University of Wollongong
Research Online

Centre for Statistical & Survey Methodology Working Paper Series

Faculty of Engineering and Information Sciences

2009

Marginal longitudinal semiparametric regression via penalized splines

Mohammad Alkadiri
ma981@uow.edu.au

R. J. Carroll
Texas A&M University

M. P. Wand
University of Wollongong, mwand@uow.edu.au

Follow this and additional works at: https://ro.uow.edu.au/cssmwp

Recommended Citation

Alkadiri, Mohammad; Carroll, R. J.; and Wand, M. P., Marginal longitudinal semiparametric regression via penalized splines, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 05-09, 2009, 13p.
https://ro.uow.edu.au/cssmwp/25

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

05-09

Marginal Longitudinal Semiparametric Regression via Penalized Splines.

Alkadiri, M., Carroll, R.J. and Wand, M.P.

Copyright © 2008 by the Centre for Statistical & Survey Methodology, UOW. Work in progress, no part of this paper may be reproduced without permission from the Centre.

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Marginal longitudinal semiparametric regression via penalized splines

BY M. ALKADIRI[1], R.J. CARROLL[2] AND M. P. WAND[1]

[1] *Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, New South Wales, Australia*

[2] *Department of Statistics, Texas A&M University, College Station, Texas, USA*

16th July, 2009

ABSTRACT

We study the marginal longitudinal nonparametric regression problem and some of its semiparametric extensions. We point out that, while several elaborate proposals for efficient estimation have been proposed, a relative simple and straightforward one, based on penalized splines, has not. After describing our approach we then explain how Gibbs sampling and the `BUGS` software can be used to achieve quick and effective implementation. Illustrations are provided for nonparametric regression and additive models.

Keywords: Additive models; Best prediction; Maximum likelihood; Gibbs sampling; Nonparametric regression; Restricted maximum likelihood; Varying coefficient models.

1 Introduction

The past decade has seen a great deal of interest and activity in nonparametric regression for longitudinal data. A prominent component of this research is the *marginal longitudinal nonparametric regression* problem in which the covariance matrix of the responses for each subject is not modelled conditionally, and instead is an unspecified parameter to be estimated.

Ruppert, Wand & Carroll (2009; Section 3.9) provide a summary of research on this problem up until about 2008. Whilst Zeger & Diggle (1994) is an early reference for marginal longitudinal nonparametric regression, the area started to heat up in response to Lin & Carroll (2001), where it was shown that ordinary kernel smoothers are more efficient if so-called working independence is assumed. This spawned a flurry of activity on the problem. Relevant references include: Welsh, Lin & Carroll (2002), Wang (2003), Linton, Mammen, Lin & Carroll (2003), Lin, Wang, Welsh & Carroll (2004), Carroll, Hall, Apanasovich & Lin (2004), Hu, Wang & Carroll (2004), Chen & Jin (2005), Wang, Carroll & Lin (2005), Lin & Carroll (2006) and Fan, Huang & Li (2007), Sun, Zhang & Tong (2007) and Fan & Wu (2008).

In this article we describe a relatively simple approach to the marginal longitudinal regression problem and its semiparametric extensions. Our approach is the natural one arising from the mixed model representation of penalized splines (e.g. Brumback, Ruppert & Wand, 1999; Ruppert, Wand & Carroll, 2003) with estimation and inference done using maximum likelihood and best prediction. There is also the option of adopting a Bayesian standpoint and calling upon Markov chain Monte Carlo to achieve approximate inference. An interesting aspect of our marginal longitudinal semiparametric regression models is that Gibbs sampling applies with draws from standard distributions. The Bayesian version of our models means that the `BUGS` inference engine (Lunn *et al.* 2000) can be used for fitting, and we provide some illustrative code.

The penalized spline/mixed model approach means that semiparametric extensions of the marginal longitudinal regression problem can be handled straightforwardly. We describe extensions to additive and varying coefficient models, although other extensions can be handled similarly.

Section 2 describes the penalized spline approach and identifies the mixed model structures required to handle marginal longitudinal semiparametric regression problems. In Section 3 we discuss fitting via maximum likelihood and best prediction. Section 4 describe Bayesian inference via Gibbs sampling and `BUGS`. Illustrations are provided in Section 5 and closing discussion is given in Section 6.

2 Marginal Longitudinal Nonparametric Regression and Extensions

For $1 \leq i \leq m$ subjects we observe $1 \leq j \leq n$ ($n \ll m$) scalar responses y_{ij} and predictors x_{ij}. Let $\boldsymbol{y}_i$ be the vector of responses for the ith subject and $\boldsymbol{x}_i$ be defined similarly. The covariance matrix of a random vector $\boldsymbol{v}$ is denoted by $\text{Cov}(\boldsymbol{v})$. The marginal longitudinal nonparametric regression model is then

$$E(y_{ij}) = f(x_{ij}), \qquad \text{Cov}\{\boldsymbol{y}_i|f(\boldsymbol{x}_i)\} = \boldsymbol{\Sigma}, \quad 1 \leq i \leq m,\ 1 \leq j \leq n \tag{1}$$

for some real-valued smooth function f and $n \times n$ covariance matrix $\boldsymbol{\Sigma}$. The notation $f(\boldsymbol{x}_i)$ means that the function f is applied element-wise to each of the entries of $\boldsymbol{x}_i$. We use $\text{Cov}\{\boldsymbol{y}_i|f(\boldsymbol{x}_i)\}$ rather than $\text{Cov}(\boldsymbol{y}_i)$ to allow for the possibility that $f(\boldsymbol{x}_i)$ is random according to the model, although this is not a requirement.

Figure 1 shows a simulated data set for model (1), with $m = 100$, $n = 10$,

$$f(x) = 1 + \tfrac{1}{2}\Phi((2x - 36)/5) \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 0.122 & 0.098 & 0.078 & 0.063 & 0.050 \\ 0.098 & 0.122 & 0.098 & 0.078 & 0.063 \\ 0.078 & 0.098 & 0.122 & 0.098 & 0.078 \\ 0.063 & 0.078 & 0.098 & 0.122 & 0.098 \\ 0.050 & 0.063 & 0.078 & 0.098 & 0.122 \end{bmatrix}, \tag{2}$$

where Φ is the standard normal distribution function. The main problem is efficient estimation of f from data such as that shown in Figure 1. Estimation of $\boldsymbol{\Sigma}$ may also be of interest.

Our approach to function estimation involves spline models for f of the form

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k z_k(x) \tag{3}$$

where $z_1, \ldots, z_K$ is a rich set of spline basis functions. A simple basis arises from setting $z_k(x) = (x - \kappa_k)_+$ where $\kappa_1, \ldots, \kappa_K$ is a dense set of knots placed over the range of the x_is. However, we recommend a smoother and more numerically stable choice for z_k, such as those described in Welham, Cullis, Kenward & Thompson (2007) and Wand & Ormerod (2008). The number of basis functions K has a minor effect on the efficacy of (3) and, for most signals arising in practice, $K = 25$ is sufficient. Li & Ruppert (2008) give some interesting asymptotics that provide support for this maxim.

To avoid over-fitting the spline coefficients u_k, $1 \leq k \leq K$, need to be penalized in some way. A convenient penalisation mechanism is to treat the u_k as a random sample from a distribution with mean zero and variance σ^2. This permits the following linear mixed model representation of (1) and (3):

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} \tag{4}$$

where

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_1 \\ \vdots & \vdots \\ \mathbf{1} & \boldsymbol{x}_m \end{bmatrix}, \quad \boldsymbol{Z} = \begin{bmatrix} z_1(\boldsymbol{x}_1) & \cdots & z_K(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ z_1(\boldsymbol{x}_m) & \cdots & z_K(\boldsymbol{x}_1) \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix}.$$

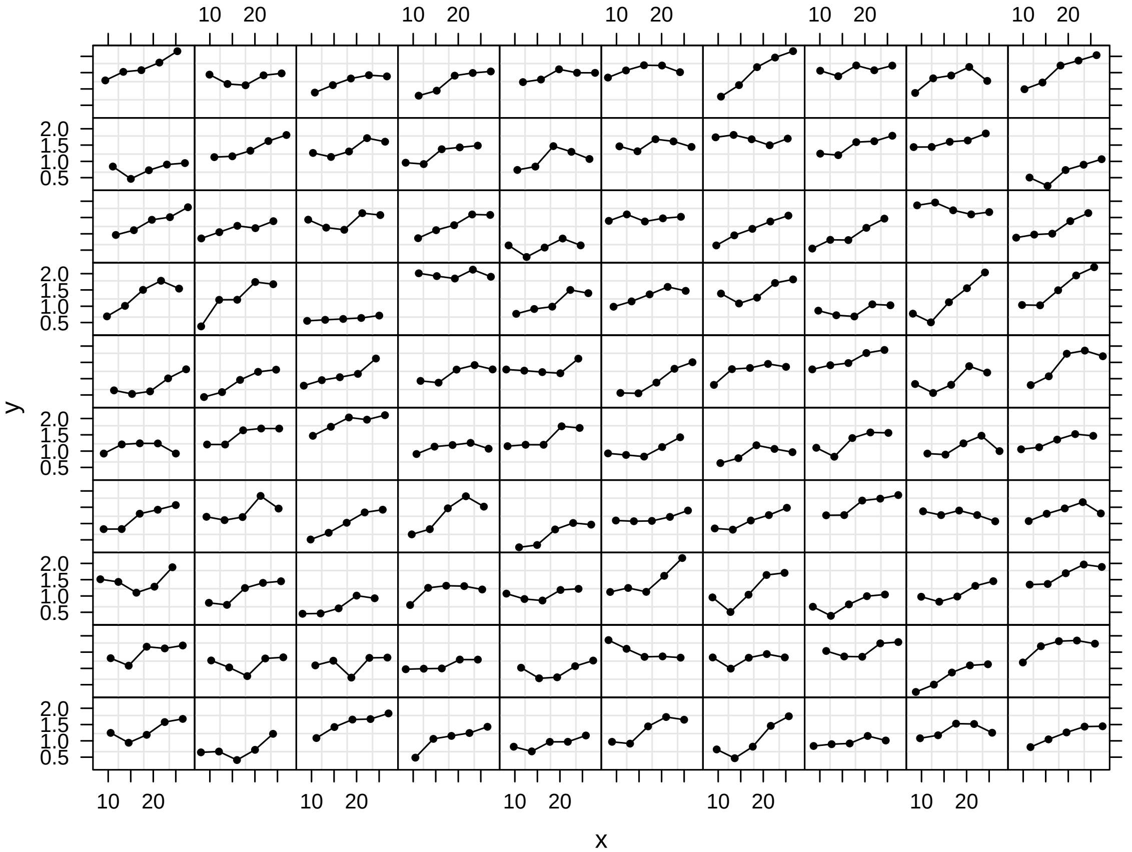

Figure 1: *A data set simulated from a version of the marginal longitudinal nonparametric regression model (1) with $m = 100$, $n = 5$ and f and $\boldsymbol{\Sigma}$ as described in the text.*

The random vectors on the right-hand side of (4) have mean zero and covariance matrix:

$$\mathrm{Cov}\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{bmatrix} = \begin{bmatrix} \sigma^2\boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{\Sigma} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{\Sigma} \end{bmatrix} = \begin{bmatrix} \sigma^2\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_m \otimes \boldsymbol{\Sigma} \end{bmatrix}.$$

For fixed values of σ^2 and $\boldsymbol{\Sigma}$ we can call upon best linear unbiased prediction (e.g. Robinson, 1991) to estimate $\boldsymbol{\beta}$ and $\boldsymbol{u}$ and, hence, the regression function f. In practice, though, both σ^2 and $\boldsymbol{\Sigma}$ need to be estimated and a convenient assumption for achieving this aim is

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \sigma^2\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_m \otimes \boldsymbol{\Sigma} \end{bmatrix}\right). \tag{5}$$

From now on we will assume that the *Gaussian* linear mixed model (4) and (5) is reasonably assumed. Sections 3 and 4 describe two approaches to fitting and inference. Before getting to that we describe some semiparametric extensions of (1).

2.1 Additive Models Extension

Suppose now that, corresponding to each y_{ij}, several predictor variables are available. There are a number of semiparametric regression extensions of (1) that could be considered. In this section we focus on the additive model extension. To keep the notation simple we restrict discussion to the situation where there are two continuous predictors with the jth measurement on subject i denoted by x_{1ij} and x_{2ij}. The *marginal longitudinal additive model* for such data is

$$E(y_{ij}) = \beta_0 + f_1(x_{1ij}) + f_2(x_{2ij}), \qquad \text{Cov}\{\boldsymbol{y}_i | f_1(\boldsymbol{x}_{1i}), f_2(\boldsymbol{x}_{2i})\} = \boldsymbol{\Sigma}, \quad 1 \le i \le m,\ 1 \le j \le n \quad (6)$$

where f_1 and f_2 are smooth functions. If each of these is modelled as a penalized spline:

$$f_1(x_1) = \beta_{11} x_1 + \sum_{k=1}^{K_1} u_{1k} z_{1k}(x_1) \quad \text{and} \quad f_2(x_2) = \beta_{21} x_2 + \sum_{k=1}^{K_2} u_{2k} z_{2k}(x_2) \tag{7}$$

with coefficients independently subject to

$$u_{1k} \text{ i.i.d. } N(0, \sigma_1^2) \quad \text{and} \quad u_{2k} \text{ i.i.d. } N(0, \sigma_2^2)$$

then a Gaussian linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}$$

arises. The differences between this model and that of Section 2 are that the design matrices are now

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_{11} & \boldsymbol{x}_{21} \\ \vdots & \vdots & \vdots \\ \mathbf{1} & \boldsymbol{x}_{1m} & \boldsymbol{x}_{2m} \end{bmatrix}, \qquad \boldsymbol{Z} = \begin{bmatrix} z_{11}(\boldsymbol{x}_{11}) & \cdots & z_{1K_1}(\boldsymbol{x}_{11}) & z_{21}(\boldsymbol{x}_{21}) & \cdots & z_{2K_2}(\boldsymbol{x}_{21}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ z_{11}(\boldsymbol{x}_{1m}) & \cdots & z_{1K_1}(\boldsymbol{x}_{1m}) & z_{21}(\boldsymbol{x}_{2m}) & \cdots & z_{2K_2}(\boldsymbol{x}_{2m}) \end{bmatrix}$$

where $\boldsymbol{x}_{1i}$ is the $n \times 1$ vector containing the the x_{1ij} measurements and $\boldsymbol{x}_{2i}$ is defined similarly. The coefficient vectors are

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_{11} \\ \beta_{12} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix}$$

where $\boldsymbol{u}_1$ is the $K_1 \times 1$ vector containing the u_{1k} and $\boldsymbol{u}_2$ is defined similarly. The covariance matrix of the spline coefficients and errors is now

$$\begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_1^2 \boldsymbol{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_m \otimes \boldsymbol{\Sigma} \end{bmatrix} \right). \tag{8}$$

Fitting via maximum likelihood and best prediction is analogous to that described in Section 3. The main difference is that there are two variance parameters σ_1^2 and σ_2^2 (and in extensions to additive models with d smooth functions there will be d such variance components) as well as the error covariance matrix $\boldsymbol{\Sigma}$. Maximum likelihood fitting, described in Section 3, requires an expression for $\boldsymbol{V} \equiv \text{Cov}(\boldsymbol{y})$. For the current model, this matrix takes the form

$$\boldsymbol{V} = \boldsymbol{V}(\sigma_1^2, \sigma_2^2, \boldsymbol{\Sigma}) = \sigma_1^2 \boldsymbol{Z}_{[1]} \boldsymbol{Z}_{[1]}^T + \sigma_2^2 \boldsymbol{Z}_{[2]} \boldsymbol{Z}_{[2]}^T + \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}$$

where $\boldsymbol{Z}_{[1]}$ and $\boldsymbol{Z}_{[2]}$ correspond to the column-wise partitioning of $\boldsymbol{Z}$ according to the basis functions for f_1 and f_2 (i.e. $\boldsymbol{Z} = [\boldsymbol{Z}_{[1]}\ \boldsymbol{Z}_{[2]}]$).

Before closing this section we briefly mention that the model

$$E(y_{ij}) = \beta_0 + \beta_1 x_{1ij} + f_2(x_{2ij}), \qquad \text{Cov}(\boldsymbol{y}_i) = \boldsymbol{\Sigma}, \quad 1 \le i \le m,\ 1 \le j \le n \tag{9}$$

is a simpler type of additive model than (6) since it only has one smooth function component. This is a *bona fide* semiparametric regression model since the right-hand side has the effect of the x_{1ij}s modelled parametrically and the effect of the x_{2ij}s modelled nonparametrically. However, the linear mixed model attached with this model is on par with that treated in Section 2. In particular, the random component structure (5) applies to (9).

2.2 Varying Coefficient Models Extension

Another type of multiple-predictor semiparametric regression model is that involving varying coefficients. The simplest marginal longitudinal varying coefficient model is

$$E(y_{ij}) = f_0(s_{ij}) + f_1(s_{ij})\,x_{ij}, \qquad \text{Cov}\{\boldsymbol{y}_i|f_0(\boldsymbol{s}_i), f_1(\boldsymbol{s}_i)\} = \boldsymbol{\Sigma}, \quad 1 \le i \le m,\ 1 \le j \le n \tag{10}$$

where the s_{ij} are longitudinal measurements on a continuous predictor variable s and the x_{ij} are measurements on a second predictor x. The modifying effect of s on the linear relationship between $E(y)$ and x is modelled flexibly through the varying coefficients $f_0(s)$ and $f_1(s)$. Sun, Zhang & Tong (2007) paid particular attention to models of this type.

Interestingly, the Gaussian linear mixed model for fitting the varying coefficient model (10) takes the same form as that for fitting the additive model (6). In particular, the covariance matrix of the random effects and error vectors is exactly the same as that given at (8). The only difference is that the fixed effect matrices are now

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{s}_1 & \boldsymbol{x}_1 & \boldsymbol{s}_1 \odot \boldsymbol{x}_1 \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{1} & \boldsymbol{s}_m & \boldsymbol{x}_m & \boldsymbol{s}_m \odot \boldsymbol{x}_m \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \beta_0 \\ \beta_{01} \\ \beta_{10} \\ \beta_{11} \end{bmatrix}$$

whilst the design matrix for the random effects component is

$$\boldsymbol{Z} = \begin{bmatrix} z_1(\boldsymbol{s}_1) & \cdots & z_K(\boldsymbol{s}_1) & \boldsymbol{x}_1 \odot z_1(\boldsymbol{s}_1) & \cdots & \boldsymbol{x}_1 \odot z_K(\boldsymbol{s}_1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ z_1(\boldsymbol{s}_m) & \cdots & z_K(\boldsymbol{s}_m) & \boldsymbol{x}_m \odot z_1(\boldsymbol{s}_m) & \cdots & \boldsymbol{x}_m \odot z_K(\boldsymbol{s}_m) \end{bmatrix},$$

with $\boldsymbol{a} \odot \boldsymbol{b}$ denoting the element-wise product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$.

3 Maximum Likelihood Estimation and Best Prediction

Each of the marginal longitudinal semiparametric regression models in the previous section, and their extensions to d smooth functions, can be handled using the Gaussian linear mixed model

$$\boldsymbol{y}|\boldsymbol{u} \sim N(\boldsymbol{X\beta} + \boldsymbol{Zu}, \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}), \quad \boldsymbol{u} \sim N(\mathbf{0}, \underset{1\le\ell\le d}{\text{blockdiag}}(\sigma_\ell^2 \boldsymbol{I}_{K_\ell})). \tag{11}$$

Here K_ℓ corresponds to the number of spline basis functions used in the ℓth smooth function estimate. Let $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_d^2)$ be the vector of variance parameters. Then the log-likelihood of $\boldsymbol{y}$ under (11) is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\Sigma}) = -\tfrac{1}{2}\left\{n\log(2\pi) + \log|\boldsymbol{V}| + (\boldsymbol{y} - \boldsymbol{X\beta})^T \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})\right\} \tag{12}$$

where

$$\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma}) \equiv \text{Cov}(\boldsymbol{y}) = \sum_{\ell=1}^{d} \sigma_\ell^2 \boldsymbol{Z}_{[\ell]} \boldsymbol{Z}_{[\ell]}^T + \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}$$

and $[\boldsymbol{Z}_{[1]} \cdots \boldsymbol{Z}_{[d]}]$ is the partition of $\boldsymbol{Z}$ corresponding to the basis functions for each smooth function estimate.

For any fixed values of $\boldsymbol{\sigma}^2$ and $\boldsymbol{\Sigma}$ the fixed effects solution is

$$\widetilde{\boldsymbol{\beta}}(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma}) = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{y} \tag{13}$$

On substitution into (12) we obtain the *profile log-likelihood* for $(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma})$ as:

$$\ell_P(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma}) = -\tfrac{1}{2}\left[\log|\boldsymbol{V}| + \boldsymbol{y}^T \boldsymbol{V}^{-1}\{\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1}\}\boldsymbol{y}\right] - \tfrac{n}{2}\log(2\pi). \tag{14}$$

However, the *restricted log-likelihood* (Patterson & Thompson, 1971)

$$\ell_R(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma}) = \ell_P(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma}) - \tfrac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}| \tag{15}$$

is usually preferred since it accounts for estimation of the fixed effects vector $\boldsymbol{\beta}$. The maximizers of $\ell_R(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma})$ are often labelled the *restricted maximum likelihood* or *REML* estimates of $\boldsymbol{\sigma}^2$ and $\boldsymbol{\Sigma}$.

Likelihood-based estimation of the model parameters $\boldsymbol{\beta}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\Sigma}$ thus involves:

1. Obtain the REML estimates $\widehat{\boldsymbol{\sigma}}^2$ and $\widehat{\boldsymbol{\Sigma}}$ by maximising $\ell_R(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma})$.

2. Obtain the maximum likelihood estimate of $\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}(\widehat{\boldsymbol{\sigma}}^2, \widehat{\boldsymbol{\Sigma}})$ according to (13).

Step 1. is by far the more challenging since it involves multivariate numerical optimisation.

Lastly, there is the problem of estimating spline coefficients $\boldsymbol{u}$. Since $\boldsymbol{u}$ is random we cannot appeal to maximum likelihood and instead have to rely on best prediction:

$$\widetilde{\boldsymbol{u}}(\boldsymbol{\sigma}^2, \boldsymbol{\Sigma}) \equiv E(\boldsymbol{y}|\boldsymbol{u}) = \boldsymbol{G}_{\boldsymbol{\sigma}^2}\boldsymbol{Z}^T\boldsymbol{V}(\sigma^2, \boldsymbol{\Sigma})^{-1}\{\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}(\sigma^2, \boldsymbol{\Sigma})\}$$

where $\boldsymbol{G}_{\boldsymbol{\sigma}^2} = \text{blockdiag}_{1\le\ell\le d}(\sigma_\ell^2\boldsymbol{I}_{K_\ell})$. An appropriate estimator for $\boldsymbol{u}$ in this context is the empirical best predictor

$$\widehat{\boldsymbol{u}} = \boldsymbol{G}_{\widehat{\boldsymbol{\sigma}}^2}\boldsymbol{Z}^T\boldsymbol{V}(\widehat{\boldsymbol{\sigma}}^2, \widehat{\boldsymbol{\Sigma}})^{-1}\{\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}(\widehat{\boldsymbol{\sigma}}^2, \widehat{\boldsymbol{\Sigma}})\}.$$

It is straightforward to construct estimates of the regression function f at arbitrary locations $x \in \mathbb{R}$ using $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{u}}$.

Despite (11) being a relatively simple linear mixed model, we have not yet been successful in fitting it with standard mixed model software such as `lme()` (Pinheiro *et al.* 2008) in the `R` computing language (R Core Development Team, 2009). This led us to also consider the Bayesian inference version and implementation via Gibbs sampling, as the next section describes.

4 Bayesian Inference

An alternative inference strategy, which permits more direct implementation in standard software, involves working with a hierarchical Bayesian version of the Gaussian linear mixed model (11). This entails treating $\boldsymbol{\beta}$, $\boldsymbol{\sigma}^2$ and $\boldsymbol{\Sigma}$ as random and setting prior distributions for each of them. The most convenient choice, because of conjugacy properties, are priors of the form:

$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \boldsymbol{F}), \quad \sigma_\ell^2 \sim \text{Inverse-Gamma}(A_\ell, B_\ell) \quad \text{and} \quad \boldsymbol{\Sigma} \sim \text{Inverse-Wishart}(a, \boldsymbol{B}) \tag{16}$$

where A_ℓ, B_ℓ, $1 \le \ell \le d$, are positive constants and $\boldsymbol{F}$ and $\boldsymbol{B}$ both positive definite matrices. Throughout this section let $[x]$ denote the density function of x. Then the notation $\sigma^2 \sim$ Inverse-Gamma(A, B) means that

$$[\sigma^2] = \frac{B^A}{\Gamma(A)}(\sigma^2)^{-A-1}e^{-B/\sigma^2}, \quad \sigma^2, A, B > 0.$$

The notation $\boldsymbol{\Sigma} \sim$ Inverse-Wishart$(a, \boldsymbol{B})$, where $\boldsymbol{\Sigma}$ is $n \times n$, means that

$$[\boldsymbol{\Sigma}] = C_{n,a}^{-1}|\boldsymbol{B}|^{a/2}|\boldsymbol{\Sigma}|^{-(a+n+1)/2}\exp\{-\tfrac{1}{2}\text{tr}(\boldsymbol{B}\Sigma^{-1})\}, \quad a > 0,\ \boldsymbol{\Sigma}, \boldsymbol{B} \text{ both positive definite}$$

where $C_{n,a} \equiv 2^{an/2}\pi^{n(n-1)/4}\prod_{i=1}^n \Gamma(\frac{a+1-i}{2})$.

Bayesian inference is based on the posterior density functions:

$$[\boldsymbol{\beta}|\boldsymbol{y}], \quad [\boldsymbol{u}|\boldsymbol{y}] \quad \text{and} \quad [\boldsymbol{\Sigma}|\boldsymbol{y}]. \tag{17}$$

The probability calculus required to obtain each of these is unwieldy and, in practice, either analytic or Monte Carlo approximations need to be called upon. As shown in Section 4.1, the Markov Chain Monte Carlo method *Gibbs sampling* is straightforward to implement for the Bayesian version of (11) and and the priors (16) and, upon convergence, yields samples of arbitrary size from the posterior densities (17). The software package `BUGS` (Lunn *et al.* 2000) facilitates this approach to approximate Bayesian inference and illustrative code is given in Section 4.2.

A final, albeit important, aspect of this approach to fitting and inference is choice of the hyperparameters $\boldsymbol{F}$, A_ℓ, B_ℓ, a and $\boldsymbol{B}$. If the analyst has specific prior beliefs about the model parameters then there is the opportunity to choose the hyperparameters so that the prior densities reflect those beliefs. More often than not such prior beliefs are absent and vague priors should be used. Reasonable choices for the fixed effects and variance hyperparameters, assuming that the data have been suitably standardized, are:

$$\boldsymbol{F} = 10^8\boldsymbol{I} \quad \text{and} \quad A_\ell = B_\ell = 0.01. \tag{18}$$

Reasonable choices for the hyperparameters associated with $\boldsymbol{\Sigma}$ are

$$a = n \quad \text{and} \quad \boldsymbol{B} = 0.01\boldsymbol{I}_n. \tag{19}$$

4.1 Gibbs Sampling Scheme

The hierarchical Bayesian model specified by (11) and (16) can be fitted using a Gibbs sampling scheme with draws from standard distributions. We give the details here.

First, we note (e.g. Robert & Casella, 2004, p. 371) that Gibbs sampling requires successive draws from the full conditional distributions for each member of a particular partition of the parameters in the model. For the present model we use the partition:

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}, \sigma_1^2, \ldots, \sigma_d^2, \boldsymbol{\Sigma}.$$

As an example, the full conditional distribution for σ_1^2 is

$$\sigma_1^2 \Big| \boldsymbol{y}, \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}, \sigma_2^2, \ldots, \sigma_d^2, \boldsymbol{\Sigma}.$$

We denote this by '$\sigma_1^2|$ rest' for short. Let $\boldsymbol{C} \equiv [\boldsymbol{X}\ \boldsymbol{Z}]$ and, as before, let $\boldsymbol{G}_{\boldsymbol{\sigma}^2} \equiv \text{blockdiag}_{1\le\ell\le d}(\sigma_\ell^2 \boldsymbol{I}_{K_\ell})$. Then the required full conditionals for Gibbs sampling are:

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \Big| \text{rest} \sim N((\boldsymbol{C}^T(\boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{C} + \boldsymbol{G}_{\boldsymbol{\sigma}^2})^{-1}\boldsymbol{C}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}, (\boldsymbol{C}^T(\boldsymbol{I} \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{C} + \boldsymbol{G}_{\boldsymbol{\sigma}^2})^{-1}),$$

$$\sigma_\ell^2 | \text{ rest} \sim \text{Inverse-Gamma}(A_\ell + \tfrac{1}{2}K_\ell, B_\ell + \tfrac{1}{2}\|\boldsymbol{u}_\ell\|^2), \ 1 \le \ell \le d$$

and $$\boldsymbol{\Sigma} | \text{ rest} \sim \text{Inverse-Wishart}(a + m, \boldsymbol{B} + (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})^T).$$

Provided that a is an integer then the Inverse-Wishart draws for $\boldsymbol{\Sigma}$ can be achieved by setting

$$\boldsymbol{\Sigma} = \left(\sum_{i=1}^{a+m} \boldsymbol{v}_i \boldsymbol{v}_i^T \right)^{-1}$$

where the $\boldsymbol{v}_i$ are independent $N(\boldsymbol{0}, \{\boldsymbol{B} + (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zu})^T\}^{-1})$ random vectors.

Interestingly, the fact that $\boldsymbol{\Sigma}$ is unstructured means that Gibbs sampling is exact. This is not the case if $\boldsymbol{\Sigma}$ is structured (e.g. autoregressive) and more complicated Markov chain Monte Carlo schemes are then required.

4.2 Implementation in BUGS

The BUGS language supports implementation of our Bayesian marginal longitudinal semiparametric regression models. It is recommended that the spline basis functions be set up outside of BUGS. We do this in R and then call BUGS using the BRugs package (Ligges *et al.* 2007). We pass the regression data to BUGS using matrices. For example, the variable yMat is an $m \times n$ matrix with (i, j) entry containing y_{ij}. Our BUGS code for fitting the marginal longitudinal nonparametric regression model is:

```
model
{
   for (i in 1:m)
   {
      for (j in 1:n)
      {
         mu[i,j] <- beta0 + beta1*xMat[i,j] + inprod(u[],Z[(i-1)*n+j,])
      }
      yMat[i,1:n] ~ dmnorm(mu[i,],Omega[1:n,1:n])
   }
   for (k in 1:K)
   {
      u[k] ~ dnorm(0,tau)
   }
   beta0  ~ dnorm(0,1.0E-8) ; beta1  ~ dnorm(0,1.0E-8)
   tau ~ dgamma(0.01,0.01)
   Omega[1:n,1:n] ~ dwish(R[,],n)
   for (i in 1:n)
   {
      for (j in 1:n)
      {
         R[i,j] <- 0.01*equals(i,j)
      }
   }
   sigma <- 1/sqrt(tau)
   Sigma[1:n,1:n] <- inverse(Omega[,])
}
```

Note that BUGS uses precision matrices rather than covariance matrices in its multivariate normal distribution specification. Hence, the above code uses the variable Omega, corresponding to $\Omega = \mathbf{\Sigma}^{-1}$. Similarly, the precision parameter tau corresponds to $\tau = 1/\sigma^2$ where σ^2 is the spline penalisation variance component.

5 Illustrations

We tested out BUGS fitting of the three types of models presented in Section 2 on several sets of simulated data. The simulation aspect also allows for comparisons were done with the true functions and marginal covariance matrix that generated the data. We now present some of these results as illustration of the methodology and its good performance.

5.1 Illustration for Nonparametric Regression

We fitted the penalized spline model (11) to the data of Figure 1. The y_{ij} were generated according to (2). The x_{ij} are equally spaced but with the starting positions x_{i1} were generated uniformly from the interval $(8, 12)$. We used the diffuse priors given by (18) and (19). A burn-in period of 5000 was used, followed by 5000 iterations with a thinning factor of 5 – resulting in samples of size 1000 being retained for inference.

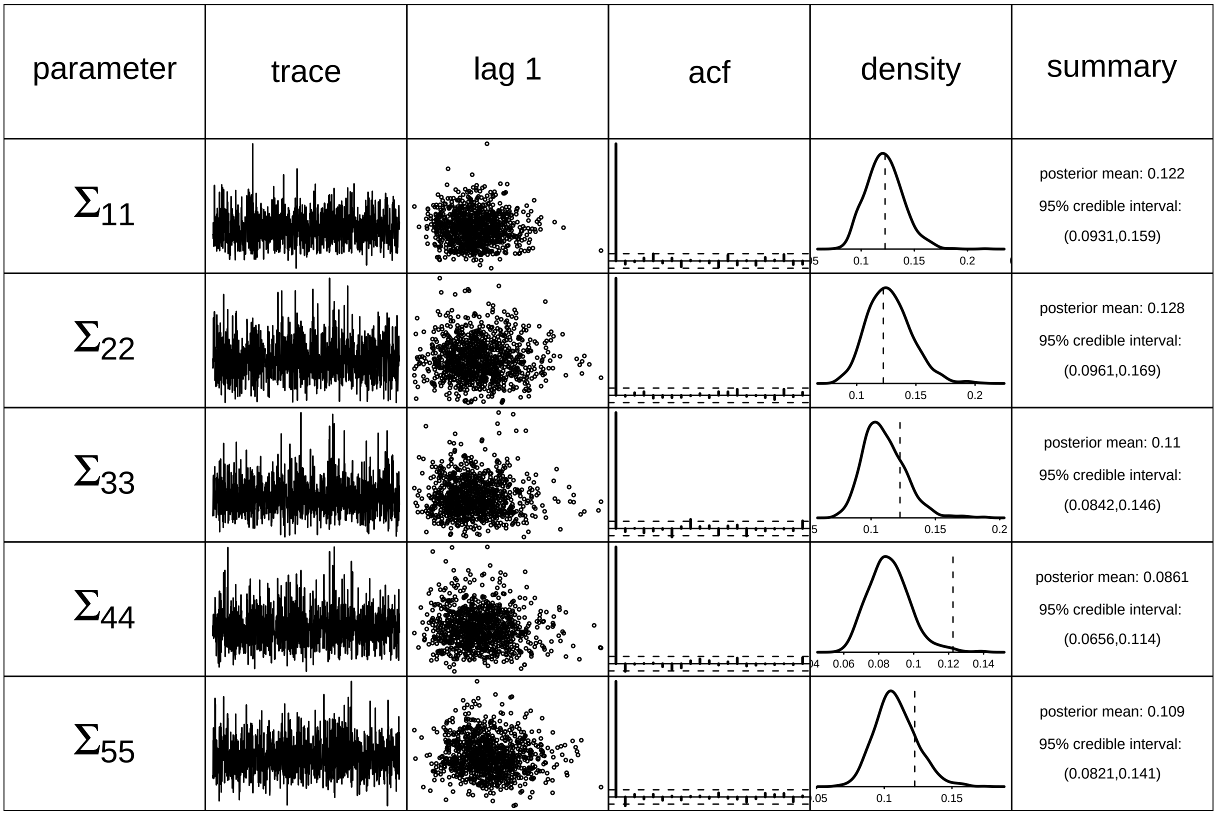

Figure 2: *Summary of MCMC-based inference for the diagonal entries of $\mathbf{\Sigma}$ in the fitted marginal longitudinal nonparametric regression model . The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates of posterior density and basic numerical summaries. True values of the parameters are shown as vertical dashed lines in the posterior density estimate.*

Figure 2 summarizes the `BUGS` output for the diagonal entries of $\mathbf{\Sigma}$. The chains mix quite well with no significant autocorrelation. In addition, the true values of Σ_{ii} are captured by the 95% credible intervals in four out of the five cases.

The results for the off-diagonal entries of $\mathbf{\Sigma}$ are summarized in Figure 3. All ten of the true values of Σ_{ij} are captured by the 95% credible intervals in four out of the five cases.

The Bayesian penalized spline estimate of f is shown in Figure 4, with and without the data. The thick solid curves correspond to the posterior means of 3 over a grid of xs. The dashed curves are corresponding 95% credible sets. The regression function from which the data were generated is shown for comparison. The gridwise posteriors are seen to cover the true f quite well.

The good results presented in this section are typical of the performances we observed over several runs, as well as different choices for f and $\mathbf{\Sigma}$. An interesting future project would be a large scale simulation study that compares this approach with existing methods.

5.2 Illustration for Additive Models

We simulated data according to the model

$$\begin{aligned} E(y_{ij}) &= \sin(2\pi(x_{1ij}^2 - 0.1)) + \sin(3\pi(0.05 - x_{2ij})), \\ \text{Cov}(\boldsymbol{y}_i) &= 0.36\mathbf{1}_5\mathbf{1}_5^T + 0.25\,\boldsymbol{I}_5 \end{aligned}$$

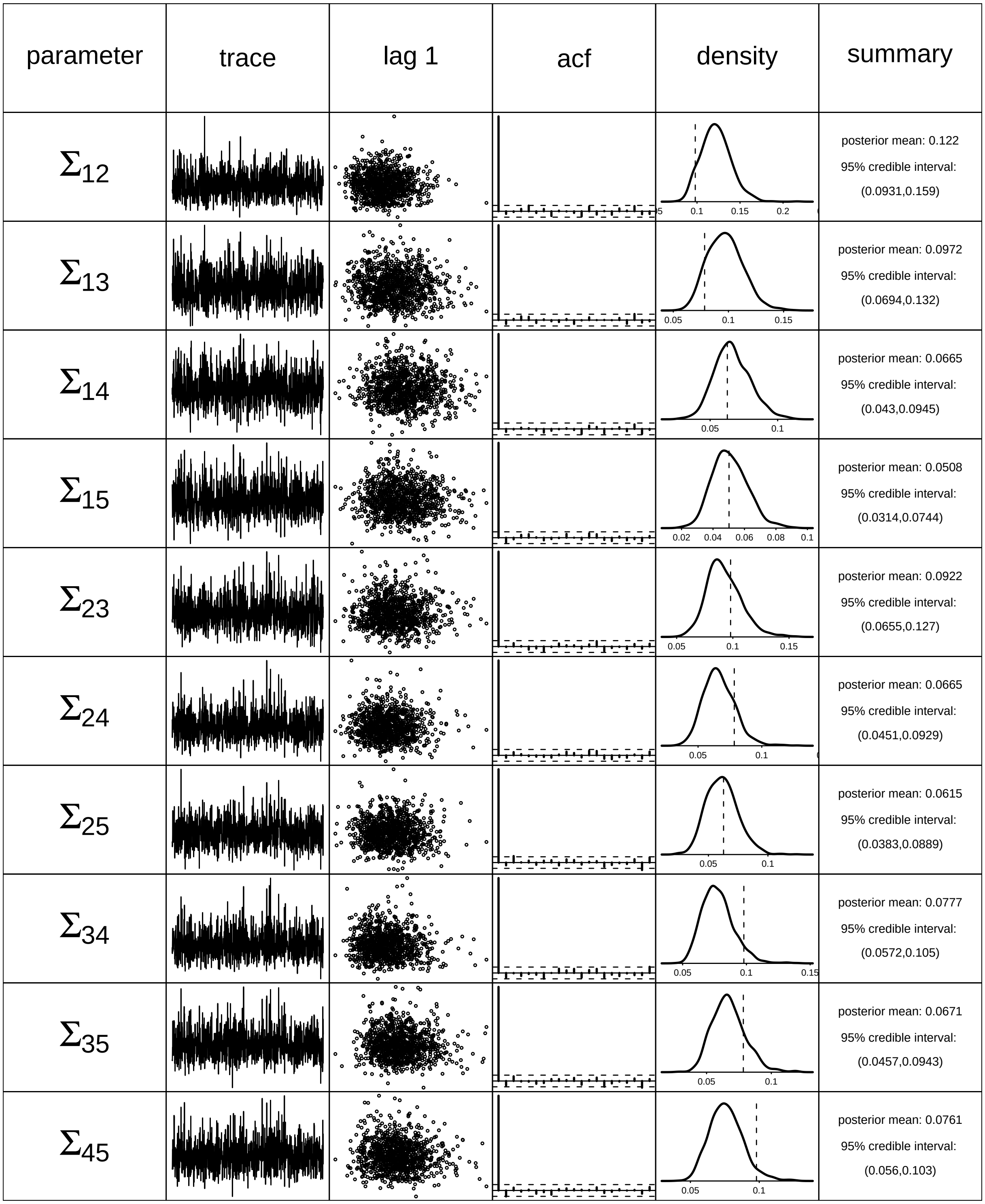

Figure 3: *Summary of MCMC-based inference for the off-diagonal entries of* $\mathbf{\Sigma}$ *in the fitted marginal longitudinal nonparametric regression model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates of posterior density and basic numerical summaries. True values of the parameters are shown as vertical dashed lines in the posterior density estimate.*

for $1 \le i \le 200$ and $1 \le j \le 5$. Here $\mathbf{1}_d$ denotes the $d \times 1$ vector of ones. For each i we generated the x_{1ij} to be

$$x_{1ij} = r_i + 1/n, \ j = 1, \ldots, n$$

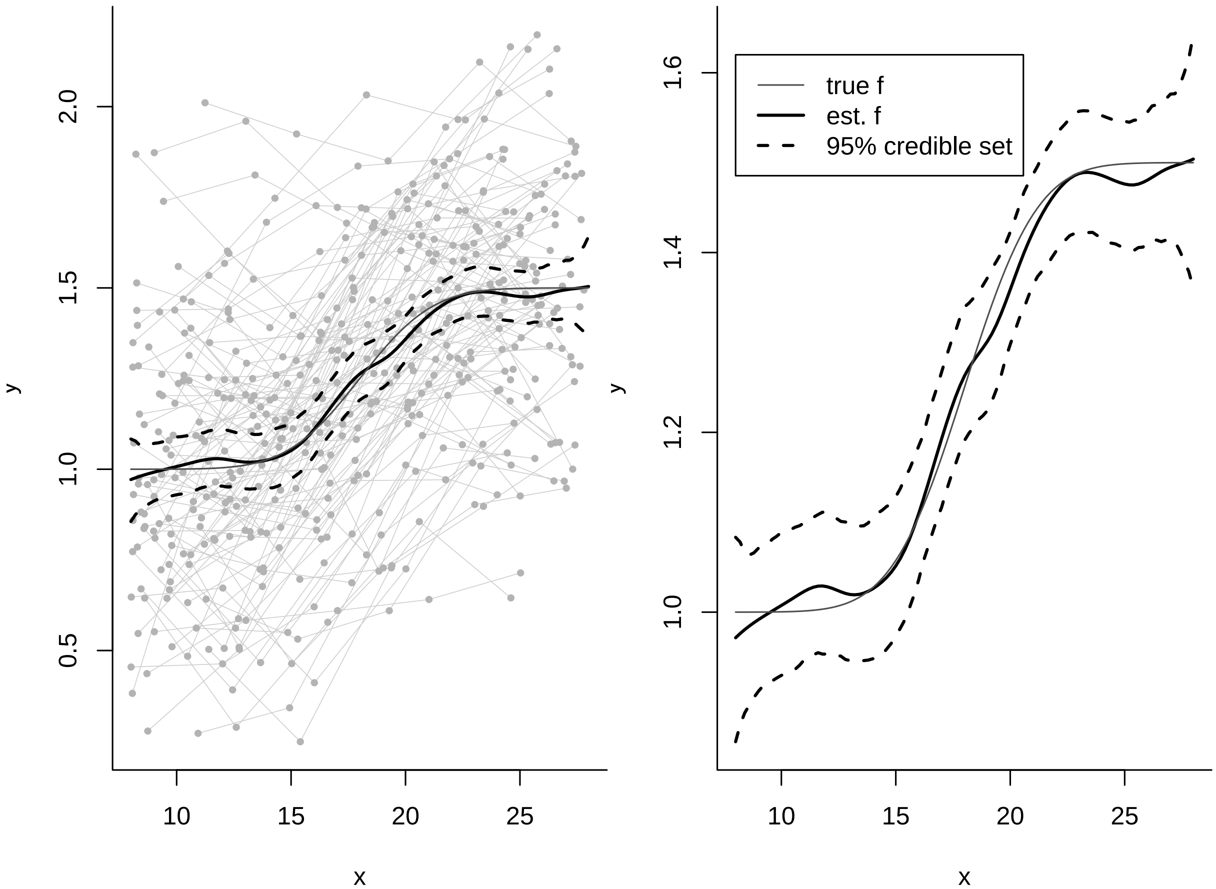

Figure 4: *Left panel: Estimated regression function f, together with the longitudinal data on which it is based. Dashed curves correspond to pointwise 95% credible sets. The true f is shown as a thin grey curve. Right panel: Estimated regression function f, with data omitted to allow better visualisation.*

where r_i is uniformly distributed on $(0, 1/n)$. An identical strategy was used for the x_{2ij} Even though the x_{1ij} and x_{2ij} were generated from a random process, they are considered fixed in the present analysis.

Figure 5 shows the posterior means of f_1 and f_2 and accompanying pointwise 95% credible sets. These answers were obtained via `BUGS`. Agreement with the true f_1 and f_2 is seen to be very good. The numerical summaries for the posterior of $\mathbf{\Sigma}$ are consistent with the truth from which the data were generated, although these are not shown because of space considerations.

6 Discussion

It is somewhat of a quirk that the mixed model-based penalized spline approach to marginal longitudinal nonparametric regression has not been explored in depth until now. Nevertheless, as we have illustrated in the previous section, it is a viable approach that is readily implemented in standard software. Another advantage of this approach is that complications such as missingness can be handled within the same likelihood-based or Bayesian framworks. It would be interesting to see if the asymptotic efficiency results established for other approaches (e.g. Wang, Carroll & Lin, 2005) also apply here.

Acknowledgments

Wand's research was partially supported by Australian Research Council Discovery Project DP0877055. Carroll's research was supported by a grant from the US National Cancer Insti-

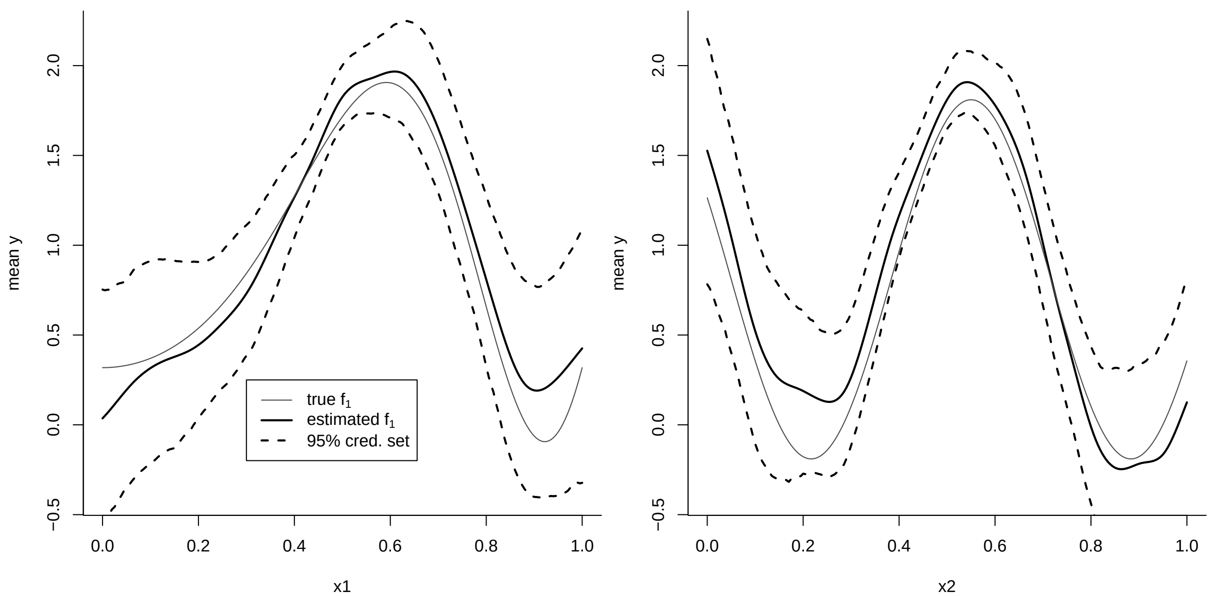

Figure 5: *Posterior mean estimates of f_1 and f_2. Dashed curves correspond to pointwise 95% credible sets. The true f_1 and f_2 are shown as thin grey curves. Vertical alignment is achieved by plotting $f_1(x_1) + f_2(\overline{x}_2)$ versus x_1 in the left panel, where $\overline{x}_2$ is the average of the x_{2ij}s. The right panel is $f_1(\overline{x}_1) + f_2(x_2)$ versus x_2.*

tute (CA57030) and by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology, Saudi Arabia.

References

Brumback, B.A., Ruppert, D. & Wand, M.P. (1999). Comment on paper by Shively, Kohn & Wood. *Journal of the American Statistical Association*, **94**, 794–797.

Carroll, R.J., Hall, P., Apanasovich, T.V. & Lin, X. (2004). Histospline method in nonparametric regression models with application to clustered/longitudinal data. *Statistica Sinica*, **14**, 649–674.

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

Fan, J., Huang, T. & Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association*, **102**, 632–641.

Fan, J. & Wu, Y. (2008). Semiparametric estimation of covariance matrixes for longitudinal data. *Journal of the American Statistical Association*, **103**, 1520–1533.

Hu, Z.H., Wang, N. & Carroll, R.J. (2004). Profile-kernel versus backfitting in the partially linear models for longitudinal/clustered data. *Biometrika*, **91**, 251–262.

Li, Y. & Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, **95**, 415–436.

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

Lin, X. & Carroll, R.J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society, Series B*, **68**, 68–88.

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

Linton, O.B., Mammen, E., Lin, X. & Carroll, R.J. (2003). Accounting for correlation in marginal longitudinal nonparametric regression. In *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*, Eds. D. Y. Lin and P. J. Heagerty, pp. 23-33, New York: Springer.

Ligges, U., Thomas, A., Spiegelhalter, D., Best, N., Lunn, D., Rice, K. & Sturtz, S. (2007). `BRugs 0.5`: Analysis of graphical models using MCMC techniques. R package.

Lunn, D.J., Thomas, A., Best, N. & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**, 325–337.

Patterson, H.D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & the R Core team. (2009). nlme 3.1: linear and nonlinear mixed effects models. R package. `www.R-project.org`.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. `www.R-project.org`.

Robert, C.P. & Casella, G. (2004). *Monte Carlo Statistical Methods, Second Edition.* New York: Springer-Verlag.

Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, **6**, 15–51.

Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Ruppert, D., Wand, M.P. & Carroll, R.J. (2009). Semiparametric regression during 2003-2007. *Journal of the American Statistical Association*, tentatively accepted.

Sun, Y., Zhang, W., Tong, H. (2007). Estimation of the covariance matrix of random effects in longitudinal studies. **35**, 2795–2814.

Wand, M.P. & Ormerod, J.T. (2008). On O'Sullivan penalised splines and semiparametric regression. *Australian and New Zealand Journal of Statistics*, **50**, 179–198.

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

Wang, N., Carroll, R.J. & Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, **100**, 147–157.

Welham, S.J., Cullis, B.R., Kenward, M.G. & Thompson, R. (2007). A comparison of mixed model splines for curve fitting. *Australian and New Zealand Journal of Statistics*, **49**, 1–23.

Welsh, A. H., Lin, X. & Carroll, R.J. (2002). Marginal longitudinal nonparametric regression: locality and efficiency of spline and kernel methods. *Journal of the American Statistical Association*, **97**, 482–493.

Zeger, S. & Diggle, P.J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, **50**, 689–699.