 1, 2$; and

$$\begin{aligned}
E\{U_{\ell\ell}(j,j,r)\} &= \int_{x_\ell - \frac{1}{2}h}^{x_\ell + \frac{1}{2}h} (x - x_\ell)^r \, f_j(x) \, dx \\
&= h^{r+1} \int_{-\frac{1}{2}}^{\frac{1}{2}} u^r \, f_j(x_\ell + u) \, du = o(h^3) + \frac{1}{12} h^3 \times \begin{cases} f_j'(x_\ell) & \text{if } r = 1 \\ f_j(x_\ell) & \text{if } r = 2 \end{cases},
\end{aligned}$$

uniformly in j, ℓ. Combining the results in this paragraph, and recalling that $U_{\ell_1 \ell_2}(j,k,r) = 0$ if $\ell_1 \neq \ell_2$ and $j = k$, we deduce that

$$\begin{aligned}
&\sum_{r=1}^{2} \frac{1}{r} \theta^{(r)}(x_{\ell_1}) \, (\widehat{U}_r)_{\ell_1 \ell_2} \\
={}& \sum_{j=1}^{m} \sum_{k=1}^{m} \widehat{\sigma}^{jk} \Big[\theta'(x_{\ell_1}) \Big\{ (1 - \delta_{\ell_1 \ell_2})(1 - \delta_{jk}) \tfrac{1}{12} h^4 \, f_{jk}^{(1,0)}(x_{\ell_1}, x_{\ell_2}) \\
&\quad + \delta_{\ell_1 \ell_2} \delta_{jk} \tfrac{1}{12} h^3 \, f_j'(x_{\ell_1}) \Big\} + \theta''(x_{\ell_1}) \Big\{ (1 - \delta_{\ell_1 \ell_2})(1 - \delta_{jk}) \\
&\quad \times \tfrac{1}{12} h^4 \, f_{jk}(x_{\ell_1}, x_{\ell_2}) + \delta_{\ell_1 \ell_2} \delta_{jk} \tfrac{1}{12} h^3 \, f_j(x_{\ell_1}) \Big\} \Big] + o_p(h^4 + \delta_{\ell_1 \ell_2} h^3) \\
={}& \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk} \Big[\delta_{\ell_1 \ell_2} \delta_{jk} \tfrac{1}{12} h^3 \left\{ \theta'(x_{\ell_1}) \, f_j'(x_{\ell_1}) + \theta''(x_{\ell_1}) \, f_j(x_{\ell_1}) \right\} \\
&\quad + (1 - \delta_{\ell_1 \ell_2})(1 - \delta_{jk}) \tfrac{1}{12} h^4 \big\{ \theta'(x_{\ell_1}) \, f_{jk}^{(1,0)}(x_{\ell_1}, x_{\ell_2}) \\
&\quad + \theta''(x_{\ell_1}) \, f_{jk}(x_{\ell_1}, x_{\ell_2}) \big\} \Big] + o_p(h^4 + \delta_{\ell_1 \ell_2} h^3) \,, \tag{A.5}
\end{aligned}$$

uniformly in $1 \leq \ell_1, \ell_2 \leq L$.

Combining (A.4) and (A.5), and noting the assumption that $n^{(1/4)-\delta} h \to \infty$ for some $\delta > 0$, we deduce that

$$(1^{\mathrm{T}} \widehat{S}_1)_\ell = \sum_{\ell_1 = 1}^{L} \theta(x_{\ell_1}) \, \widehat{a}_{\ell_1 \ell} + \tfrac{1}{12} h^3 \left\{ \xi_1(x_\ell) + \xi_2(x_\ell) \right\} + o_p \Big(h^3 + h^2 \sum_{\ell_1 = 1}^{L} |\widehat{a}_{\ell_1 \ell}| \Big) , \tag{A.6}$$

21

uniformly in $1 \le \ell \le L$, where

$$
\xi_1(x_\ell) = \sum_{j=1}^{m} v^{jj} \{\theta'(x_\ell) f_j'(x_\ell) + \theta''(x_\ell) f_j(x_\ell)\},
$$

$$
\xi_2(x_\ell) = \sum_{j=1}^{m} \sum_{k=1}^{m} (1 - \delta_{jk}) v^{jk} \int_0^1 \{\theta'(u) f_{jk}^{(1,0)}(u, x_\ell) + \theta''(u) f_{jk}(u, x_\ell)\} du.
$$

Define $V_{jk\ell} = n^{-1/2} \sum_i I(X_{ik} \in \mathcal{B}_\ell) \epsilon_{ij}$, and note that $\mathrm{var}(V_{jk\ell})$ is asymptotic to a constant multiple of h. Since the ϵ_{ij}'s have zero mean and finite $(2+\delta)$th moments, for some $\delta > 0$, then for $C > 0$ sufficiently large,

$$
\max_{j,k,l} P\{|V_{jkl}| > C (h \log n)^{1/2}\} = o(h).
$$

See Petrov (1975, p. 254), and note assumption (4.9). Therefore,

$$
P\Big\{ \max_{j,k,l} |V_{jkl}| > C (h \log n)^{1/2} \Big\} \to 0. \tag{A.7}
$$

Using this property, and recalling the assumptions $\hat{\sigma}^{jk} = v^{jk} + O_p(h^{1/2})$ and $n^{(1/4)-\delta} h \to \infty$ for some $\delta > 0$, we deduce that

$$
\begin{aligned}
(1^{\mathrm{T}} \hat{S}_2)_\ell &= n^{-1/2} \sum_{j=1}^{m} \sum_{k=1}^{m} \hat{\sigma}^{jk} V_{jkl} \\
&= n^{-1/2} \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk} V_{jkl} + O_p\{n^{-1/2} h^{1/2} (h \log n)^{1/2}\} \\
&= n^{-1/2} \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk} V_{jkl} + o_p(h^3),
\end{aligned} \tag{A.8}
$$

uniformly in $1 \le \ell \le L$.

Combining (A.6) and (A.8), and writing $\xi = \frac{1}{12}(\xi_1 + \xi_2)$, we find that

$$
\begin{aligned}
(1^{\mathrm{T}} \hat{S})_\ell &= (1^{\mathrm{T}} \hat{S}_1)_\ell + (1^{\mathrm{T}} \hat{S}_2)_\ell \\
&= \sum_{\ell_1=1}^{L} \theta(x_{\ell_1}) \hat{a}_{\ell_1 \ell} + h^3 \xi(x_\ell) + n^{-1/2} \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk} V_{jkl} + o_p\Big(h^3 + h^2 \sum_{\ell_1=1}^{L} |\hat{a}_{\ell_1 \ell}|\Big),
\end{aligned}
$$

uniformly in $1 \le \ell \le L$. The ℓth component of the vector of equations at (2.9) that define $\hat{c}$ can therefore be written as

$$
\sum_{\ell=1}^{L} \{\hat{c}_{\ell_1} - \theta(x_{\ell_1})\} \hat{a}_{\ell_1 \ell} = h^3 \xi(x_\ell) + n^{-1/2} \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk} V_{jkl} + o_p\Big(h^3 + h^2 \sum_{\ell_1=1}^{L} |\hat{a}_{\ell_1 \ell}|\Big), \tag{A.9}
$$

uniformly in $1 \le \ell \le L$.

Equation (A.9), and indeed (2.9), is linear in $\hat{c}_{\ell_1} - \theta(x_{\ell_1})$ for $1 \le \ell \le L$. It may be proved, using (A.9) and the approximations to $\hat{a}_{\ell_1 \ell_2}$ that we develop in the next paragraph, that (A.9) admits a solution satisfying

$$
\max_{1 \le \ell \le L} |\hat{c}_\ell - \theta(x_\ell)| = O_p[\{h^2 + (nh)^{-1/2}\} n^\delta] \tag{A.10}
$$

22

for all $\delta > 0$, and that for each sufficiently small $\delta > 0$, with probability converging to 1 as $n \to \infty$ there is a unique solution within radius $\{h^2 + (nh)^{-1/2}\} n^\delta$ of the first-mentioned solution. Therefore, with probability converging to 1 the matrix $\hat{A} = (\hat{a}_{\ell_1 \ell_2})$ is invertible, and so (A.10) holds, for all $\delta > 0$, for the unique solution of (2.9) that arises with probability converging to 1 as $n \to 0$.

Using (A.2) and (A.3) with $r = 0$ we deduce that

$$\hat{a}_{\ell_1 \ell_2} = \sum_{j=1}^{m} \sum_{k=1}^{m} \hat{\sigma}^{jk} E\{U_{\ell_1 \ell_2}(j, k, 0)\} + o_p(h^3 + \delta_{\ell_1 \ell_2} h^{5/2}),$$

uniformly in $1 \leq \ell_1, \ell_2 \leq L$. Now,

$$\begin{aligned}
E\{U_{\ell\ell}(j, j, 0)\} &= P(X_{ij} \in \mathcal{B}_\ell) = h\, f_j(x_\ell) + o(h), \\
E\{U_{\ell_1 \ell_2}(j, k, 0)\} &= P(X_{ij} \in \mathcal{B}_{\ell_1}, X_{ik} \in \mathcal{B}_{\ell_2}) = h^2\, f_{jk}(x_{\ell_1}, x_{\ell_2}) + O(h^3),
\end{aligned}$$

the first formula holding uniformly in j and ℓ and the second uniformly in $j \neq k$ and $1 \leq \ell_1, \ell_2 \leq L$. Therefore,

$$\hat{a}_{\ell_1 \ell_2} = \delta_{\ell_1 \ell_2}\, h\, \eta(x_{\ell_1}) + h^2\, \zeta(x_{\ell_1}, x_{\ell_2}) + O_p(h^3) + o_p(\delta_{\ell_1 \ell_2}\, h) = O_p(h^2 + h\, \delta_{\ell_1 \ell_2}),$$

where $\eta = \sum_j v^{jj} f_j$, $\zeta = \sum\sum_{j \neq k} v^{jk} f_{jk}$, and all remainders are of the stated orders uniformly in $1 \leq \ell_1, \ell_2 \leq L$. Substituting these formulae into the left-hand side of (A.9), and using (A.10), we deduce that

$$\{1 + o_p(1)\} \{\hat{c}_\ell - \theta(x_\ell)\} \eta(x_\ell) + \sum_{\ell_1 = 1}^{L} \{\hat{c}_{\ell_1} - \theta(x_{\ell_1})\} \zeta(x_{\ell_1}, x_\ell)$$
$$= h^2\, \xi(x_\ell) + n^{-1/2}\, h^{-1} \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk}\, V_{jkl} + o_p(h^2), \tag{A.11}$$

uniformly in $1 \leq \ell \leq L$.

Result (A.7) implies that

$$\max_{j,k,l} |V_{jkl}| = O_p\{(h\, \log n)^{1/2}\}. \tag{A.12}$$

Therefore, if the function b satisfies

$$\int_0^1 b(u)\, \zeta(u, v)\, du = \xi(v) - b(v)\, \eta(v)$$

then there exists a solution to (A.11) which satisfies both (A.10) and

$$\hat{c}_\ell = \theta(x_\ell) + h^2 b(x_\ell) + \frac{1 + o_p(1)}{n^{1/2}\, h\, \eta(x_\ell)} \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk}\, V_{jkl} + o_p(h^2), \tag{A.13}$$

where the remainders are of the stated orders uniformly in $1 \leq \ell \leq L$. This must be the solution which is unique, with probability converging to 1 as $n \to \infty$, and so (4.2) is proved. Combining (A.12) and (A.13) we deduce that (A.12) holds with the right-hand side replaced by $O_p\{h^2 + (nh)^{-1}(\log n)^{1/2}\}$, which establishes (4.3).

23

A.2 Proof of Theorem 4.2

The terms in $\theta''(x)$ and $b(x)$ in (4.4) and (4.5) follow directly from (4.2), noting the definition of $\widetilde{\theta}_{\mathrm{lin}}$. To derive the coefficients of $Z_n(x)$, observe that when $0 < x < 1$ the error-about-the-mean contribution to an expansion of $\widetilde{\theta}_{\mathrm{lin}}(x)$ equals

$$\frac{1 + o_p(1)}{n\,h^2\,\eta(x)} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk} \left\{ (x_{\ell+1} - x)\, I(X_{ik} \in \mathcal{B}_\ell) + (x - x_\ell)\, I(X_{ik} \in \mathcal{B}_{\ell+1}) \right\} \epsilon_{ij} \,,$$

the asymptotic variance of which equals $(nh)^{-1}\eta(x)^{-2}$ multiplied by

$$h^{-3} \sum_{j_1=1}^{m} \sum_{k_1=1}^{m} \sum_{j_2=1}^{m} \sum_{k_2=1}^{m} v^{j_1 k_1}\, v^{j_2 k_2}\, \sigma_{j_1 j_2} \delta_{k_1 k_2} \{ (x_{\ell+1} - x)^2\, P(X_{ik_1} \in \mathcal{B}_\ell)$$
$$+ (x - x_\ell)^2\, P(X_{ik_1} \in \mathcal{B}_{\ell+1}) \}$$
$$\sim h^{-2} \left\{ (x_{\ell+1} - x)^2 + (x - x_\ell)^2 \right\} \sum_{j=1}^{m} (V^{-1} \Sigma V^{-1})^{jj} f_j(x) \,.$$

This gives (4.4), and (4.5) may be derived similarly. Result (4.6) follows directly from (4.3).

A.3 Proof of Theorem 4.3

For simplicity we derive only (4.7). Our starting point is (A.9) and the approximations derived in the paragraph containing (A.10). From the latter approximations we see that if we define

$$\widehat{\eta}_n(x_\ell) \;=\; \sum_{j=1}^{m} \widehat{\sigma}^{jj}\, h^{-1}\, P(X_{ij} \in \mathcal{B}_\ell) \,, \tag{A.14}$$

$$\widehat{\zeta}_n(x_{\ell_1}, x_{\ell_2}) \;=\; \sum_{1 \le j,\, k \le m} \widehat{\sigma}^{jk}\, h^{-2}\, P(X_{ij} \in \mathcal{B}_{\ell_1}, X_{ik} \in \mathcal{B}_{\ell_2}) \tag{A.15}$$

then (A.11) holds in the form

$$\{\widehat{c}_\ell - \theta(x_\ell)\}\, \widehat{\eta}_n(x_\ell) + h \sum_{\ell_1=1}^{L} \{\widehat{c}_{\ell_1} - \theta(x_{\ell_1})\}\, \widehat{\zeta}_n(x_{\ell_1}, x_\ell)$$
$$= h^2\, \xi(x_\ell) + n^{-1/2}\, h^{-1} \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk}\, V_{jkl} + o_p(h^2) \,,$$

uniformly in $1 \le \ell \le L$. From this formula, and since $\widehat{\sigma}^{jk} = v^{jk} + O_p(h^{1/2})$ for $1 \le j, k \le m$, we see that if we define η_n and ζ_n as at (A.14) and (A.15) but with $\widehat{\sigma}^{jk}$ replaced by v^{jk}, then

$$\widehat{c}_\ell - \theta(x_\ell) \;=\; h \sum_{\ell_1=1}^{L} \{\widehat{c}_{\ell_1} - \theta(x_{\ell_1})\}\, \alpha_n(x_{\ell_1}, x_\ell)$$
$$+ h^2\, \beta_n(x_\ell) + n^{-1/2}\, h^{-1} \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk}\, W_{jkl} + o_p(h^2) \,, \tag{A.16}$$

24

uniformly in $1 \le \ell \le L$, where $\alpha_n(u,v) = \zeta_n(u,v)/\eta_n(v)$, $\beta_n = \xi/\eta_n$ and $W_{jk\ell} = V_{jk\ell}/\eta_n(x_\ell)$.

Now pass our local linear smoother through both sides of (A.16), obtaining:

$$\sum_{\ell=1}^{L} \{\widehat{c}_\ell - \theta(x_\ell)\}\, w_l(x) + h \sum_{\ell=1}^{L} \sum_{\ell_1=1}^{L} \{\widehat{c}_{\ell_1} - \theta(x_{\ell_1})\}\, \alpha_n(x_{\ell_1}, x_\ell)\, w_l(x) = T(x) + O_p(h^2)\,,$$

where

$$T(x) = n^{-1/2}\, h^{-1} \sum_{\ell=1}^{L} \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk}\, W_{jkl}\, w_\ell(x)$$

and $w_\ell(x)$ is as defined at (4.7). The variable

$$h \sum_{\ell=1}^{L} \sum_{\ell_1=1}^{L} \{\widehat{c}_{\ell_1} - \theta(x_{\ell_1})\}\, \alpha_n(x_{\ell_1}, x_\ell)\, w_l(x)$$

equals $O_p(h^2) + o_p\{(nh_1)^{-1/2}\}$, and $T(x)$ is asymptotically Normally distributed with zero mean and variance $(nh_1)^{-1}\kappa\tau_0(x)^2$. Therefore,

$$\widehat{\theta}(x) = \sum_{\ell=1}^{L} \widehat{c}_\ell\, w_\ell(x) = \sum_{\ell=1}^{L} \theta(x_\ell)\, w_\ell(x) + (nh_1)^{-1/2}\, \kappa^{1/2}\, \tau_0(x)\, Z_n(x) + O_p(h^2)\,,$$

where $Z_n(x)$ is asymptotically Normal N(0,1). Result (4.7) follows on noting that

$$\sum_{\ell=1}^{L} \theta(x_\ell)\, w_\ell(x) = \theta(x) + \tfrac{1}{2}\, \kappa_2\, \theta''(x)\, h_1^2 + o(h_1^2)\,.$$

A.4 Estimating Equations for Model at (1.2)

Using a working covariance matrix V rather than the true Σ, and writing $V^{-1} = (v^{jk})$, the negative loglikelihood becomes:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{\ell_1=1}^{L} \sum_{\ell_2=1}^{L} (Y_{ij} - \beta^{\mathrm{T}} Z_{ij} - c_{\ell_1})\, v^{jk}\, (Y_{ik} - \beta^{\mathrm{T}} Z_{ik} - c_{\ell_2})$$
$$\times I(X_{ij} \in \mathcal{B}_{\ell_1}, X_{ik} \in \mathcal{B}_{\ell_2})\,. \tag{A.17}$$

First eliminate β by differentiating with respect to β_t and equating to zero:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} \beta^{\mathrm{T}} Z_{ij}\, v^{jk}\, Z_{ikt} = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{\ell=1}^{L} (Y_{ij} - c_\ell)\, v^{jk}\, Z_{ikt}\, I(X_{ij} \in \mathcal{B}_\ell)\,.$$

Equivalently, $\beta^{\mathrm{T}} \widehat{A} = \widehat{s}^{\mathrm{T}}$, where $\widehat{A} = (\widehat{a}_{rs})$, $\widehat{s} = \widehat{s}(c) = (\widehat{s}_r)$, $\widehat{s}_r = \widehat{d}_r - \sum_\ell \widehat{e}_{r\ell} c_\ell$,

$$\widehat{a}_{rs} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk}\, Z_{ijr}\, Z_{iks}\,,$$

$$\widehat{d}_r = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} Y_{ij}\, v^{jk}\, Z_{ikr}\,,$$

$$\widehat{e}_{rl} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk}\, Z_{ikr}\, I(X_{ij} \in \mathcal{B}_\ell)\,.$$

25

Inverting to express β as a function of $c = (c_\ell)$ we deduce that

$$\beta = (\widehat{u}_1, \ldots, \widehat{u}_p)^{\mathrm{T}} + \sum_{\ell=1}^{L} c_\ell(\widehat{v}_{\ell 1}, \ldots, \widehat{v}_{\ell p}),\tag{A.18}$$

where $\widehat{u}_r$ and $\widehat{v}_{\ell r}$ are explicitly defined functions of the data alone.

Using (A.18) to substitute for β in (A.17), differentiating with respect to c_ℓ, and equating to zero, we obtain:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{\ell_1=1}^{L} \sum_{\ell_2=1}^{L} \left(\widetilde{Y}_{ij} - \sum_{\ell_3=1}^{L} c_{\ell_3} \widetilde{Z}_{ij\ell_3} - c_{\ell_1}\right) v^{jk} \left\{\widetilde{Z}_{ikl} + I(\ell_2 = \ell)\right\}$$
$$\times I(X_{ij} \in \mathcal{B}_{\ell_1}, X_{ik} \in \mathcal{B}_{\ell_2}) = 0,$$

where $\widetilde{Y}_{ij} = Y_{ij} - \sum_r \widehat{u}_r Z_{ijr}$, $\widetilde{Z}_{ij\ell} = \sum_r \widehat{v}_{\ell r} Z_{ijr}$. Therefore $\widehat{c}^{\mathrm{T}}\widehat{Q} = \widehat{q}^{\mathrm{T}}$, where $\widehat{c} = (\widehat{c}_1, \ldots, \widehat{c}_L)^{\mathrm{T}}$, $\widehat{Q} = (\widehat{q}_{rs})$, $\widehat{q} = (\widehat{q}_r)$,

$$\widehat{q}_{rs} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk} \left\{\widetilde{Z}_{ijr} + I(X_{ij} \in \mathcal{B}_r)\right\}\left\{\widetilde{Z}_{iks} + I(X_{ik} \in \mathcal{B}_s)\right\},$$

$$\widehat{q}_r = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} v^{jk} \widetilde{Y}_{ij}\left\{\widetilde{Z}_{ikr} + I(X_{ij} \in \mathcal{B}_r)\right\}.$$

This give us an explicit formula for $\widehat{c}$ in terms of the data alone. Substituting into (A.18) we obtain an explicit formula for $\widehat{\beta}$ in terms of the data alone.

A.5 Estimating Equations for Model at (1.3)

The negative loglikelihood may be written as

$$\sum_{i=1}^{n} \sum_{\ell_1=1}^{L} \cdots \sum_{\ell_m=1}^{L} \left(Y_i - \sum_{j=1}^{m} \beta^{j-1} c_{\ell_j}\right)^2 I(X_{ij} \in \mathcal{B}_{\ell_j}, 1 \le j \le m).\tag{A.19}$$

It is simplest to eliminate $c = (c_\ell)$ first, by expressing c as a function of β. To this end, differentiate (A.3) with respect to c_ℓ and equate to zero, obtaining:

$$\sum_{i=1}^{n} \sum_{\ell_1=1}^{L} \cdots \sum_{\ell_m=1}^{L} \sum_{j=1}^{m} \sum_{k=1}^{m} \beta^{j+k-2} c_{\ell_j} I(\ell_k = \ell; X_{ij} \in \mathcal{B}_{\ell_j}, 1 \le j \le m)$$
$$= \sum_{i=1}^{n} \sum_{\ell_1=1}^{L} \cdots \sum_{\ell_m=1}^{L} \sum_{k=1}^{m} Y_i \beta^{k-1} I(\ell_k = \ell; X_{ij} \in \mathcal{B}_{\ell_j}, 1 \le j \le m).\tag{A.20}$$

The left-hand side may be simplified to:

$$\sum_{r=1}^{L} c_r \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} \beta^{j+k-2} I(X_{ij} \in \mathcal{B}_r, X_{ik} \in \mathcal{B}_\ell),$$

26

and the right-hand side to

$$\sum_{k=1}^{m} \beta^{k-1} \sum_{i=1}^{n} Y_i \, I(X_{ik} \in \mathcal{B}_\ell) \, .$$

Therefore equation (A.20), defining $c = \widehat{c} = (\widehat{c}_\ell)$, may equivalently be written as $\widehat{c}^{\mathrm{T}} \widehat{A} = \widehat{s}^{\mathrm{T}}$, where $\widehat{A} = \widehat{A}(\beta) = (\widehat{a}_{rs})$, $\widehat{s} = \widehat{s}(\beta) = (\widehat{s}_r)$,

$$\begin{aligned}
\widehat{a}_{rs} &= \sum_{j=1}^{m} \sum_{k=1}^{m} \beta^{j+k-2} \frac{1}{n} \sum_{i=1}^{n} I(X_{ij} \in \mathcal{B}_r, X_{ik} \in \mathcal{B}_s) \, , \\
\widehat{s}_r &= \sum_{j=1}^{m} \beta^{j-1} \frac{1}{n} \sum_{i=1}^{n} Y_i \, I(X_{ij} \in \mathcal{B}_r) \, .
\end{aligned}$$

Thus we have an explicit formula for $\widehat{c}$ as a function of β and of the data. Substituting into (A.19), and minimizing over the single unknown β by either grid search or a Newton-Raphson procedure, is elementary. Having computed $\beta = \widehat{\beta}$ we now calculate $\widehat{A} = \widehat{A}(\widehat{\beta})$ and $\widehat{s} = \widehat{s}(\beta)$ as functions of the data alone, and finally, compute $\widehat{c} = \widehat{s}^{\mathrm{T}} \widehat{A}^{-1}$.

	$\sin(4x-2)$	$\exp(4x-2)$	$\sin(2(4x-2))$	$g_1(x)$	$g_2(x)$	$g_3(x)$
exchangeable correlation structure $\rho = 0.6$						
LIw	0.9273	0.8796	1.0577	1.0039	1.1362	1.2970
LIunw	0.7426	0.6953	0.8145	0.7559	0.8380	0.8614
2stkw(1)	1.3354	1.2683	1.3497	1.3289	1.1835	1.0697
2stkunw(1)	1.0313	0.9309	0.9970	1.0093	0.9269	0.8065
2stkw(2)	1.1527	1.1400	1.2308	1.2231	1.1929	1.6403
2stkunw(2)	0.9192	0.8626	0.8938	0.9152	0.9273	1.2750
2stkw(3)	1.3020	1.2795	1.3424	1.3264	1.2209	1.2108
2stkunw(3)	0.9258	0.9363	0.9268	0.9568	0.9408	0.9431
autoregressive correlation structure $\rho = 0.6$						
LIw	0.9023	0.8733	1.0257	0.9908	1.1300	1.2662
LIunw	0.7258	0.6941	0.8231	0.7951	0.8294	0.8614
2stkw(1)	1.3190	1.2521	1.3152	1.2705	1.1747	1.0517
2stkunw(1)	1.0247	0.9425	0.9858	0.9523	0.9370	0.8075
2stkw(2)	1.1439	1.1316	1.2140	1.2133	1.1907	1.3628
2stkunw(2)	0.9185	0.8597	0.8925	0.9243	0.9384	0.9725
2stkw(3)	1.2741	1.2696	1.2953	1.3089	1.9073	1.1896
2stkunw(3)	0.9270	0.9410	0.9330	0.9561	1.3933	0.9462
unstructured correlation structure where $\rho_{12} = \rho_{13} = 0.8$ and $\rho_{23} = 0.5$						
LIw	1.3051	1.2987	1.5587	1.4845	1.6703	2.3063
LIunw	0.7509	0.7057	0.8081	0.7581	0.8096	1.1179
2stkw(1)	1.8523	1.7844	1.9588	1.8879	1.6470	2.2741
2stkunw(1)	0.9950	0.9309	0.9962	1.0132	0.9351	1.2912
2stkw(2)	1.5607	1.6080	1.7778	1.7702	1.7644	1.5428
2stkunw(2)	0.9319	0.8730	0.8859	0.9003	0.9320	0.9185
2stkw(3)	1.8031	1.7932	1.9217	1.8728	1.5783	1.5319
2stkunw(3)	0.9206	0.9304	0.9390	0.9488	0.9508	0.9456

Table 1: Results of 200 simulations with $n = 1000$ clusters and $m = 3$ observations per cluster. Computed are the mean squared error (MSE) efficiencies of the various estimators relative to the working independence kernel method discussed in the text. The correlation structures were autoregressive with correlation $\rho = 0.6$, exchangeable with common correlation $\rho = 0.6$, and nearly singular with the correlations between the first and second observations and the second and third being 0.80, while the correlation between the first and third observations was 0.50. In this table, "LIw" is the linear interpolant method (2.5)–(2.6) with weighting and "LIunw" is the linear interpolant method without weighting: both used crossvalidation to estimate the number of bins. Also, "2stkw" is the two–stage kernel method with weighting, see definition above (2.7) and "2stkunw" is the two–stage kernel method without weighting. For these two–stage methods, we used three methods to select number of bins and bandwidths: (1) the number of bins selected by CV and the bandwidth selected by DPI bandwidth selection method, (2) the number of bins and the bandwidth selected by CV simultaneously (3) use the number of bins selected by CV with the bandwidth selected by DPI bandwidth selection method for the working independence kernel estimator. The functions are $g_1(x) = \sqrt{x(1-x)}\sin\{2\pi(1+2^{-3/5})/(x+2^{-3/5})\}$; $g_2(x) = \sqrt{x(1-x)}\sin\{2\pi(1+2^{-7/5})/(x+2^{-7/5})\}$; and $g_3(x) = \sin(8x-4) + 2\exp\{-256(x-.5)^2\}$.