Electronic Journal of Statistics
Vol. 10 (2016) 527–549
ISSN: 1935-7524
DOI: 10.1214/16-EJS1113

Functional models for longitudinal data with covariate dependent smoothness

David K. Mensah[*], **David J. Nott**[†], **Linda S. L. Tan**

Department of Statistics and Applied Probability
National University of Singapore, Singapore 117546
e-mail: A0082808Y@nus.edu.sg; standj@nus.edu.sg; statsll@nus.edu.sg

and

Lucy Marshall[‡]

School of Civil and Environmental Engineering
University of New South Wales, Sydney
Australia
e-mail: lucy.marshall@unsw.edu.au

Abstract: This paper considers functional models for longitudinal data with subject and group specific trends modelled using Gaussian processes. Fitting Gaussian process regression models is a computationally challenging task, and various sparse approximations to Gaussian processes have been considered in the literature to ease the computational burden. This manuscript builds on a fast non-standard variational approximation which uses a sparse spectral representation and is able to treat uncertainty in the covariance function hyperparameters. This allows fast variational computational methods to be extended to models where there are many functions to be estimated and where there is a hierarchical model involving the covariance function parameters. The main goal of this paper is to implement this idea in the context of functional models for longitudinal data by allowing individual specific smoothness related to covariates for different subjects. Understanding the relationship of smoothness to individual specific covariates is of great interest in some applications. The methods are illustrated with simulated data and a dataset of streamflow curves generated by a rainfall runoff model, and compared with MCMC. It is also shown how these methods can be used to obtain good proposal distributions for MCMC analyses.

Keywords and phrases: Functional data, Gaussian processes, longitudinal data, variational Bayes.

Received July 2014.

1. Introduction

This paper considers functional models for longitudinal data where the subjects are divided into groups, with group specific and individual specific trends de-

[*]Research supported by a Singapore International Graduate Award (SINGA).
[†]Research supported by a Singapore Ministry of Education Academic Research Fund Tier 2 grant (R-155-000-143-112).
[‡]Research supported by an Australian Research Council Future Fellowship.

scribed by Gaussian processes. The computational challenges of dealing with Gaussian processes for regression are well known (see, e.g. Rasmussen and Williams, 2006). Here we consider a non-standard variational approximation developed in Tan et al. (2015), which relies on a sparse spectral representation.

Variational approximation methods try to convert problems of posterior integration into optimization problems. The basic idea is to consider some tractable class of densities for approximating the true posterior distribution, and then via an appropriate optimization method find the density which is best in that class, where best is usually in the sense of minimizing the Kullback-Leibler divergence. Further discussion and references to the literature are given in Section 3.

The approach of Tan et al. (2015) for applying variational methods to Gaussian process regression is unique as far as we know in that it allows covariance function hyperparameter uncertainty to be handled within a fast deterministic variational inference scheme. This allows hierarchical models involving the covariance hyperparameters to be handled within this framework. The main contribution of this paper is to exploit this feature in the longitudinal setting to develop fast inferential methods in a functional model which allows individual specific smoothness of trends to be related to covariates. Modelling covariate dependent individual specific smoothness is of great interest in some applications.

There is a large literature on semiparametric methods for longitudinal data that allow flexible subject specific trends – see, for example, Zeger and Diggle (1994) for one early contribution which uses Gaussian processes within a mixed model framework. However, as noted recently by Zhu and Dunson (2012), there is a much smaller literature which deals with the study of dynamic properties of longitudinal models in the functional setting. Wang et al. (2008) propose a second order differential equation model suitable for capturing the dynamics of online auctions. Müller and Yao (2010) represent functional data using stochastic ordinary differential equations with time varying coefficients and a smooth drift process. Their empirical approach to examining the underlying dynamics avoids the need to specify any parametric form for a differential equation describing the dynamics. Reithinger et al. (2008) use a boosting algorithm to fit a semiparametric mixed model which accounts for dependence between functional observations and can handle irregularly spaced observations. Zhu, Taylor and Song (2011) consider modelling of the rate function (the derivative of the mean with respect to time) and its dependence on covariates, with the rate functions being modelled by stochastic differential equations. Zhu and Dunson (2012) use hierarchical stochastic differential equations for functional data allowing volatility to depend on covariates. By volatility they mean the conditional variance of changes in the trajectory over an infinitesimal interval. Goldsmith, Wand and Crainiceanu (2011) consider functional regression models including models for longitudinal data involving parametric random effects. They use fast variational Bayes methods for inference similar to those considered here but their work does not focus on dynamic properties of the trends and how these might vary between subjects in relation to covariates.

Here we consider subjects divided into groups, with flexible trends at the level of the group and individual, with stationary Gaussian process priors for

the trend functions. We consider models for individual level covariance function hyperparmaters that relate them to covariates. Since the covariance function hyperparameters are naturally thought of as relating to the variance of the process derivative, our approach relates the dynamic behaviour of trends for individual subjects to covariates. Gaussian processes are an increasingly popular approach to the modelling of functional data. Shi, Murray-Smith and Titterington (2005) consider modelling functional data using a mixture of Gaussian processes approach and use Hamiltonian Monte Carlo methods for computation. Shi et al. (2012) consider a semiparametric approach which combines a parametric mixed effects model with a Gaussian process functional regression model. Wang and Khardon (2012) consider a Gaussian process mixed effects model with group specific trends and where the group membership is unobserved. Their work is similar in its objectives to ours in considering sparse approximations to Gaussian processes and variational ideas to speed up computations. However, the focus of our work is different since interest centres on the case where group membership is observed and there is a model which allows individual specific smoothness related to covariates. Shi and Choi (2011) is a recent monograph length treatment of the Gaussian process approach to functional data analysis.

In Section 2 our functional model is described. Section 3 briefly describes variational Bayes methods and the non-standard variational Bayes approach that we use to handle the Gaussian process elements in our hierarchical model. Application of these ideas to our longitudinal model is given in Section 4, some examples are given in Section 5 and Section 6 concludes.

2. Functional model for longitudinal data

We consider grouped longitudinal data in which a response $y_i = (y_{i1}, \ldots, y_{in_i})^T$ is observed for individual i. The response y_{ij} is associated with a time t_{ij} and the ith individual is in group $g(i)$, where there are k groups, $g(i) \in \{1, \ldots, k\}$. Similar to Zhu and Dunson (2012), we consider a model

$$y_{ij} = G_{g(i)}(t_{ij}) + S_i(t_{ij}) + \epsilon_{ij}, \tag{1}$$

where $G_g(t)$, $g = 1, \ldots, k$, are a collection of group specific trends and $S_i(t)$, $i = 1, \ldots, n$, are individual specific trends. The errors ϵ_{ij} are independent, $N(0, \sigma_\epsilon^2)$.

We consider Gaussian process priors (Rasmussen and Williams, 2006) on the functional terms,

$$G_g(t) \sim \mathrm{GP}(0, \kappa_g(t, t')) \text{ and } S_i(t) \sim \mathrm{GP}(0, \tau_i(t, t')),$$

where $\mathrm{GP}(\mu(t), C(t, t'))$ denotes the Gaussian process with mean function $\mu(t)$ and covariance function $C(t, t')$. The covariance functions $\kappa_g(\cdot, \cdot)$ and $\tau_i(\cdot, \cdot)$ are chosen to have a stationary parametric form. In what follows, we use the Gaussian covariance function,

$$r(t - t'; \sigma^2, \theta^2) = \sigma^2 \exp\left(-\theta^2 |t - t'|^2\right), \tag{2}$$

where $\sigma^2 > 0$ is the variance and θ is a spatial dependence parameter. Note that following standard results on Gaussian processes, the derivative of a Gaussian process with covariance function (2) exists in mean square and it is a Gaussian process with covariance function $r^{''}(h;\sigma^2,\theta^2)$. The process derivative has variance $\sigma^2\theta^4$ which shows how the spatial dependence parameter relates to the dynamic properties of the trends. We let

$$\kappa_g(t,t') = r(t-t';\sigma^2_{\kappa g},\theta^2_g), \quad g = 1,\ldots,k,$$

and

$$\tau_i(t,t') = r(t-t';\sigma^2_{\tau i},\lambda^2_i), \quad i = 1,\ldots n.$$

We choose half-t priors for σ_ϵ, $\sigma_{\kappa g}$ and $\sigma_{\tau i}$ such that $\sigma_\epsilon \sim$ Half-t(A_ϵ, b_ϵ), $\sigma_{\kappa g} \sim$ Half-t$(A_{\kappa_g}, b_{\kappa_g})$ and $\sigma_{\tau i} \sim$ Half-t(A_{τ_i}, b_{τ_i}), where all hyperparameters are known and Half-t(A,b) denotes the half-t distribution with scale A and degrees of freedom b. We consider normal priors on the spatial dependence parameters, $\theta_g \sim N(\mu_{\theta 0}, \sigma^2_{\theta 0})$, where $\mu_{\theta 0}$ and $\sigma^2_{\theta 0}$ are known, and $\lambda_i \sim N(\upsilon_i^T\beta, \sigma^2_\lambda)$ where $\upsilon_i = (\upsilon_{i1},\ldots,\upsilon_{ir})^T$ is a set of individual specific covariates, β is a vector of coefficients given a normal prior $\beta \sim N(\mu_{\beta 0}, \Sigma_{\beta 0})$ where $\mu_{\beta 0}$ and $\Sigma_{\beta 0}$ are known, and σ^2_λ is given an inverse gamma prior, $\sigma^2_\lambda \sim IG(a_\lambda, b_\lambda)$ where a_λ and b_λ are known. A half-t distribution can expressed as a scale mixture of inverse gamma distributions (Wand et al., 2011) and writing u_ϵ, u_{τ_i} and u_{κ_g} for appropriate auxiliary variables, the half-t priors can be equivalently expressed as $\sigma^2_\epsilon|u_\epsilon \sim IG(b_\epsilon/2, b_\epsilon/u_\epsilon)$, $\sigma^2_{\tau_i}|u_{\tau_i} \sim IG(b_{\tau_i}/2, b_{\tau_i}/u_{\tau_i})$, $\sigma^2_{\kappa_g}|u_{\kappa_g} \sim IG(b_{\kappa_g}/2, b_{\kappa_g}/u_{\kappa_g})$, $u_\epsilon \sim IG(1/2, 1/A^2_\epsilon)$, $u_{\tau_i} \sim IG(1/2, 1/A^2_{\tau_i})$ and $u_{\kappa_g} \sim IG(1/2, 1/A^2_{\kappa_g})$. We use these alternative forms of the Half-t priors to simplify our computations.

3. Variational Bayes for Gaussian processes

We give a brief introduction to variational Bayes methods before applying them to our functional model. With data y, a likelihood $p(y|\xi)$ with parameter ξ, and $p(\xi)$ as the prior, the posterior distribution is $p(\xi|y) \propto p(\xi)p(y|\xi)$. In variational Bayes (Attias, 2000; Waterhouse, Mackay and Robinson, 1996), the posterior distribution (which is usually intractable for complex models) is approximated with a distribution $q(\xi)$ belonging to some tractable class Q. We can assume that $q(\xi)$ takes some parametric form such as multivariate normal, or we might split ξ into blocks and assume posterior independence between the blocks. We then choose $q \in Q$ to approximate $p(\xi|y)$ as well as possible, usually in the sense of minimizing the Kullback-Leibler divergence,

$$KL(q(\xi)||p(\xi|y)) = \int \log \frac{q(\xi)}{p(\xi|y)} q(\xi) d\xi. \tag{3}$$

Since $p(y) = p(\xi)p(y|\xi)/p(\xi|y)$ for all ξ, where $p(y) = \int p(\xi)p(y|\xi)d\xi$ is the marginal likelihood, multiplying and dividing the right hand side of this expression by $q(\xi)$, taking logs, multiplying by $q(\xi)$ and then integrating gives

$$\log p(y) = \int \log \frac{p(\xi)p(y|\xi)}{q(\xi)} q(\xi) d\xi + \int \log \frac{q(\xi)}{p(\xi|y)} q(\xi) d\xi. \tag{4}$$

Since the Kullback-Leibler divergence is non-negative,

$$\mathcal{L}(q) = E_q\left(\log \frac{p(\xi)p(y|\xi)}{q(\xi)}\right)$$

is a lower bound on $\log p(y)$ (where $E_q(\cdot)$ denotes expectation with respect to $q(\xi)$), and minimizing the Kullback-Leibler divergence between $q(\xi)$ and $p(\xi|y)$ is equivalent to maximizing the lower bound $\mathcal{L}(q)$. When $q(\xi) = p(\xi|y)$, $\mathcal{L}(q) = \log p(y)$. Otherwise, the lower bound will be close to $\log p(y)$ and is a good approximation to it for purposes such as Bayesian model choice if $q(\xi)$ is close to the posterior. For more background on variational Bayes methods, see Bishop (2006) and Ormerod and Wand (2010). Tan et al. (2015) discuss a non-standard variational approximation for Gaussian process regression models which makes use of the sparse spectral approximation of Lázaro-Gredilla et al. (2010). We explain how the approach of Tan et al. (2015) works for a simple Gaussian process model in Section A of the supplementary material (see Mensah et al. (2016)). In the next section, we extend this approximation to the functional model.

4. Variational approximation for functional model

For the functional model in (1), let $y_i = (y_{i1}, \ldots, y_{in_i})^T$, $G_i = (G_{g_{(i)}}(t_{i1}), \ldots, G_{g_{(i)}}(t_{in_i}))^T$, $S_i = (S_i(t_{i1}), \ldots, S_i(t_{in_i}))^T$ and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^T$. Then we have

$$y_i = G_i + S_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2 I_{n_i}). \tag{5}$$

From Lázaro-Gredilla et al. (2010), the terms $G_g(t)$ and $S_i(t)$ can be approximated using spectral approximations. We have

$$S_i(t) \approx \sum_{r=1}^{m} a_{ir}\cos(2\pi\lambda_i t\omega_r) + b_{ir}\sin(2\pi\lambda_i t\omega_r) \text{ for } i = 1, \ldots, n,$$

where a_{ir}, b_{ir} are independent $N(0, \sigma_{\tau i}^2/m)$ for $r = 1, ..., m$, and $\omega_1, \ldots, \omega_m$, are generated independently from the spectral density of $r(h; 1, 1)$. Similarly,

$$G_g(t) \approx \sum_{r=1}^{m} \alpha_{gr}\cos(2\pi\theta_g t\omega_r) + \beta_{gr}\sin(2\pi\theta_g t\omega_r) \text{ for } g = 1, \ldots, k,$$

where α_{gr}, β_{gr} are independent $N(0, \sigma_{\kappa g}^2/m)$ and $\omega_1, \ldots, \omega_m$ are defined as above. The relationship between these representations and the Gaussian processes they approximate is explained in Section A of the supplementary material. A non-random choice for the frequencies $\omega_1, \ldots, \omega_m$, involves using the expected order statistics for a sample of size m from the spectral density of $r(h; 1, 1)$ or some approximation to these. If the spectral distribution function is $F(\omega)$, we use

$$\omega_r = F^{-1}\left(\frac{r}{m+1}\right)$$

for $r = 1, \ldots, m$. The spectral density of the Gaussian covariance with $\sigma^2 = \theta^2 = 1$ is $N(0, 1/(2\pi^2))$ so ω_r can be computed easily.

We write our model as

$$y_i = Z_{\lambda_i} c_i + W_{\theta_i} \gamma_{g(i)} + \epsilon_i, \tag{6}$$

with

$$c_i = (a_{i1}, \ldots, a_{im}, b_{i1}, \ldots, b_{im})^T \sim N(0, \sigma^2_{\tau i}/mI_{2m}), \quad i = 1, \ldots, n,$$

and

$$\gamma_g = (\alpha_{g1}, \ldots, \alpha_{gm}, \beta_{g1}, \ldots, \beta_{gm})^T \sim N(0, \sigma^2_{\kappa g}/mI_{2m}), \quad g = 1, \ldots, k,$$

where Z_{λ_i} is the $n_i \times 2m$ matrix with lth row

$$(\cos(2\pi\lambda_i t_{il}\omega_1), \ldots, \cos(2\pi\lambda_i t_{il}\omega_m), \sin(2\pi\lambda_i t_{il}\omega_1), \ldots \sin(2\pi\lambda_i t_{il}\omega_m)),$$

and W_{θ_i} is the $n_i \times 2m$ matrix with lth row

$$(\cos(2\pi\theta_{g(i)} t_{il}\omega_1), \ldots, \cos(2\pi\theta_{g(i)} t_{il}\omega_m), \sin(2\pi\theta_{g(i)} t_{il}\omega_1), \ldots, \sin(2\pi\theta_{g(i)} t_{il}\omega_m)).$$

The priors are as stated in Section 2.

Writing ξ for the full set of parameters in the model, we consider the following variational approximation $q(\xi)$ to the posterior distribution $p(\xi|y)$,

$$\begin{aligned} q(\xi) = {} & q(\beta) q(\sigma^2_\epsilon) q(\sigma^2_\lambda) q(u_\epsilon) \\ & \times \left\{ \prod_{i=1}^{n} q(\lambda_i) q(\sigma^2_{\tau_i}) q(u_{\tau_i}) q(c_i) \right\} \left\{ \prod_{g=1}^{k} q(\sigma^2_{\kappa_g}) q(u_{\kappa_g}) q(\theta_g) q(\gamma_g) \right\}, \end{aligned} \tag{7}$$

where $q(\lambda_i) \sim N(\mu^q_{\lambda_i}, {\sigma^q_{\lambda_i}}^2)$, $q(\sigma^2_\epsilon) \sim IG(a^q_\epsilon, b^q_\epsilon)$, $q(\sigma^2_\lambda) \sim IG(a^q_\lambda, b^q_\lambda)$, $q(\sigma^2_{\tau_i}) \sim IG(a^q_{\tau_i}, b^q_{\tau_i})$, $q(\sigma^2_{\kappa_g}) \sim IG(a^q_{\kappa_g}, b^q_{\kappa_g})$, $q(\theta_g) \sim N(\mu^q_{\theta g}, {\sigma^q_{\theta_g}}^2)$, $q(\beta) \sim N(\mu^q_\beta, \Sigma^q_\beta)$, $q(u_{\kappa_g}) \sim IG(a^q_{u_{\kappa_g}}, b^q_{u_{\kappa_g}})$, $q(u_{\tau_i}) \sim IG(a^q_{u_{\tau_i}}, b^q_{u_{\tau_i}})$, $q(u_\epsilon) \sim IG(a^q_{u_\epsilon}, b^q_{u_\epsilon})$, $q(\gamma_g) \sim N(\mu^q_{\gamma_g}, \Sigma^q_{\gamma_g})$ and $q(c_i) \sim N\left(\mu^q_{c_i}, \Sigma^q_{c_i}\right)$. Let ϑ denote the set of all variational parameters in $q(\xi)$. We will denote the dependence $q(\xi) = q(\xi|\vartheta)$ explicitly when needed. With the above variational posterior, the lower bound can be evaluated in closed form (see Section C of the supplementary material for details). We optimize this lower bound with respect to the variational parameters ϑ via nonconjugate variational message passing as described next.

4.1. Nonconjugate variational message passing

Suppose the variational posterior distribution factorizes into independent terms for different blocks of parameters such that

$$q(\xi) = \prod_{i=1}^{p} q(\xi_i), \tag{8}$$

where $\xi = (\xi_1^T, \ldots, \xi_p^T)^T$. The optimal choice of the factor $q(\xi_j)$ (with $q(\xi_i)$ fixed for $i \neq j$), in terms of minimizing the Kullback-Leibler divergence to the true posterior, is

$$q(\xi_j) \propto \exp\left(E_{-\xi_j}(\log p(\xi, y))\right), \tag{9}$$

where $E_{-\xi_j}(\cdot)$ denotes expectation with respect to $\prod_{i \neq j} q(\xi_i)$ and $p(\xi, y)$ is the joint distribution of (ξ, y). An iterative coordinate ascent optimization based on (9) yields the mean field variational Bayes algorithm; see for example Waterhouse, Mackay and Robinson (1996), Attias (2000) and Ghahramani and Beal (2001). There are also some extensions of the approach which relax the product restriction in various ways. For a certain type of conjugate exponential model, Winn and Bishop (2005) developed an algorithmic implementation of mean field variational Bayes called variational message passing (VMP), where the factors $q(\xi_j)$ are in the exponential family. In VMP, the updates in (9) reduce to updating the natural parameters in the appropriate exponential family and these updates can be computed using local operations on the directed graph describing conditional independencies in the model. Knowles and Minka (2011) introduce nonconjugate variational message passing (NCVMP) as an extension of VMP to deal with nonconjugate models. An approximation of the form (8) is assumed and $q(\xi_i) = q(\xi_i|\vartheta_i)$ has exponential family form

$$q(\xi_i|\vartheta_i) = \exp(\vartheta_i^T t(\xi_i) - m(\vartheta_i)),$$

where ϑ_i denotes a vector of natural parameters and $t(\xi_i)$ are the sufficient statistics. A well known property of exponential families is that the covariance matrix of $t(\xi_i)$ is

$$K(\vartheta_i) = \frac{\partial^2 m(\vartheta_i)}{\partial \vartheta_i \partial \vartheta_i^T}.$$

Suppose $p(\xi, y) = \prod_{f=1}^{s} p_f(\xi, y)$. A factor p_f is said to be in the neighbourhood of ξ_i if $p_f(\xi, y)$ depends on ξ_i and we denote this by $f \in N(\xi_i)$. NCVMP gives a recipe for updating the natural parameter of the factor $q(\xi_j)$. The update rule is

$$\vartheta_i \leftarrow \sum_{f \in N(\xi_i)} K(\vartheta_i)^{-1} \frac{\partial S_f(\vartheta_i)}{\partial \vartheta_i}, \tag{10}$$

where $S_f(\vartheta_i) = E_q(\log p_f(\xi, y))$ and $E_q(\cdot)$ denotes expectation with respect to $q(\xi)$.

The NCVMP algorithm initializes the parameters ϑ_i and then uses the update rule (10), cycling through the factors $q(\xi_j|\vartheta_j)$ until convergence in an iterative coordinate ascent algorithm. Knowles and Minka (2011) show that NCVMP reduces to standard VMP when all factors are conjugate. They also show that if NCVMP converges to a fixed point, then it is a stationary point of the Kullback-Leibler divergence and will be a minimum in practice. As NCVMP updates are based on fixed point iterations and convergence is not guaranteed, Knowles and Minka (2011) suggest using damping to fix any convergence problems. Tan and Nott (2014) show that the NCVMP algorithm can be interpreted as a natural

gradient descent algorithm with step size 1. See Rohde and Wand (2015) for further discussion of NCVMP and related algorithms.

From Knowles and Minka (2011), the updates in (10) for a univariate normal factor $q_i(\xi_i) = N(\mu^q_{\xi_i}, {\sigma^q_{\xi_i}}^2)$ can be written in terms of the means and variances as

$$
\begin{aligned}
{\sigma^q_{\xi_i}}^2 &\leftarrow -\frac{1}{2}\left\{\sum_{b\in N(\xi_i)} \frac{\partial S_b(\mu^q_{\xi_i}, {\sigma^q_{\xi_i}}^2)}{\partial {\sigma^q_{\xi_i}}^2}\right\}^{-1} \quad \text{and} \\
\mu^q_{\xi_i} &\leftarrow \mu^q_{\xi_i} + {\sigma^q_{\xi_i}}^2 \sum_{b\in N(\xi_i)} \frac{\partial S_b(\mu^q_{\xi_i}, {\sigma^q_{\xi_i}}^2)}{\partial \mu^q_{\xi_i}}. \qquad (11)
\end{aligned}
$$

Wand (2014) gives a rigorous derivation of computationally efficient updates in the multivariate normal case. The NCVMP updates for nonconjugate factors in our model (those involving the spatial dependence parameters) involve only univariate normal factors. The derivation of these updates is given in Section D of the supplementary material. The remaining updates are conjugate and may be obtained in closed form directly by optimizing the lower bound given in Appendix C. The complete set of updates is given in Algorithm 1, where the following notation is used;

For $g(i) = g$, write $\theta_i = \theta_g, W_i = E_q(W_{\theta_i})$, $W_i^* = E_q(W_{\theta_i}^T W_{\theta_i})$, $Z_i = E_q(Z_{\lambda_i})$, $Z_i^* = E_q(Z_{\lambda_i}^T Z_{\lambda_i})$, $Z_{i\mu^q_c} = Z_i \mu^q_{c_i}, W_{i\mu^q_\gamma} = W_i \mu^q_{\gamma_g}, \Sigma^q_{\mu_c} = \mu^q_c {\mu^q_c}^T + \Sigma^q_c$, $\mu^q_{y_i} = \mu^q_{c_i} + W_i \mu^q_{\gamma_{g(i)}}, \Sigma^q_{\mu_\gamma} = \mu^q_\gamma {\mu^q_\gamma}^T + \Sigma^q_\gamma$, $b^*_\epsilon = \frac{(b_\epsilon a^q_{u_\epsilon})}{b^q_{u_\epsilon}}$, $D_{\theta_g} = \frac{1}{\sigma^2_{\theta_0}}\left(\mu^q_{\theta_g} - \mu_{\theta_0}\right)$, $D_{\lambda_i} = \frac{a^q_\lambda}{b^q_\lambda}\left(\mu^q_{\lambda_i} - \upsilon_i^T \mu^q_\beta\right)$, $H_g = \{i : g(i) = g\}$ and $\mathbf{N} = \sum_{i=1}^{n} n_i$. Then we have $A_i = \partial\left(Z_i, \mu^q_{\lambda_i}\right)$, $B_i = \partial\left(Z_i^*, \mu^q_{\lambda_i}\right)$, $M_i = \partial\left(Z_i, {\sigma^q_{\lambda_i}}^2\right)$, $N_i = \partial\left(Z_i^*, {\sigma^q_{\lambda_i}}^2\right)$, $F_i = \partial\left(W_i, {\sigma^q_{\theta_i}}^2\right)$, $R_i = \partial\left(W_i^*, {\sigma^q_{\theta_i}}^2\right)$, $U_i = \partial\left(W_i, \mu^q_{\theta_i}\right)$, and $Q_i = \partial\left(W_i^*, \mu^q_{\theta_i}\right)$ and $\partial(a, b) = \frac{\partial a}{\partial b}$. The expectations in the above definitions can be evaluated using Lemma 1 of Section B in the supplementary material.

4.2. Acceleration of the basic algorithm

The independence assumptions implicit in factorized mean field approximations can be detrimental to convergence of variational Bayes algorithms. In our model, there is strong coupling between spatial dependence parameters and the corresponding spectral basis function coefficients, as well as between group and individual trends. We now explore ways to accelerate convergence of the basic NCVMP algorithm.

4.2.1. Directional adaptive nonconjugate variational message passing

Inspired by the work of Salakhutdinov and Roweis (2003) and Wang et al. (2006), Tan et al. (2015) accelerated convergence of their algorithm with an adaptive

Algorithm 1 NCVMP scheme for functional model in (5)

Initialize: $b_\epsilon^q = 0.5\sum_{i=1}^n y_i^{'} y_i$, $\Sigma_\beta^q = \Sigma_{\beta 0}$, $\mu_\beta^q = \mu_{\beta 0}$, $\mu_{\lambda_i}^q = \upsilon_i^{'}\mu_{\beta_0}$, ${\sigma_{\lambda_i}^q}^2 = \frac{b_\lambda}{a_\lambda}$,
$b_{\tau i}^q = 1, \mu_{c_i}^q = 0$ for $i = 1, \ldots, n$ and $b_{\kappa_g}^q = 1$, $\mu_{\theta_g}^q = \mu_{\theta_0}$, ${\sigma_{\theta_g}^q}^2 = \sigma_{\theta_0}^2$, for $g = 1, \ldots, k$,
$b_\lambda^q = b_\lambda$.
Set $a_\lambda^q = \frac{(a_\lambda + n)}{2}, a_\epsilon^q \leftarrow \frac{b_\epsilon}{2} + \frac{\mathbf{N}}{2}$, $a_{\tau_i}^q = m + \frac{b_{\tau_i}}{2}$, for $i = 1, \ldots, n$, $a_{\kappa_g}^q = m + \frac{b_{\kappa_g}}{2}$ for $g = 1, \ldots, k, a_{u_\epsilon}^q = \frac{(b_\epsilon + 1)}{2}$, $a_{u_{\kappa_g}}^q = \frac{(b_{\kappa_g} + 1)}{2}, a_{u_{\tau_i}}^q = \frac{(b_{\tau_i} + 1)}{2}$.
Do until the change in the lower bound is less than a specified tolerance:

- For $i = 1, \ldots, n, g = 1, \ldots, k$,
  $b_{u_{\kappa_g}}^q \leftarrow \frac{(b_{\kappa_g} a_{\kappa_g}^q)}{b_{\kappa_g}^q} + \frac{1}{A_{\kappa_g}^2}$ $b_{u_{\tau_i}}^q \leftarrow \frac{(b_{\tau_i} a_{\tau_i}^q)}{b_{\tau_i}^q} + \frac{1}{A_{\tau_i}^2}$ $b_{u_\epsilon}^q \leftarrow \frac{(b_\epsilon a_\epsilon^q)}{b_\epsilon^q} + \frac{1}{A_\epsilon^2}$
- For $g = 1, \ldots, k$,
  $\Sigma_{\gamma_g}^q \leftarrow \left[m\frac{a_{\kappa_g}^q}{b_{\kappa_g}^q} \mathrm{I}_{2m} + \frac{a_\epsilon^q}{b_\epsilon^q} \left(\sum_{i \in H_g} W_i^* \right) \right]^{-1}$ $\mu_{\gamma_g}^q \leftarrow \Sigma_{\gamma_g}^q \left(\frac{a_\epsilon^q}{b_\epsilon^q} \sum_{i \in H_g} W_i^T \right) \left[y_i^T - Z_i \mu_{c_i}^q \right]$
- For $i = 1, \ldots, n$,
  $\Sigma_{c_i}^q \leftarrow \left[m\frac{a_{\tau_i}^q}{b_{\tau_i}^q} \mathrm{I}_{2m} + \frac{a_\epsilon^q}{b_\epsilon^q} Z_i^* \right]^{-1}$ $\mu_{c_i}^q \leftarrow \Sigma_{c_i}^q (\frac{a_\epsilon^q}{b_\epsilon^q} Z_i^T) \left[y_i^T - W_i \mu_{\gamma_{g(i)}}^q \right]$
- For $i = 1, \ldots, n$,
  $b_\epsilon^q \leftarrow \frac{1}{2} \sum_{i=1}^n \left\{ y_i^T y_i - 2 y_i^T \mu_{y_i}^q + \mathrm{tr}\left(\Sigma_{\mu_{c_i}}^q Z_i^* \right) + \mathrm{tr}\left(\Sigma_{\mu_{\gamma_g}}^q W_i^* \right) + 2 Z_{i\,\mu_c^q} W_{i\,\mu_\gamma^q}^T \right\} + b_\epsilon^*$
- $b_\lambda^q \leftarrow \frac{1}{2} \sum_{i=1}^n \left(\left(\mu_{\lambda_i}^q - \upsilon_i^T \mu_\beta^q \right)^2 + {\sigma_{\lambda_i}^q}^2 + \upsilon_i^T \Sigma_\beta^q \upsilon_i \right) + b_\lambda$
- For $i = 1, \ldots, n$,
  $b_{\tau_i}^q \leftarrow \frac{m}{2} \left[{\mu_{c_i}^q}^T \mu_{c_i}^q + \mathrm{tr}(\Sigma_{c_i}^q) \right] + \frac{(b_{\tau_i} a_{u_{\tau_i}}^q)}{b_{u_{\tau_i}}^q}$
- For $g = 1, \ldots, k$,
  $b_{\kappa_g}^q \leftarrow \frac{m}{2} \left[{\mu_{\gamma_g}^q}^T \mu_{\gamma_g}^q + \mathrm{tr}(\Sigma_{\gamma_g}^q) \right] + \frac{(b_{\kappa_g} a_{u_{\kappa_g}}^q)}{b_{u_{\kappa_g}}^q}$
- $\Sigma_\beta^q \leftarrow \left\{ \frac{a_\lambda^q}{b_\lambda^q} \sum_{i=1}^n \upsilon_i \upsilon_i^T + \Sigma_{\beta_0}^{-1} \right\}^{-1}$ $\mu_\beta^q \leftarrow \Sigma_\beta^q \left\{ \frac{a_\lambda^q}{b_\lambda^q} \sum_{i=1}^n \mu_{\lambda_i}^q \upsilon_i + \Sigma_{\beta_0}^{-1} \mu_{\beta_0} \right\}$
- For $i = 1, \ldots, n$,
  ${\sigma_{\lambda_i}^q}^2 \leftarrow -\frac{1}{2} \left[\frac{a_\epsilon^q}{b_\epsilon^q} \left\{ \left(y_i - W_i \mu_{\gamma_{g(i)}}^q \right)^T M_i \mu_{c_i}^q \right\} - \frac{a_\epsilon^q}{2 b_\epsilon^q} \mathrm{tr}\left(\Sigma_{\mu_{c_i}}^q N_i \right) - \frac{a_\lambda^q}{2 b_\lambda^q} \right]^{-1}$
  $\mu_{\lambda_i}^q \leftarrow \mu_{\lambda_i}^q + {\sigma_{\lambda_i}^q}^2 \left[\frac{a_\epsilon^q}{b_\epsilon^q} \left\{ \left(y_i - W_i \mu_{\gamma_{g(i)}}^q \right)^T A_i \mu_{c_i}^q \right\} - \frac{a_\epsilon^q}{2 b_\epsilon^q} \mathrm{tr}\left(\Sigma_{\mu_{c_i}}^q B_i \right) - D_{\lambda_i} \right]$
- For $g = 1, \ldots, k$,
  ${\sigma_{\theta_g}^q}^2 \leftarrow -\frac{1}{2} \left[\frac{a_\epsilon^q}{b_\epsilon^q} \sum_{H_g} \left\{ \left(y_i - F_i \mu_{\gamma_{g(i)}}^q \right)^T Z_i \mu_{c_i}^q \right\} - \frac{a_\epsilon^q}{2 b_\epsilon^q} \sum_{H_g} \mathrm{tr}\left(\Sigma_{\mu_{\gamma_g}}^q R_i \right) - \frac{1}{2\sigma_{\theta_0}^2} \right]^{-1}$
  $\mu_{\theta_g}^q \leftarrow \mu_{\theta_g}^q + {\sigma_{\theta_g}^q}^2 \left[\frac{a_\epsilon^q}{b_\epsilon^q} \sum_{H_g} \left\{ \left(y_i - U_i \mu_{\gamma_{g(i)}}^q \right)^T Z_i \mu_{c_i}^q \right\} - \frac{a_\epsilon^q}{2 b_\epsilon^q} \sum_{H_g} \mathrm{tr}\left(\Sigma_{\mu_{\gamma_g}}^q Q_i \right) - D_{\theta_g} \right]$

scheme that employs the variational lower bound to determine whether to increase or decrease a step size. The adaptive step is utilized only in the non-conjugate updates and the step size is magnified by a pre-specified factor when

the lower bound increases, reverting to 1 when it decreases. We adapt the approach of Tan et al. (2015) here. The difference between our algorithm and theirs is that the adaptive step is applied to the updates of all parameters and not just those in the non-conjugate updates. We also experimented with a method based on the pattern search algorithm of Honkela, Valpola and Karhunen (2003). However, this requires line searches which we find do not result in a favourable trade-off between iterations to convergence and computational effort per iteration.

Recall that ϑ denotes the set of all variational parameters. At iteration t of Algorithm 1, $\vartheta^{(t)}$ is updated to $\vartheta^{(t+1)}$. We can consider a more general update of the form

$$\vartheta^{(t+1)} = \vartheta^{(t)} + d_t(\vartheta^{(t+1)} - \vartheta^{(t)}),$$

where d_t is a step size. Clearly $d_t = 1$ corresponds to the original update. Tan et al. (2015) adapt step sizes by increasing step sizes by a multiplicative factor as long as the lower bound is increasing and reverting to $d_t = 1$ when the lower bound decreases. As the NCVMP algorithm is not guaranteed to converge, it may also be of interest to consider step sizes less than one to fix convergence problems. However, we did not found any need for this in our examples. Let the multiplicative factor we use to increase step sizes be $\delta > 0$. We initialize $d_1 = 1$. Then at iteration t, we set $d_{t+1} = \delta d_t$ if a step of size d_{t+1} results in an increase in the lower bound, otherwise $d_{t+1} = 1$. Note that the same step size is applied to all variational parameters to circumvent the difficulty of adapting step size parameters in the NCVMP updates for each λ_i, $i = 1, \ldots, n$ and θ_g, $g = 1, \ldots, k$. We have experimented with $\delta \in \{1.01, 1.02, 1.08, 1.2\}$. When all parameters are being adapted at once, and ϑ is high-dimensional, a small value value of δ seems to work best, with 1.02 being close to optimal over a range of examples. We also experimented with a step halving scheme when the lower bound decreases, but the additional lower bound evaluations are not worthwhile compared with the simpler strategy of reverting to step size 1. Following Honkela, Valpola and Karhunen (2003), we transform positive parameters by logs to ensure positivity. Specifically if ζ is a positive parameter, the update at iteration t is $\tilde{\zeta}^{(t+1)} = \exp(\log \zeta^{(t)} + d_t(\log \zeta^{(t+1)} - \log \zeta^{(t)}))$. We did not transform the covariance matrix parameters in the updates as this did not cause any violations of the positive definiteness condition in our examples. When this strategy is employed, it is advisable to check for positive definiteness of covariance matrix parameters and to revert the step size to 1 in the case of any violation. Alternatively, the covariance matrices could be reparametrized in terms of, for example, the Cholesky factor.

5. Posterior inference via MCMC

A standard approach to Bayesian inference in complex models is to use Monte Carlo methods such as Markov chain Monte Carlo (MCMC) to generate samples from the posterior distribution, which can then be used to approximate relevant

expectations and probabilities. See Gelman et al. (2013) for an introductory account. A common algorithm for conditionally conjugate models popular for its automated character is the Gibbs' sampler, where we update parameter blocks by sampling from their posterior full conditional distributions. For the model we have considered here, a natural choice of blocks leads to tractable Gibbs' updates for most parameter blocks. However, the full conditional distributions for the parameters λ_i, $i = 1, \ldots, n$ and θ_g, $g = 1, \ldots, k$ do not have a standard form and it is natural to update them using Metropolis-Hastings steps, resulting in a so-called Metropolis within Gibbs scheme. Readers unfamiliar with these algorithms are referred to Gelman et al. (2013) for further background and discussion. For the Metropolis-Hastings steps, a proposal distribution is required and variational Bayes approximations have been suggested as one way to obtain good proposals (de Freitas et al., 2001). Another possibility is to use a so-called adaptive MCMC scheme where a good proposal value is learnt from the samples as the algorithm proceeds. See Andrieu and Thoms (2008) for a review of adaptive MCMC and some discussion of what is required for validity of these schemes. Adaptive schemes usually require some initialization of proposal variances and recovery from a poor initial choice can be very slow. We consider initializing an adaptive MCMC scheme using the variational posterior. The detailed algorithm is given in Section E of the supplementary material. Although there are many adaptive MCMC schemes in the literature, the adaptive steps that we employ are similar to Algorithm 5 of Andrieu and Thoms (2008). However, we do not adapt the scaling parameter in the proposal along with the proposal mean and variance. Andrieu and Thoms (2008) is also a good introduction to adaptive MCMC methods generally and describes a unifying stochastic approximation framework for such algorithms.

6. Examples

We evaluate the performance of the proposed methods in comparison with MCMC through simulation studies and then apply the methodology to a dataset of streamflow curves generated from a rainfall-runoff model. The initialization of the directional adaptive algorithm in all examples follows the default in the basic algorithm except $\Sigma^q_{c_i}$, $\mu^q_{c_i}$ and $\Sigma^q_{\gamma_j}$, for which we used one step of their NCVMP updates starting from the prior while $b^q_{u_\lambda}$, $b^q_{u_{\kappa_g}}$, $b^q_{u_{\tau_i}}$ and $b^q_{u_\epsilon}$ are set to $1/A^2_\lambda$, $1/A^2_{\kappa_g}$, $1/A^2_{\tau_i}$ and $1/A^2_\epsilon$ for $i = 1, \ldots, n$, $g = 1, \ldots, k$ respectively in addition to the defaults in Algorithm 1. The algorithms are stopped when the relative change in the variational lower bound is less than 10^{-4} for the simulated examples. For the streamflow application the tolerance is set to 10^{-5}. Initialization of the adaptive variational MCMC algorithm uses the variational posterior mean values as starting values and the variational posterior variances as initial proposal variances. All codes were written in R and run on an Intel (R) quadruplet processor Windows PC 3.40 GHz workstation. R code to implement the methods is available upon request.

6.1. Simulated dataset

This example considers a dataset of 100 functional curves comprising 30 observations per subject on 100 subjects generated from model (1). We set $\upsilon_i = (\upsilon_{i1}, \upsilon_{i2})$ where $\upsilon_{i1} = 1$ if $g(i) = 1$, otherwise 0, and $\upsilon_{i2} = 1 - \upsilon_{i1}$. The vector of regression coefficients β is set as (0.8, 0.5). With υ_i and β given, the spatial dependence parameters, λ_i and θ_g, are drawn independently from $N(\upsilon_i^T\beta, 0.01^2)$ for $i = 1, \ldots, n$, and $N(0.7, 0.01)$ for $g = 1, \ldots, k$, respectively. Then $S_i(t)$ and $G_g(t)$ are drawn from their Gaussian process conditional prior distributions with these covariance hyperparameters. We consider a two group model with $g(i) = 1$ for $i = 1, \ldots, 50$, and $g(i) = 2$ for $i = 51, \ldots, 100$. We use 30 equally spaced time points in the interval $[-5, 5]$ for all subjects. In simulating the Gaussian process functional parameters, we used $\sigma^2_{\tau_i} = 0.5$ for $i = 1, \ldots, n$, and $\sigma^2_{\kappa_g} = 0.25$ for $g = 1, \ldots, k$. Finally, the observations are drawn from model (1) with σ^2_ϵ set as 0.02^2. For the normal priors, we used $\theta_g \sim N(0.92, 0.3^2)$ and $\beta_r \sim N(0.92, 0.3^2)$ independently for $r = 1, 2$. These priors were chosen to give probability 0.95 to lag 1 correlations of the Gaussian processes being in the range $[0.1, 0.9]$. Note that if we were to use a very diffuse prior, that prior would correspond to strong prior information, putting a large prior mass on very weak dependence.

For the inverse gamma hyperparameters, we set $\sigma^2_\lambda \sim IG(3, 0.045)$. Note that σ^2_λ controls the amount of variation of individual specific covariance hyperparameters around the conditional prior mean $\upsilon_i^T\beta$. The elements of β in this application are the group means for these hyperparameters with prior standard deviation 0.3. The hyperparameters for σ^2_λ are chosen so that $E(\sigma^2_\lambda) = (0.5 \times 0.3)^2 = 0.0225$, which roughly makes the variation about the conditional mean $\upsilon_i^T\beta$ in the prior similar in magnitude to the standard deviation. For the half-t scale parameters, we use $A_\epsilon = A_{\tau_i} = A_{\kappa_g} = 25$ and we set the degree of freedom parameters $b_\epsilon = b_{\tau_i} = b_{\kappa_g} = 1$ for $i = 1, \ldots, n$, $g = 1, 2$.

Figure 1 shows a plot of the simulated data set. The two groups are evident with one group having higher frequency individual specific variations about the group trends. We fit model (6) using our MCMC approach and the variational algorithms with spectral samples of size $m \in \{20, 30, 40, 50, 60\}$. For the MCMC simulation, we use chains of length 40000 with a burnin of size 10000 for posterior analysis, and 30000 with a burnin of size 10000 for computational time comparisons.

First, we examine the performance of Algorithm 1 and its MCMC counterpart in recovering the underlying trends. The first two rows of Figure 2 shows for the run corresponding to $m = 40$, a plot of the trends of four subjects together with 95% Bayesian credible intervals, two per group over time. The two subjects plotted in each group are those for which the true λ_i takes its minimum and maximum value within that group. The third row shows a similar plot for the estimated group specific trends. We are able to recover the group and individual specific trends well with both NCVMP and MCMC. However, the VB credible intervals are narrow and often fail to capture the underlying true trends as compared to the MCMC-based credible intervals. Although VB performs poorly here in terms of posterior inference it does give good point predictions. Hence for

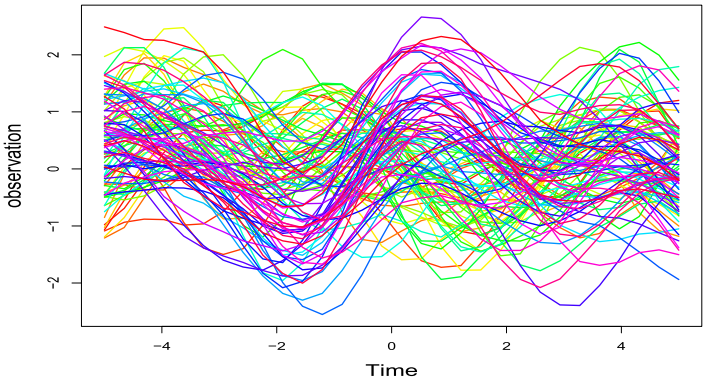

FIG 1. *Plot of simulated dataset.*

problems where interest centres on predictive inference and for which speed is important, the VB approach might be preferred. VB can also be used to obtain good proposal distributions for MCMC, and this point is illustrated later.

Figure 3 shows the estimated marginal variational posterior distributions for β_1 and β_2 estimated by variational Bayes and MCMC. For comparison, the marginal variational posterior distribution of the intercept parameter in an intercept only model for λ_i is shown as well. Although predictive inference of individual curves is not affected much by the omission of covariates the regression model helps us to understand differences in smoothness between the functional groups, something that is very important in the application of the next section.

Figure 4 shows for the run with $m = 40$, the attained lower bound at each iteration for Algorithm 1 and its adaptive variant (left), as well as the step size at each iteration for the adaptive algorithm using $\delta = 1.02$ (right).

The efficiency of the basic scheme and its adaptive variant in terms of the lower bound attained at convergence and the number of iterations required for convergence over a range of spectral basis frequencies (m) is illustrated in Figure 5. The directional adaptive NCVMP requires fewer iterations to converge on the average, yielding a significant reduction in computation time as compared to NCVMP. We have also found in other experiments that sometimes better local modes can be attained although the two algorithms perform similarly here. The additional computational effort in the adaptive scheme corresponds essentially to just one additional lower bound evaluation per iteration so that the number of iterations to convergence is a good guide to the total computational effort.

Table 1 reports the time to convergence for NCVMP, its adaptive variant with $\delta = 1.02$ and the variational MCMC sampler. It is apparent that NCVMP and its adaptive variant are faster than MCMC by an order of magnitude, and the directional adaptive scheme improves upon NCVMP.

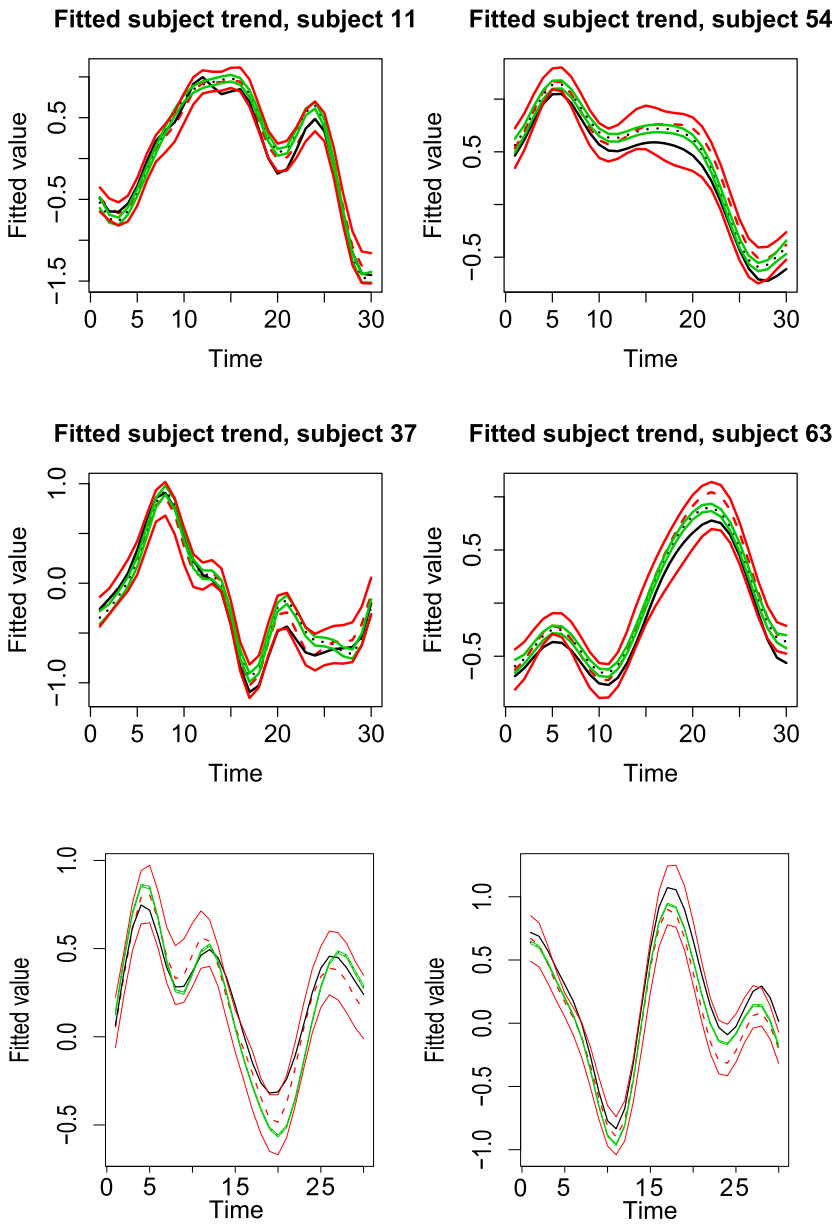

FIG 2. *Simulated data. NCVMP and MCMC-based Bayesian credible intervals for* $m = 40$. *Red and green solid curves show* 95% *Bayesian credible intervals for MCMC and NCVMP respectively. Red dashed and black dotted curves are the MCMC and NCVMP fitted trends respectively while the black solid curves correspond to the true trends. The first two rows are subject specific trends, two per group corresponding respectively to minimum and maximum value of* λ_i *within each group. The third row gives group specific trends for group 1 (left) and group 2 (right).*

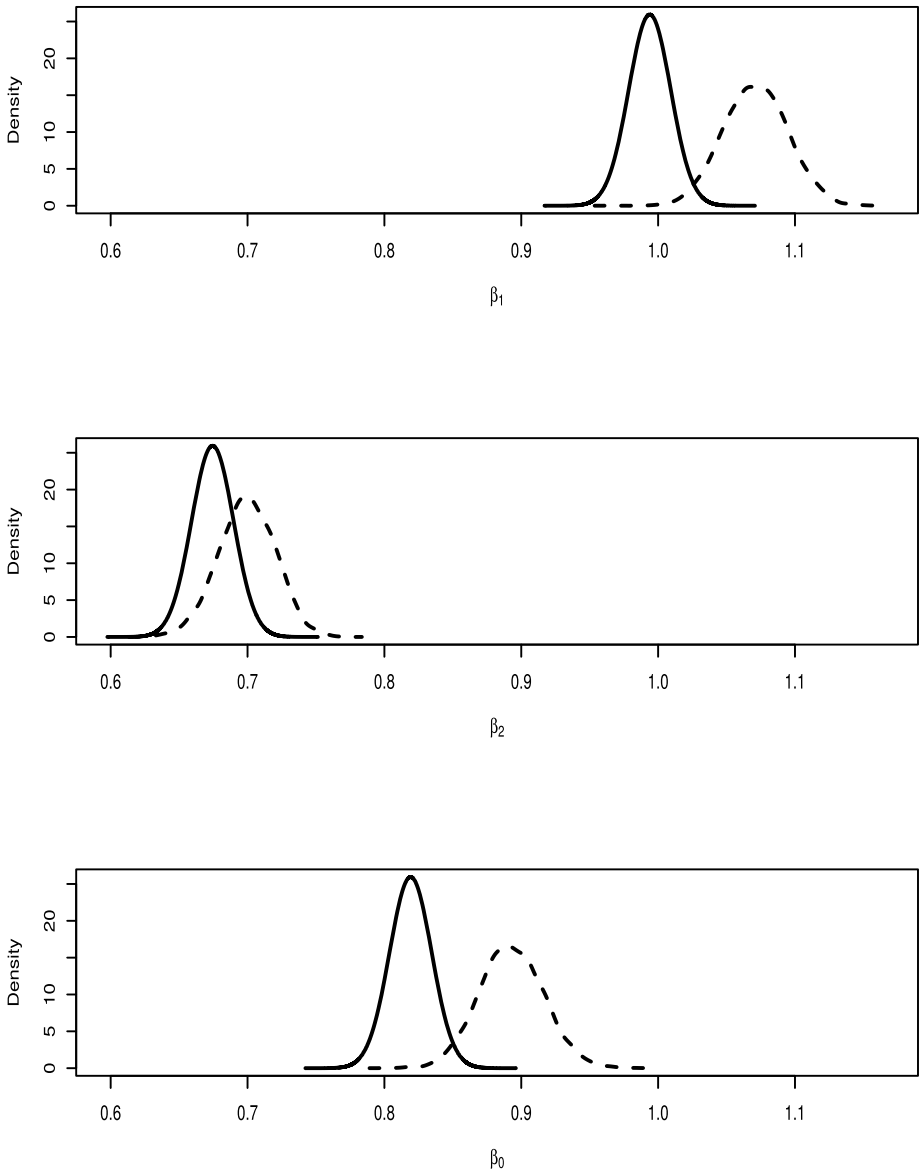

FIG 3. *Simulated data. Marginal variational posterior distribution of β_1 (top) and β_2 (middle) obtained using variational Bayes (solid curves) and MCMC (dashed curves) for model including covariates, and for intercept in an intercept only model (bottom).*

6.2. French Broad River catchment streamflow dataset

Hydrologic models take as input time series of climatic variables (typically rainfall and evapotranspiration) to simulate time series of streamflow. These mod-

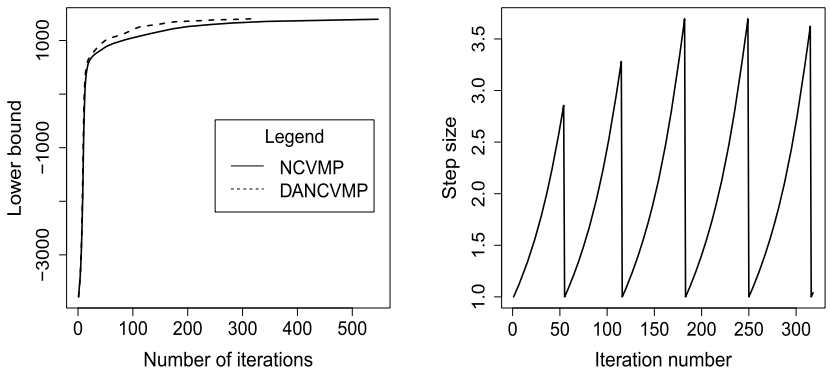

FIG 4. *Simulated data. Plot of lower bound (left) and adaptive step size d_t (right) against iterations in the directional adaptive scheme with $m = 40$. Dashed line is DANCVMP with $\delta = 1.02$ and solid line is NCVMP.*

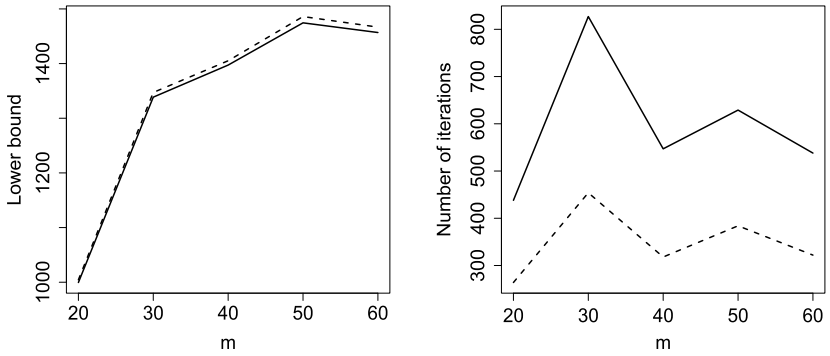

FIG 5. *Simulated data. Plot of lower bound attained at convergence (left) and iterations to converge (right) versus number of spectral basis frequencies m. In each plot, solid line is NCVMP and dashed line is DANCVMP with $\delta = 1.02$.*

TABLE 1

Comparison of variational Bayes and MCMC for simulated dataset. First three columns gives the time to convergence of NCVMP, its adaptive variant and MCMC (30000 iterations) in seconds. The fourth column gives the ratio of the times to convergence for NCVMP against DANCVMP. The last two columns give the ratios of the time to convergence for variational MCMC against those for NCVMP and DANCVMP respectively

	Time					
m	NCVMP	DANCVMP	MCMC	Ratio 1	Ratio 2	Ratio 3
20	2757.71	2159.01	48844.30	1.28	17.71	22.62
30	9598.91	7818.13	99438.91	1.23	10.36	12.72
40	14848.83	8246.37	188324.94	1.80	12.68	22.84
50	20177.43	10384.00	237140.91	1.94	11.75	22.84
60	29089.80	18603.74	381902.00	1.56	13.13	20.53

els aim to mathematically represent common hydrologic processes such as soil water storage, surface runoff and baseflow. Hydrologic models are used for ap-

plications such as flood forecasting, climate change impact studies, or predictions in ungauged basins. Fitting models such as the ones we have developed here to streamflow data for different catchments can be a way of assessing different ways of grouping the catchments in terms of their dynamic properties; such groupings can be used as a surrogate for regions without data. This relates to the problem of prediction in ungauged basins, as we discuss further below. Our model provides a formal way of assessing the meaningfulness of catchment groupings in terms of representing different dynamics in such an application.

Characterizing the uncertainties affecting hydrologic models is a considerable challenge in hydrologic science and practice (Renard et al., 2010). Of particular concern is the impact of uncertain rainfall inputs (due to measurement error or inadequate spatial sampling) on parameter estimates and runoff forecasts. Rainfall errors can be modelled via storm-dependent multiplicative terms, where it may be assumed that rainfall multipliers follow a lognormal distribution (Kavetski, Kuczera and Franks, 2006).

For this study, we sample a series of storms from six-hourly rainfall data and evapotranspiration estimates for the French Broad River at Asheville, North Carolina. The data are modelled via a widely used hydrologic model known as the Probability Distribution Model (PDM). The PDM uses inputs of rainfall and evapotranspiration to produce time series of streamflow. The model represents the spatial variability of soil water capacity via a Pareto distribution, and our version of the model additionally incorporates parameters representing fast and slow reservoir routing (Smith and Marshall, 2009). This version of the PDM consists of 6 model parameters, of which 3 were kept fixed for all catchments in the study. (See Smith and Marshall, 2009, for a complete description of the model and parameters). A more detailed description of the model is given in Section F of the supplementary material.

To represent a collection of catchments, we specified three different groups of catchments with 100 members for each group. Recent research in hydrology has sought to classify catchments in this way, such that members of each group could be considered to have similar hydrologic processes and thus similar modelled behaviour (Sawicz et al., 2011; Wagener et al., 2007). For this study, we fixed three PDM parameters and allowed three parameters to vary between each catchment: the maximum storage capacity (C_{max}), the surface runoff outflow rate (T_q), and the baseflow outflow rate (T_s). Assuming groups were normally distributed, a mean and standard deviation was specified for each group (see Table 1 of supplementary material). These values were selected based on the physical bounds and typical values of these parameters in other hydrologic studies. One hundred parameter sets were then randomly sampled within each group representing individual catchments. Streamflow was then simulated with the PDM for every catchment using the same rainfall.

To represent different levels of measurement uncertainty in the rainfall data for each catchment, we specified a lognormal distribution of multiplicative rainfall errors that were storm dependent (i.e. the same rainfall multiplier is used over a whole storm and storms are defined as a period of continuing rain with

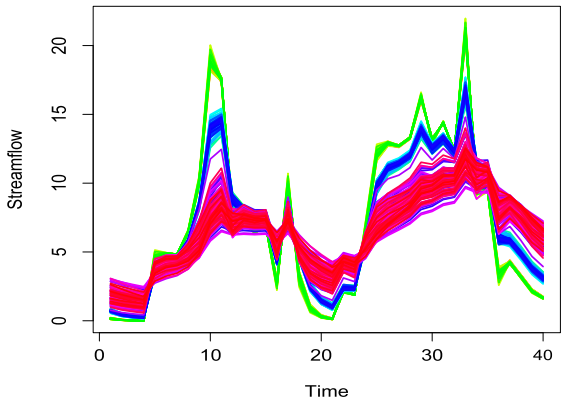

FIG 6. *Plot of streamflow dataset.*

no breaks). The mean of the sampled rainfall multipliers was kept fixed at 1.0, but the variance of the multipliers varied between groups. For simplicity, the "true" rainfall was assumed to be the same for each catchment.

Denoting the streamflow output as y, we use y as the response. We fit our model to a subset of the data for 40 time steps chosen to cover a major rainfall event using algorithm 1. We use the simulated hydrologic catchment parameters, namely, the maximum storage capacity $\mathrm{C}_{\max}$, the surface runoff outflow rate T_q and the baseflow outflow rate T_s as well as catchment group indicators as covariates. Precisely, we set $\upsilon_i = (\upsilon_{i1}, \upsilon_{i2}, \upsilon_{i3}, \upsilon_{i4}, \upsilon_{i5}, \upsilon_{i6})$ where υ_{i1}, υ_{i2}, υ_{i3} are indicators for the catchment groups and υ_{i4}, υ_{i5} and υ_{i6} are respectively centred and scaled versions of $\mathrm{C}_{\max}$, T_q and T_s. For the prior distributions, we used the same prior settings as in example 1 except for the regression coefficients corresponding to υ_{i4}, υ_{i5} and υ_{i6} which were set as $N(0, 0.3^2)$ independently. For this application we initialized the NCVMP algorithm with the default settings except $b^q_{\tau_i}$ and $b^q_{\kappa_g}$ which were set as half the scale parameter values of $\sigma^2_{\tau_i}$ and $\sigma^2_{\kappa_g}$.

Figure 6 shows a plot of the streamflow data set. A referee has pointed out that this data contains non-stationary features, and this is certainly the case. However, we emphasize that although our prior distributions on functional terms are stationary, the corresponding posterior distributions after updating are not. In any case very flexible non-stationary models may be difficult to fit in longitudinal models with few observations per functional term. The three catchment groups are evident in the plot of the data; the groups exhibit different levels of high frequency variability following the different noise levels in the rainfall inputs. We report results here for our model using $m = 90$ spectral points.

Figures 7 and 8 show variational Bayes estimates of the posterior distributions of regression coefficients. The coefficients β_1, β_2, β_3 represent the mean of λ_i for the three different groups for average values of the other covariates. We see in particular that the third catchment group contains much smoother functional observations, something that is expected here on hydrological grounds.

Figures 9 shows the fitted catchment group specific trends and Figure 10

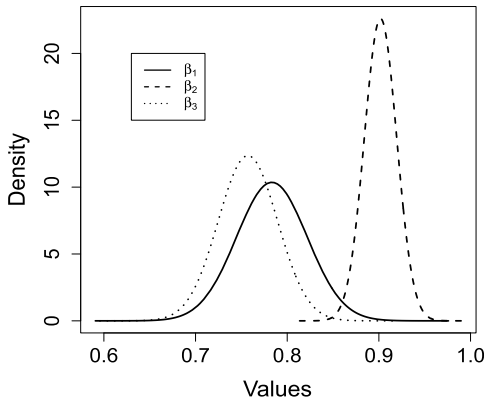

FIG 7. *Streamflow data. Plot of marginal variational posterior distributions of regression coefficients for the catchment indicators.*

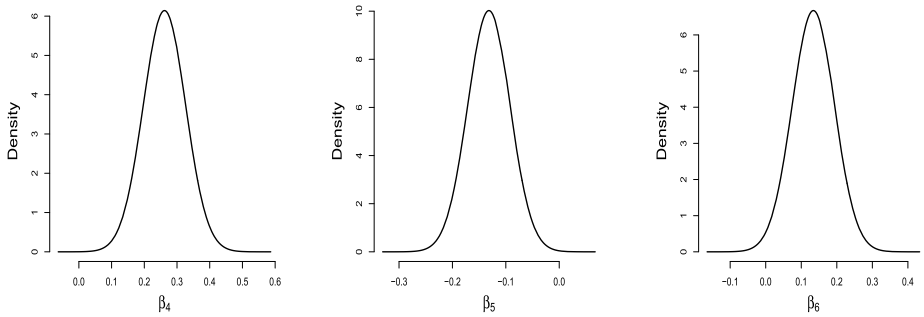

FIG 8. *Streamflow data. Plot of marginal variational posterior distributions of regression coefficients* β_4, β_5 *and* β_6.

plots catchment specific trends for the first three catchments in each catchment group. The varying smoothness between the 3 groups is evident. MCMC results from this application are similar (Results not presented).

The ability to identify group-specific trends in measurement error and streamflow dynamics is particularly important from a hydrologic standpoint. A recent and ongoing concern in hydrologic science is streamflow forecasting in catchments without available observations (Sivapalan et al., 2003). To address this, a myriad of studies have focused on regionalization methods that aim to identify natural catchment groupings (as expressed in streamflow dynamics) so that catchment groups may act as surrogates for regions without data (e.g. Wagener and Wheater, 2006). These methods are often impacted by the presence of potentially strong measurement error, affecting the ability to appropriately identify catchment groups and to estimate typical group behavior or functioning. The methods presented here provide insight to how catchment groups vary in terms of their streamflow dynamics and could thus be related to catchment physical

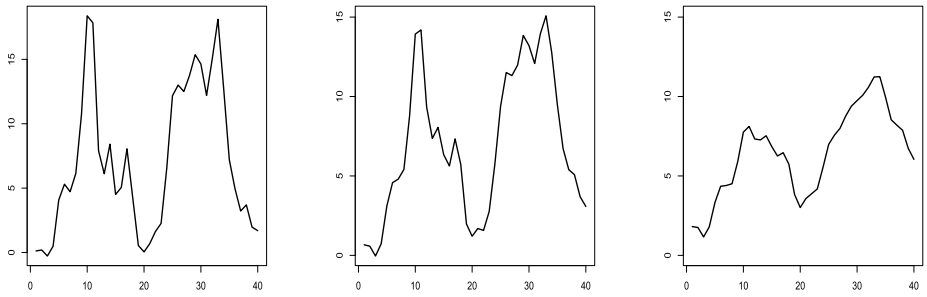

FIG 9. *Streamflow data. Plot of fitted group specific streamflow from Algorithm 1 (NCVMP) with* $m = 90$. *Arranged from left to right are Groups 1, 2 and 3, x-axis are time points and y-axis are fitted values. MCMC results are similar.*

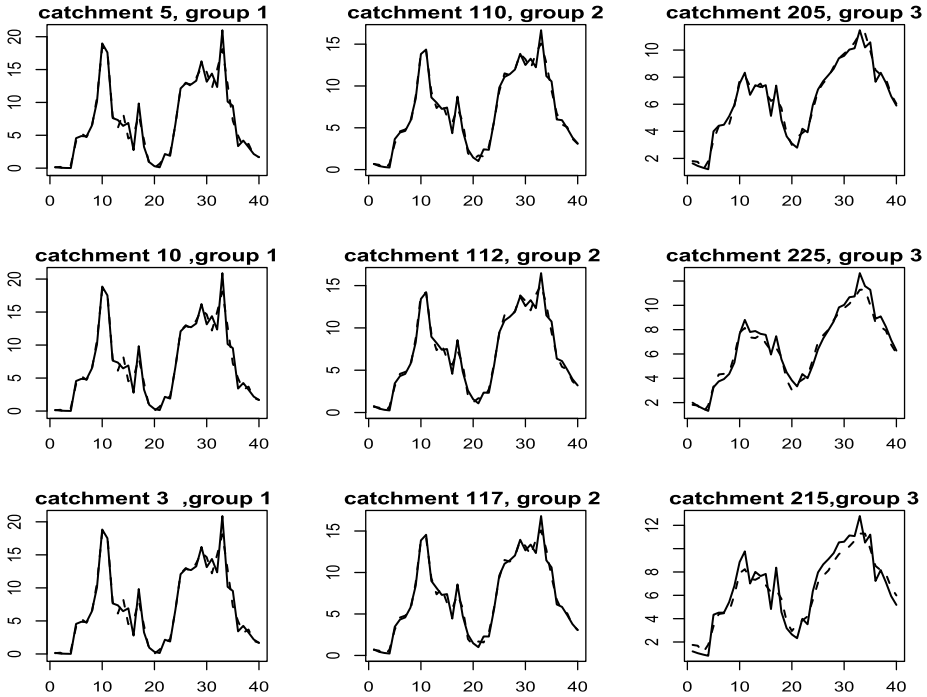

FIG 10. *Streamflow data. Plot of fitted catchment specific streamflow from Algorithm 1 (NCVMP), x-axis are time points and y-axis are fitted values. Solid curves correspond to observed streamflow and dash curves represent fitted streamflow. The three catchments presented in each group are the first three in each catchment group.*

properties. The ability to identify trends in measurement error between groups is particularly novel and can suggest how errors impact subsequent hydrologic models and forecasts. Of interest would then be the natural extension of these methods to varying input rainfall.

7. Conclusion

We have presented a novel Gaussian process approach to grouped functional modelling for longitudinal data. This approach models subject and group specific trends with Gaussian processes and relates individual specific smoothness to covariates for different subjects. We have developed fast variational inference methods using a sparse spectral approximation. We have also explored the joint use of variational Bayes methods and MCMC sampling algorithms. A referee has suggested looking at the inclusion of covariates into the model at the observation level. We agree that this is a worthy extension but also believe it will be a non-trivial one, due to the way the additional flexibility may make it more difficult to identify subject specific smoothness depending on covariates.

Supplementary Material

Supplementary material for "Functional models for longitudinal data with covariate dependent smoothness"
(doi: 10.1214/16-EJS1113SUPPA; .pdf).

References

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing* **18** 343–373. MR2461882

Attias, H. (2000). A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12* 209–215. MIT Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. MR2247587

de Freitas, N., Højen-Sørensen, P., Jordan, M. I. and Russell, S. (2001). Variational MCMC. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. UAI'01* 120–127. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition. Chapman & Hall/CRC Texts in Statistical Science*. Taylor & Francis. MR3235677

Ghahramani, Z. and Beal, M. J. (2001). Propagation algorithms for variational Bayesian learning. In *Advances in Neural Information Processing Systems 13* (T. K. Leen, T. G. Dietterich and V. Tresp, eds.) 507–513. MIT Press.

Goldsmith, J., Wand, M. P. and Crainiceanu, C. (2011). Functional regression via variational Bayes. *Electronic Journal of Statistics* **5** 572–602. MR2813555

Honkela, A., Valpola, H. and Karhunen, J. (2003). Accelerating cyclic update algorithms for parameter estimation by pattern searches. *Neural Processing Letters* **17** 191–203.

Kavetski, D., Kuczera, G. and Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 1.Theory. *Water Resources Research* **42** 1–9.

Knowles, D. A. and Minka, T. P. (2011). Non-conjugate variational message passing for multinomial and binary regression. In *Advances in Neural Information Processing Systems 24* (J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. Pereira and K. Q. Weinberger, eds.) 1701–1709.

Lázaro-Gredilla, M., Quiñonero-Candela, J., Rasmussen, C. E. and Figueiras-Vidal, A. R. (2010). Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research* **11** 1865–1881. MR2660655

Mensah, D. K., Tan, L. S. L., Marshall, L. and Nott, D. J. (2016). Supplementary material for "Functional models for longitudinal data with covariate dependent smoothness". DOI: 10.1214/16-EJS1113SUPPA.

Müller, H. G. and Yao, F. (2010). Empirical dynamics for longitudinal data. *The Annals of Statistics* **38** 3458–3486. MR2766859

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician* **64** 140–153. MR2757005

Rasmussen, C. E. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. The MIT Press. MR2514435

Reithinger, F., Jank, W., Tutz, G. and Shmueli, G. (2008). Smoothing sparse and unevenly sampled curves using semiparametric mixed models: An application to online auctions. *Journal of the Royal Statistical Society, Series C* **57** 127–148. MR2420433

Renard, B., Kavetski, D., Kuczera, G., Thyer, M. and Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research* **46**. W05521.

Rohde, D. and Wand, M. P. (2015). Semiparametric mean field variational Bayes: General principles and numerical issues. Available at `http://matt-wand.utsacademics.info/RohdeWand.pdf`.

Salakhutdinov, R. and Roweis, S. (2003). Adaptive overrelaxed bound optimization methods. In *Proceedings of the 20th International Conference on Machine Learning* (T. Fawcett and N. Mishra, eds.) 664–671. AAAI Press.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. and Carrillo, G. (2011). Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrology and Earth System Sciences* **15** 2895–2911.

Shi, J. Q. and Choi, T. (2011). *Gaussian process regression analysis for functional data*. Chapman and Hall/CRC. MR2932935

Shi, J. Q., Murray-Smith, R. and Titterington, D. M. (2005). Hierarchical Gaussian process mixtures for regression. *Statistics and Computing* **15** 31–41. MR2137215

Shi, J. Q., Wang, B., Will, E. J. and West, R. M. (2012). Mixed-effects Gaussian process functional regression models with application to dose-response curve prediction. *Statistics in Medicine* **31** 3165–3177. MR2993619

Sivapalan, M., Takeuchi, K., Franks, S., Gupta, V., Karam-

BIRI, H., LAKSHMI, V., LIANG, X., MCDONNELL, J., MENDIONDO, E., O'CONNELL, P. et al. (2003). IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal* **48** 857–880.

SMITH, T. and MARSHALL, L. (2009). Exploring uncertainty and model predictive performance concepts via a modular snowmelt-runoff modeling framework. *Environmental Modelling & Software* **26** 691–701.

TAN, L. S. L. and NOTT, D. J. (2014). A Stochastic Variational Framework for Fitting and Diagnosing Generalized Linear Mixed Models. *Bayesian Anal.* **9** 963–1004. MR3293964

TAN, L. S. L., ONG, V. M. H., NOTT, D. J. and JASRA, A. (2015). Variational inference for sparse spectrum Gaussian process regression. *Statistics and Computing*, to appear.

WAGENER, T. and WHEATER, H. S. (2006). Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty. *Journal of Hydrology* **320** 132–154.

WAGENER, T., SIVAPALAN, M., TROCH, P. and WOODS, R. (2007). Catchment classification and Hydrologic Similarity. *Geography Compass* **1** 901–931.

WAND, M. P., ORMEROD, J. T., PADOAN, S. A., FUHRWIRTH, R., et al. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis* **6** 847–900. MR2869967

WAND, M. P. (2014). Fully simplified multivariate normal updates in non-conjugate variational message passing. *Journal of Machine Learning Research* **15** 1351–1369. MR3214787

WANG, Y. and KHARDON, R. (2012). Nonparametric Bayesian Mixed-effect model: A sparse Gaussian process approach. http://arxiv.org/abs/1211.6653.

WANG, B., TITTERINGTON, D. et al. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis* **1** 625–650. MR2221291

WANG, S., JANK, W., SHMUELI, G. and SMITH, P. (2008). Modeling price dynamics in eBay auctions using principal differential analysis. *Journal of the American Statistical Association* **103** 1100–1118. MR2528829

WATERHOUSE, S., MACKAY, D. and ROBINSON, T. (1996). Bayesian methods for mixture of experts. In *Advances in Neural Information Processing Systems 8* 351–357. MIT Press.

WINN, J. and BISHOP, C. M. (2005). Variational message passing. *Journal of Machine Learning Research* **6** 661–694. MR2249835

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

ZHU, B. and DUNSON, D. B. (2012). Stochastic volatility regression for functional data dynamics. http://arxiv.org/abs/1212.0181.

ZHU, B., TAYLOR, J. M. G. and SONG, P. X. K. (2011). Semiparametric stochastic modeling of the rate function in longitudinal studies. *Journal of the American Statistical Association* **106** 1485–1495. MR2896851