ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2016)
65, *Part* 2, *pp.* 215–236

Assessing systematic effects of stroke on motor control by using hierarchical function-on-scalar regression

Jeff Goldsmith and Tomoko Kitago

Columbia University, New York, USA

[Received February 2014. Revised May 2015]

Summary. This work is concerned with understanding common population level effects of stroke on motor control while accounting for possible subject level idiosyncratic effects. Upper extremity motor control for each subject is assessed through repeated planar reaching motions from a central point to eight prespecified targets arranged on a circle. We observe the kinematic data for hand position as a bivariate function of time for each reach. Our goal is to estimate the bivariate function-on-scalar regression with subject level random functional effects while accounting for potential correlation in residual curves; covariates of interest are severity of motor impairment and target number. We express fixed effects and random effects by using penalized splines, and we allow for residual correlation by using a Wishart prior distribution. Parameters are jointly estimated in a Bayesian framework, and we implement a computationally efficient approximation algorithm using variational Bayes methods. Simulations indicate that the method proposed yields accurate estimation and inference, and application results suggest that the effect of stroke on motor control has a systematic component observed across subjects.

Keywords: Bayesian regression; Bivariate data; Gibbs sampler; Penalized splines; Variational Bayes method

1. Introduction

Stroke is the leading cause of long-term disability in the USA, with an incidence of over 795000 events each year (Go *et al.*, 2013)—a rate that is expected to grow to over 1 million by 2025 (Broderick, 2004). Disability induced by stroke is manifested in many activities including motor control, speech and cognitive performance. Between 30% and 66% of stroke patients have clinically apparent motor deficits involving the upper extremity at 6 months (Kwakkel *et al.*, 2003). Because of this, there is a great need for development of neurorehabilitative therapies to improve arm movements after stroke. One of the challenges in developing and testing therapeutic interventions to promote motor recovery after stroke is that it remains unclear to what extent these motor deficits are idiosyncratic (or subject specific) rather than common across affected patients, and how these deficits vary according to the severity of motor impairment. A better understanding of these factors would allow for therapies that target the specific motor deficits that are shared by stroke patients, but perhaps also to tailor therapies based on individual characteristics, such as severity of the stroke.

In this current study we focus on the effects of stroke on motor control, which we define as the ability to make accurate, goal-directed movements. We use a planar reaching task designed to test

Address for correspondence: Jeff Goldsmith, Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10023, USA.
E-mail: jeff.goldsmith@columbia.edu

 0035–9254/16/65215

a fundamental level of motor control (Kitago *et al.*, 2013) and explore the relationship between patients' performance on this task and the Fugl-Meyer upper extremity motor assessment score FM-UE (Fugl-Meyer *et al.*, 1974), which is a clinical measure of the severity of arm motor impairment, in a population of chronic stroke patients with residual arm paresis. Elderly healthy controls are included as a reference group.

In the reaching task, observations at the subject level are repeated two-dimensional motion trajectories to eight target directions parameterized by time. Our analytical approach for these multilevel bivariate functional data is to model jointly main effects for motor impairment and target direction, subject level random effects and residual correlation in a Bayesian function-on-scalar regression.

1.1. Two-dimensional planar reaching data

We now describe the scientific setting and data structure in more detail. Our study population consists of patients who had a first-time ischaemic or haemorrhagic stroke 6 or more months in the past and have residual paresis of the affected arm (FM-UE less than the maximum score of 66). Exclusion criteria include multiple-stroke events, haemorrhagic stroke, traumatic brain injury, major non-stroke medical illness that alters brain function, an orthopaedic or neurological condition that interferes with arm function or inability to give informed consent. Selected patients exhibit moderate-to-severe motor impairment in the affected arm. Healthy controls with an age distribution similar to that in stroke patients are included as a reference group.

As a measurement of upper extremity motor control, subjects make repeated centre-out arm reaching movements to eight targets in the following experimental design. After subjects have been seated to align the shoulder, elbow and hand in the horizontal plane, the trunk is comfortably secured and the wrist and hand are immobilized with a splint. The forearm is supported on an air sled system to reduce the effects of friction and gravity, diminishing the effect of strength deficits on motions and isolating motor control. Subjects make reaching movements from a central starting point to eight targets arranged equidistantly on a circle of radius 8 cm around the starting point. The centre-out reaching movements that are required can be performed by all but the most severely impaired subjects. Before data acquisition, a short introductory period familiarizes subjects with the experiment configuration. These data were collected as part of baseline assessments for two longitudinal stroke intervention studies (Kitago *et al.*, 2013; Huang *et al.*, 2012) and a study of cerebral blood flow after stroke (unpublished data), which were approved by the Columbia University Medical Center Institutional Review Board.

Kinematic data are recorded for each motion made by each subject, i.e. we observe the X- and Y-co-ordinate of the hand position as a function of time, giving bivariate functional observations $(P_{ij}^X(t), P_{ij}^Y(t))$ for subjects i and motions j. Our data set consists of 24 healthy controls, 25 mildly affected stroke patients (FM-UE 44 and above) and eight severely affected stroke patients (FM-UE < 44); all participants make 22 reaching motions with both their dominant and their non-dominant hands to each of the eight targets, giving 352 motions for each subject and roughly 20000 overall bivariate functional observations (due to technical errors in recording, some motions have been removed from the data set). Although the data are inherently functional in nature, existing analyses have primarily focused on scalar summaries of observed trajectories including the deviation of the end point from the target, peak velocity and curvature (Levin, 1996; Lang *et al.*, 2006; Coderre *et al.*, 2010).

Fig. 1 shows the observed data for one healthy control in Figs 1(a)–1(c) and one severely affected stroke patient in Figs 1(d)–1(f). In Figs 1(a) and 1(d) are complete trajectories, illus-

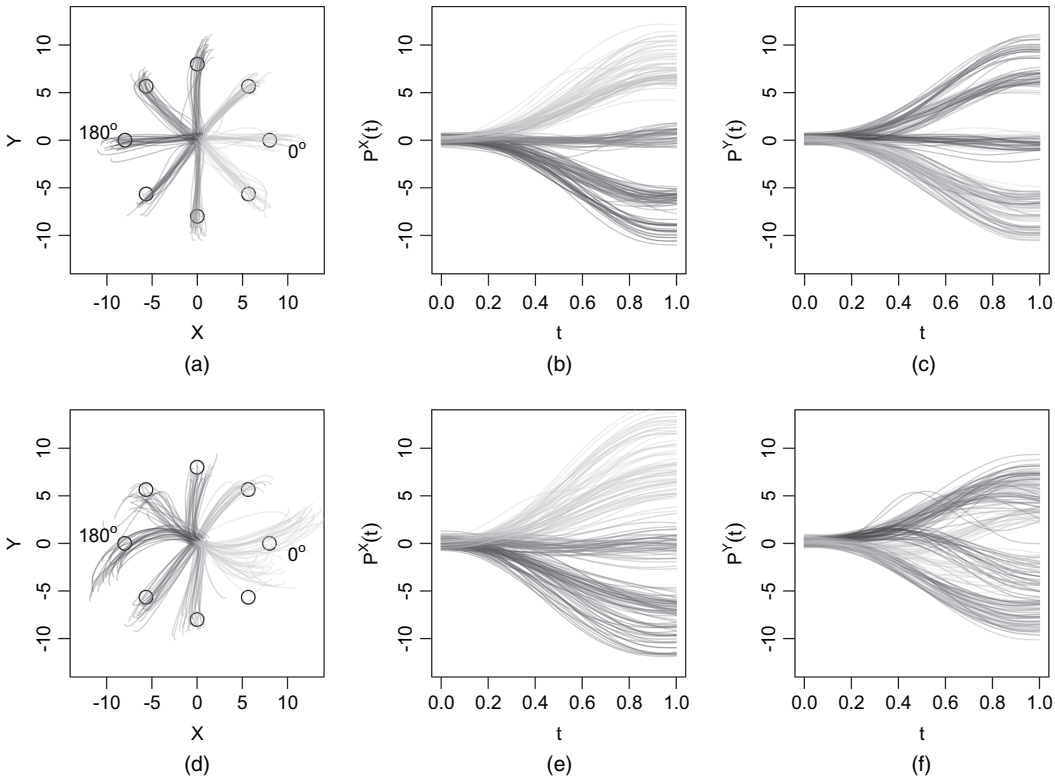

Fig. 1. Observed data for two subjects: (a)–(c) the dominant hand of a healthy control; (d)–(f) the affected dominant hand of a severe stroke patient; (a), (d) observed kinematic data for all reaches observed in the dominant hand; (b), (e) X-position for all reaches; (c), (f) Y-position for all reaches

trating the full path of each reaching motion coloured according to target. There are clear differences comparing the healthy control and stroke patient, particularly in the average motion that is made to each target: for instance, for the target at 0° the stroke patient exhibits both overextension and increased curvature with respect to the control subject. Figs 1(b) and 1(e), and 1(c) and 1(f) show the constituent functions $P_{ij}^X(t)$ and $P_{ij}^Y(t)$ that make up the kinematic data for each trajectory—we shall model these by using a combination of population level fixed effects, subject level random effects and curve level functional principal component effects. The stroke patient has unilateral tissue damage due to blockage of the right middle cerebral artery, which results in disrupted motor skill in the dominant arm.

Our goal in this analysis is to explore the extent to which the effects of stroke on motor control are shared across subjects or are subject specific through a regression analysis using a combination of subject level scalar covariates, such as the Fugl-Meyer measure of impairment severity, as predictors of interest. Evidence for systematic effects of stroke on motor control would indicate that the induced control abnormality is not entirely subject specific, but rather that disrupting the motor cortex or its descending pathways leads to predictable deficits in upper extremity motor control. The data structure necessitates correctly accounting for subject level effects through the inclusion of random functional intercepts, and accurate inference depends on incorporating residual correlation. Throughout, our outcome is the bivariate kinematic function for hand position over time.

1.2. Statistical methods

Conceptually, we observe functional data $[P_{ij}^X(t), P_{ij}^Y(t), \mathbf{w}_i]$ for subjects $i = 1, \ldots, I$ and visits $j = 1, \ldots J_i$ for a total number of observations $n = \Sigma_i J_i$. In our application $P_{ij}^X(t)$ and $P_{ij}^Y(t)$ are the X- and Y-position curves indexed by time $t \in [0, 1]$ and $\mathbf{w}_i = (w_{i1}, \ldots, w_{ip})$ is a length p vector of scalar covariates. We propose the model

$$\begin{aligned} P_{ij}^X(t) &= \beta_0^X(t) + \sum_{k=1}^{p} w_{ik}\, \beta_k^X(t) + b_i^X(t) + \epsilon_{ij}^X(t), \\ P_{ij}^Y(t) &= \beta_0^Y(t) + \sum_{k=1}^{p} w_{ik}\, \beta_k^Y(t) + b_i^Y(t) + \epsilon_{ij}^Y(t) \end{aligned} \tag{1}$$

where $\beta_k^X(t)$ and $\beta_k^Y(t)$ are fixed effects associated with scalar covariates, $b_i^X(t)$ and $b_i^Y(t)$ are subject-specific random effects, and $\epsilon_{ij}^X(t)$ and $\epsilon_{ij}^Y(t)$ are potentially correlated residual curves. Penalized splines are used to estimate fixed and random effects; although many options are possible, we shall use a cubic B-splines basis with a combined zeroth-and second-derivative penalty throughout. All parameters are modelled in a Bayesian framework that allows the joint modelling of the mean structure (through fixed and random effects) and residual correlation (through the error covariance matrix) in a single Gibbs sampler. Importantly, a variational Bayes algorithm provides a computationally efficient and accurate approximation to the full sampler. Model (1) is analogous to a standard mixed model with subject level random intercepts; the identifiability of fixed and random effects will depend on the prior specification, hyperparameter selection and sampling framework, which we discuss in Sections 2.1 and 2.3.

In practice, observations are not truly functional but are observed as structured discrete vectors. For notational simplicity, we assume that functional observations lie on a dense grid of the domain $[0, 1]$ with D elements, and that this grid is common to all subjects. In our application, trajectories are observed at 120 Hz. Although efforts were made to ensure uniform motion times of approximately 0.5 s, motions take different times to complete and we use a linear registration to obtain a common, evenly spaced observation grid of length $D = 25$ before analysis. After registration, motions are observed from $t = 0$ to $t = 1$ with 0 and 1 indicating the beginning and end of the motion respectively. Despite being observed as vectors, objects that are functional in nature will be denoted as $f(t)$ where appropriate to emphasize the structure underlying the data.

There is a large body of existing work for the analysis of functional outcome models. We broadly consider two methodological categories, the first of which consists of approaches that seek to estimate each curve in the data set. Brumback and Rice (1998) posed a function-on-scalar regression in which population level coefficients and curve level deviations are modelled by using penalized splines; for computational convenience, intercepts and slopes for curve level effects were treated as fixed effects. Guo (2002) extended this approach by formulating curve-specific deviations as random effects. Because of the difficulty in estimating all curves by using penalized splines, these approaches can be computationally intensive for large data sets. Functional principal component methods for cross-sectional data (Yao *et al.*, 2005), as well as recent extensions for multilevel (Di *et al.*, 2009), longitudinal (Greven *et al.*, 2010) and spatially correlated data (Staicu *et al.*, 2010), have modelled curve-specific deviations from a population mean by using low dimensional basis functions estimated from the empirical covariance matrix. These methods did not focus on the flexible estimation of the population mean surface; moreover in assessing uncertainty these methods implicitly conditioned on estimated decomposition objects, which can lead to the understatement of total variability (Goldsmith *et al.*, 2013).

Alternatively, one can view individual curves as errors around the population or subject level mean of interest. This approach is described in Ramsay and Silverman (2005), section 13.4, in which fixed effects at the population level were estimated by using penalized splines but individual curves were not directly modelled. Reiss *et al*. (2010) built on this approach by taking advantage of the inherent connection between penalized splines and ridge regression to develop fast methods for leave-one-out cross-validation to select tuning parameters. Scheipl *et al*. (2015) proposed a very flexible class of functional outcome models, allowing cross-sectional or multilevel data as well as scalar or functional predictors and estimating effects in a mixed model framework; a robust software implementation of this method is provided in the `refund` R package (Crainiceanu *et al*., 2012). A drawback of these approaches is the assumption that error curves consist only of uncorrelated measurement error despite clear correlation in the functional domain. One alternative, which was proposed by Reiss *et al*. (2012), is an iterative procedure to estimate the mean structure and then, using this mean, the residual covariance matrix followed by a re-estimation of the mean by using generalized least squares. Doing so necessarily increases the computational burden and does not allow joint estimation of the mean and covariance; additionally, coverage properties of this approach have not been presented.

Several Bayesian methods for function-on-scalar regression exist in the literature. Morris *et al*. (2003) developed wavelet-based functional mixed models assuming that residual curves consist of independent measurement errors; Morris and Carroll (2006) extended this to allow correlated residual curves. Both approaches used a discrete wavelet transform of the observed data and model coefficients in the wavelet domain by using spike-and-slab priors. It was assumed that errors in the wavelet domain are independent, which was justified heuristically by the whitening property of the discrete wavelet transform. After modelling using Markov chain Monte Carlo methods, the inverse discrete wavelet transform provided estimates in the observed space. A penalized spline approach for functional mixed models was taken in Baladandayuthapani *et al*. (2007), whereas Baladandayuthapani *et al*. (2010) used a piecewise constant basis; in both cases, correlated residual curves were explicitly modelled rather than treated as errors around the mean. For cross-sectional functional data observed sparsely at the subject level, Montagna *et al*. (2012) developed a Bayesian latent factor model in which predictors are incorporated at the latent factor level; a potentially large number of factors are allowed, and sparsity is induced through a shrinkage prior on the basis coefficients. The computation burden of the Bayesian procedures can be prohibitive for data exploration and model building even for moderate data sets, which has contributed to the slow adoption of Bayesian methods in functional data analysis. As an example, a comparison of the Bayesian penalized spline method in Baladandayuthapani *et al*. (2007) to a method based on functional principal component analysis on simulated data found computation times of 5 h *versus* 5 s (Staicu *et al*., 2010).

This paper presents several methodological advancements. We develop a Bayesian framework for penalized spline function-on-scalar regression, allowing the joint modelling of population level fixed effects, subject level random effects and residual covariance. Dramatic computational improvements compared with the fully Bayesian and, surprisingly, with a frequentist mixed model approach are obtained through a variational Bayes approximation that is fast and accurate. This algorithm enables model selection and comparison, which for large data sets is infeasible with competing approaches. Novelly in multilevel function-on-scalar regression, we consider bivariate functional data as the outcome of interest. Finally, the size and structure of the motivating data set—which consists of nearly 20 000 trajectories, nested within subjects and depending on target, severity of impairment and affected hand as covariates—is unique in the functional data analysis literature.

The remainder of the paper is organized as follows. We discuss the model formulation and the variational Bayes approximation in Section 2. In Section 3 we conduct simulations designed to mimic the motivating data. Section 4 presents the analysis of the complete data set. We close with a discussion in Section 5. The Web-based supporting materials present the complete Gibbs sampler and variational Bayes algorithm. R implementations of all proposed methods and complete simulation code are available from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Methods

We begin by focusing on a simplification of model (1) for univariate functional data in Sections 2.1, 2.2 and 2.3. These sections develop our methodology for estimating fixed effects, random effects and residual covariance when a single functional response is observed. Once this has been established, we consider the bivariate model in Section 2.4.

2.1. Full model

For now, assume that the data are $[Y_{ij}(t), \mathbf{w}_i]$ for subjects $i = 1, \ldots, I$ and visits $j = 1, \ldots J_i$, giving a total of $n = \Sigma_i J_i$ functional observations. Univariate functional outcomes $Y_{ij}(t)$ are observed on a regular grid of length D for all subjects and visits. We pose the outcome model

$$\mathbf{y}_{ij} = \mathbf{w}_i\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{b} + \boldsymbol{\epsilon}_{ij} \qquad \boldsymbol{\epsilon}_{ij} \sim N(0, \Sigma) \tag{2}$$

where $\mathbf{y}_{ij}$ is the $1 \times D$ observed functional outcome; $\mathbf{w}_i$ and $\mathbf{z}_{ij}$ are fixed and random-effect design vectors of size $1 \times p$ and $1 \times I$ respectively; $\boldsymbol{\beta}$ and $\mathbf{b}$ are fixed and random-effect coefficient matrices of size $p \times D$ and $I \times D$ respectively; and $\boldsymbol{\epsilon}_{ij}$ is a $1 \times D$ vector of residual curves distributed $N(0, \Sigma)$ where Σ is a $D \times D$ covariance matrix. Errors $\boldsymbol{\epsilon}_{ij}$ are conditionally independent given fixed effects and subject-specific random effects, and are independently and identically distributed across subjects and visits.

We express the functional effects in the rows of $\boldsymbol{\beta}$ and $\mathbf{b}$ by using a spline expansion. Let Θ denote a $D \times K_\theta$ cubic B-spline evaluation matrix with K_θ basis functions. Further let B_W and B_Z denote the matrices whose columns are basis coefficients for $\boldsymbol{\beta}$ and $\mathbf{b}$ respectively, so that $\boldsymbol{\beta} = (\Theta B_W)^{\mathrm{T}}$ and $\mathbf{b} = (\Theta B_Z)^{\mathrm{T}}$. (This notation is drawn from (Ramsay and Silverman (2005), section 13.4.3, and Reiss *et al.* (2010); readers should note potential conflicts in notation, particularly with Montagna *et al.* (2012) and references therein.)

Penalization is a commonly used technique to avoid overfitting and to induce smoothness in functional effects. For spline coefficients in the kth column $B_{W,k}$ of B_W and ith column $B_{Z,i}$ of B_Z, we assume that $B_{W,k} \sim N(0, \sigma^2_{W,k} P^{-1})$ and $B_{Z,i} \sim N(0, \sigma^2_Z P^{-1})$. P is a known penalty matrix shared across fixed and random effects to enforce a common penalty structure. Variances $\{\sigma^2_{W,k}\}$ are unique to each coefficient function to allow unique levels of smoothness, and σ^2_Z is shared across random effects so that they are draws from a common population. Notationally, the 'zeroth' column $B_{W,0}$ of B_W and the variance $\sigma^2_{W,0}$ correspond to the intercept $\beta_0(t)$. The connection between this prior specification and a quadratic roughness penalty is well known; see Ruppert *et al.* (2003), chapter 4.9, for a detailed treatment. The choice of penalty matrix P is discussed in Section 2.3. In this penalized spline framework, the number and position of knots are typically unimportant provided that the number is sufficient to model the complexity of the coefficient functions (Ruppert, 2002). Using this specification, model (2) can be expressed as

$$\left.\begin{aligned} \mathbf{y}_{ij} &= \mathbf{w}_i B_W^{\mathrm{T}} \Theta^{\mathrm{T}} + \mathbf{z}_{ij} B_Z^{\mathrm{T}} \Theta^{\mathrm{T}} + \boldsymbol{\epsilon}_{ij}, \\ \boldsymbol{\epsilon}_{ij} &\sim N(0, \Sigma), \\ B_{Z,i} &\sim N(0, \sigma_Z^2 P^{-1}) \qquad \text{for } i = 1, \ldots, I, \\ B_{W,k} &\sim N(0, \sigma_{W,k}^2 P^{-1}) \qquad \text{for } k = 0, \ldots, p, \end{aligned}\right\} \tag{3}$$

which has a form that is similar to a traditional mixed model.

The variance components $\{\sigma_{W,k}^2\}$ and σ_Z^2 are assigned inverse gamma priors, and our model specification is completed by using an inverse Wishart prior for the residual covariance matrix Σ. Although these priors are convenient for the development of a straightforward Gibbs sampler and for the derivation of the variational approximation in Section 2.2, they have been criticized by several researchers (Gelman, 2006; Yang and Berger, 1994). Inverse gamma priors can be sensitive to the choice of hyperparameters a and b, especially for 'uninformative' values like $a = b = 0.001$ that place a large prior mass near zero, and inverse Wishart priors may not shrink eigenvalues as expected. In Section 2.3 we suggest using the data to help to choose hyperparameters in a reasonable way, but we emphasize the need for sensitivity analyses for these choices.

Following Gelfand *et al.* (1995), we pursue an alternative parameterization of model (3) to improve sampling performance through hierarchical recentring. For a simple example of this idea, consider the (non-functional) model $y_{ij} = \mu + b_i + \epsilon_{ij}$ with priors for μ and b_i having mean 0; alternatively one could let $Y_{ij} = \eta_i + \epsilon_{ij}$ with the prior for η_i having mean α and the prior for α having mean 0. The latter parameterization often results in better behaviour of the posterior chains and increased identifiability of fixed and random effects. For our function-on-scalar regression model, let Y be an $n \times D$ matrix of row-stacked functional outcomes, Z be the random-effects design matrix, W be the fixed effects matrix constructed by row-stacking the $\mathbf{w}_i$ and '$\otimes$' represent the Kronecker product operator. Using hierarchical recentring, we reparameterize model (3) by using

$$\left.\begin{aligned} Y &= Z B_Z^{\mathrm{T}} \Theta^{\mathrm{T}} + \epsilon, \\ \epsilon &\sim N(0, \Sigma \otimes I_n) \qquad \Sigma \sim \mathrm{IW}(\nu, \Psi), \\ B_{Z,i} &\sim N(B_W \mathbf{w}_i^{\mathrm{T}}, \sigma_Z^2 P^{-1}) \qquad \text{for } i = 1, \ldots, I,\ \sigma_Z^2 \sim \mathrm{IG}(a_Z, b_Z), \\ B_{W,k} &\sim N(0, \sigma_{W,k}^2 P^{-1}), \qquad \sigma_{W,k}^2 \sim \mathrm{IG}(a_{W,k}, b_{W,k}) \quad \text{for } k = 1, \ldots, p. \end{aligned}\right\} \tag{4}$$

Full conditionals for all model parameters are straightforward to obtain by using vector notation and Kronecker products. For matrix M and vector c, let $\mathrm{vec}(M)$ be the vector that is formed by concatenating the columns of M and $\mathrm{diag}(c)$ be the matrix with elements of c on the main diagonal and 0 elsewhere. Further let 'rest' include both the observed data and all parameters that are not currently under consideration. As an example of the full conditional distributions resulting from model (4), it can be shown that

$$p\{\mathrm{vec}(B_W) | \mathrm{rest}\} \propto N(\mu_{B_W}, \Sigma_{B_W})$$

where

$$\Sigma_{B_W} = \left\{ \frac{1}{\sigma_Z^2} (W \otimes I_{K_\theta})^{\mathrm{T}} (I_I \otimes P)(W \otimes I_{K_\theta}) + \mathrm{diag}\left(\frac{1}{\sigma_{W,k}^2} \right) \otimes P \right\}^{-1}$$

and

$$\mu_{B_W} = \Sigma_{B_W} \left\{ \frac{1}{\sigma_Z^2} (W \otimes I_{K_\theta})^{\mathrm{T}} (I_I \otimes P) \mathrm{vec}(B_Z) \right\}.$$

Additionally, we have that

$$p(\sigma^2_{W,k}|\text{rest}) \propto \text{IG}\left(a_{W,k} + \frac{K_\theta}{2}, b_{W,k} + B^{\text{T}}_{W,k} P B_{W,k}\right).$$

Complete derivations of this and all other full conditional distributions are provided in the Web-based supplementary materials.

Model (4) contains a large number of parameters, particularly in the spline coefficient matrices B_W and B_Z. Typically the available data for estimation in the $n \times D$ outcome matrix Y will dwarf the number of parameters in B_W and B_Z, meaning that these can be well estimated. However, in some cases it may be necessary to use a low dimensional spline basis or other parametric approach for the estimation of fixed and random effects. This is possible by using a simplification of model (4), but it should be noted that the choice of parametric form will be an implicit tuning parameter that replaces the explicit penalization implemented above. When the number of observed points per curve D is large the covariance matrix Σ will also be large, and the number of parameters in this matrix could be substantial; again, a modification of model (4) to use a parametric form for Σ is possible and perhaps advisable in this case, although the *caveats* regarding the introduction of a parametric structure still apply. In our motivating data set n is large (about 20 000) whereas D is modest (50 for bivariate curves), and model (4) is reasonable.

2.2. Variational Bayes methods

Variational Bayes methods are regularly used in the computer science literature, and to a more limited extent in the statistics literature, to provide approximate solutions to intractable inference problems (Jordan, 2004; Jordan *et al.*, 1999; Titterington, 2004; Ormerod and Wand, 2012). These tools have also been used somewhat rarely in functional data analysis (Goldsmith *et al.*, 2011; McLean *et al.*, 2013; van der Linde, 2008). Here we review variational Bayes methods only as much as needed to develop an iterative algorithm for approximate Bayesian inference in penalized function-on-scalar regression; for a more detailed overview see Ormerod and Wand (2010) and Bishop (2006), chapter 10. We emphasize that the variational Bayes approach is not intended to supplant a more complete Markov chain Monte Carlo sampler but rather is an appealing computationally efficient approximation that is useful for model building and data exploration.

Let $\mathbf{y}$ and ϕ represent respectively the full data and parameter collection. The goal of variational Bayes methods is to approximate the full posterior $p(\phi|\mathbf{y})$ by using $q(\phi)$, where q is restricted to a class of functions more tractable than the full posterior distribution. From the restricted class of functions, we wish to choose the element q^* that minimizes the Kullback–Leibler distance from $p(\phi|\mathbf{y})$. Divergence between $p(\phi|\mathbf{y})$ and $q(\phi)$ is measured by using

$$L_q = \int q(\phi) \log\left\{\frac{p(\mathbf{y}, \phi)}{q(\phi)}\right\} \text{d}\phi,$$

the q-specific lower bound on the marginal log-likelihood $\log\{p(\mathbf{y})\}$; maximizing L_q across the class of candidate functions gives the best possible approximation to the full posterior distribution. To make the approximation tractable, the candidate functions $q(\phi)$ are products over a partition of ϕ, so that $q(\phi) = \Pi_{l=1}^{L} q_l(\phi_l)$, and each q_l is a parametric density function. It can be shown that the optimal q_l^*-densities are given by

$$q_l^*(\phi_l) \propto \exp[E_{\phi_{-l}} \log\{p(\mathbf{y}, \phi)\}] \propto \exp[E_{\phi_{-l}} \log\{p(\phi_l|\text{rest})\}]$$

where, again, $\text{rest} \equiv \{\mathbf{y}, \phi_1, \ldots, \phi_{l-1}, \phi_{l+1}, \ldots, \phi_L\}$ is the collection of all remaining parameters and the observed data. In practice, we set initial values for each of the ϕ_l and update the respective optimal densities iteratively, similarly to a Gibbs sampler, while monitoring the q-specific lower bound L_q for convergence.

For the function-on-scalar regression model (4), we assume that

$$q(B_Z, B_W, \sigma^2_{W,0}, \ldots, \sigma^2_{W,p}, \sigma^2_Z, \Sigma) = q(B_Z)\, q(B_W)\, q(\sigma^2_{W,0}, \ldots, \sigma^2_{W,p}, \sigma^2_Z, \Sigma)$$

where the functions q are distinguished by their argument rather than by subscript l. The additional factorization

$$q(B_Z, B_W, \sigma^2_{W,0}, \ldots, \sigma^2_{W,p}, \sigma^2_Z, \Sigma) = \left\{ \prod_{i=1}^{I} q(B_{Z,i}) \right\} q(B_W) \left\{ \prod_{k=0}^{p} q(\sigma^2_{W,k}) \right\} q(\sigma^2_Z)\, q(\Sigma)$$

is induced by the conditional independence properties of the joint distribution (see Bishop (2006), section 10.2.5; a directed acyclic graph of our model appears in Fig. A.1). Using this factorization, it can be shown that the optimal density $q^*\{\mathrm{vec}(B_W)\}$ is $N(\mu_{q(B_W)}, \Sigma_{q(B_W)})$, where

$$\Sigma_{q(B_W)} = \{\mu_{q(1/\sigma^2_Z)} (W \otimes I_{K_\theta})^{\mathrm{T}} (I_I \otimes P)(W \otimes I_{K_\theta}) + \mathrm{diag}(\mu_{q(1/\sigma^2_{W,k})}) \otimes P\}^{-1}$$

and

$$\mu_{q(B_W)} = \Sigma_{q(B_W)} \{\mu_{q(1/\sigma^2_Z)} (W \otimes I_{K_\theta})^{\mathrm{T}} (I_I \otimes P)\, \mathrm{vec}(\mu_{q(B_Z)})\}.$$

In this equation, the notation $\mu_{q(\phi)}$ and $\Sigma_{q(\phi)}$ indicates the mean and variance of the density $q(\phi)$. Thus, the optimal density $q^*\{\mathrm{vec}(B_W)\}$ is normal with mean and variance completely determined by the data and the parameters of the remaining densities. Similar expressions are obtained for all model parameters. Together, these forms suggest an iterative algorithm in which each density is updated in turn by using the parameters from the remaining densities; convergence of this algorithm is monitored through L_q, the q-specific lower bound of the marginal log-likelihood. The iterative algorithm and the form for L_q are provided in the Web-based supplementary materials.

2.3. Choice of penalty matrix, hyperparameters and initial values

In both the Gibbs sampler that was presented in Section 2.1 and density updates for the variational Bayes algorithm described in Section 2.2, it is not necessary that the penalty matrix P be of full rank: although this introduces improper priors for the functional effects, the posteriors are proper. However, the lower bound L_q, which is used to monitor convergence of the variational Bayes algorithm, contains a term of the form $\log(|P^{-1}|)$, thus requiring P to be full rank. For this reason, we propose to use $P = \alpha P_0 + (1-\alpha) P_2$, where P_0 and P_2 are the matrices corresponding to zeroth- and second-derivative penalties. The P_2 penalty matrix is commonly used in functional data analysis and enforces smoothness in the estimated function but is non-invertible. The P_0 penalty matrix is the identity matrix, induces general shrinkage and is full rank. Details on the construction of P_0 and P_2 are available in Eilers and Marx (1996). Selecting $0 < \alpha \leqslant 1$ balances smoothness and shrinkage, and results in a full rank penalty matrix. There can be some sensitivity to the choice of α, with large values shrinking estimates toward 0; we recommend a relatively small value ($\alpha = 0.01$ or smaller) in keeping with the tendency to enforce smoothness rather than shrinkage.

We use the following procedure based on model (4) to choose hyperparameters. First, we estimate B_Z by using ordinary least squares from the regression $E[Y] = Z B_Z^{\mathrm{T}} \Theta^{\mathrm{T}}$ to obtain B_Z^{OLS}. We estimate the error covariance Σ by using a functional principal components decomposition of the residuals from this regression to obtain $\hat{\Sigma}$ (Yao *et al.*, 2005), and we use $\nu = \Sigma_i J_i$ and $\Psi = \Sigma_i J_i \hat{\Sigma}$ as hyperparameters in the prior for Σ. Next, we estimate B_W by using weighted least squares

from the regression $E[(B_Z^{\mathrm{OLS}})^{\mathrm{T}}] = WB_W^{\mathrm{T}}$ with weight matrix P^{-1} to obtain B_W^{WLS}. We choose

$$a_z = \frac{IK_\theta}{2},$$

$$b_z = \tfrac{1}{2}\,\mathrm{tr}\{((B_Z^{\mathrm{OLS}})^{\mathrm{T}} - W(B_W^{\mathrm{WLS}})^{\mathrm{T}})^{\mathrm{T}}\, P((B_Z^{\mathrm{OLS}})^{\mathrm{T}} - W(B_W^{\mathrm{WLS}})^{\mathrm{T}})\}$$

and

$$a_{W,k} = \frac{K_\theta}{2},$$

$$b_{W,k} = \tfrac{1}{2}(B_{W,k}^{\mathrm{WLS}})^{\mathrm{T}} P(B_{W,k}^{\mathrm{WLS}})$$

motivated by the form of the full conditionals for these variance components.

This procedure avoids the default 'uninformative' choice $a_z = b_z = a_{W,k} = b_{W,k} = 0.001$, which places a large prior mass near zero for all variance components. Such a choice favours over-shrinkage of fixed and random effects towards zero. In our application, we found that using this default tended to result in the incorporation of subject level random effects in the error variance Σ. Sensitivity to the choice of tuning parameters should be assessed in each application and weakly informative priors used where possible; see Section 4.2 and the Web-based supplementary materials for details of a sensitivity analysis for our application.

Initial values are sampled from an $N(0, 100)$ distribution for spline coefficients and from a uniform$(0.1, 10)$ distribution for variance components. The starting value for Σ is $\sigma^2 I_D$ where σ^2 is uniform$(0.1, 10)$ distributed. For other applications, different starting values may be needed.

2.4. Bivariate data

In the preceding discussion we have focused on a univariate outcome for clarity of exposition while introducing methods. In this section we describe the bivariate outcome model. Only straightforward modifications to the Gibbs sampler and variational Bayes updates given in Sections 2.1 and 2.2 respectively are needed for this setting. Similar extensions to three or more curves observed over a common domain proceed similarly.

Let $Y = [Y_1 Y_2]$ be the concatenation of two outcome matrices Y_1 and Y_2 (in our example, we concatenate the X- and Y-position curves so that $Y = [P^X P^Y]$). Next, let $B_W^{\mathrm{T}} = (B_{1,W}^{\mathrm{T}}, B_{2,W}^{\mathrm{T}})$ so that the columns of B_W concatenate the fixed effect spline coefficients for Y_1 and Y_2. Similarly, let $B_Z^{\mathrm{T}} = (B_{1,Z}^{\mathrm{T}}, B_{2,Z}^{\mathrm{T}})$ so that the columns of B_Z concatenate the random-effect spline coefficients for Y_1 and Y_2. Then $B_Z^{\mathrm{T}}(I_2 \otimes \Theta^{\mathrm{T}})$ contains subject-specific random effects for the overall outcome matrix Y. For spline coefficients in columns $B_{W,k}$ we assume that $B_{W,k} \sim N\{0, \mathrm{diag}(\sigma^2_{1,W_k}, \sigma^2_{2,W_k}) \otimes P^{-1}\}$, so that fixed effects for Y_1 and Y_2 are penalized separately and assumed independent *a priori*. A similar specification is used for the columns $B_{Z,i}$, and variance components are assigned inverse gamma priors. Again using a hierarchical formulation, for bivariate data our complete model is

$$\left.\begin{aligned}
Y &= ZB_Z^{\mathrm{T}}(I_2 \otimes \Theta^{\mathrm{T}}) + \epsilon, \\
\epsilon &\sim N(0, \Sigma \otimes I_n) \qquad \Sigma \sim \mathrm{IW}(\nu, \Psi), \\
B_{Z,i} &\sim N\{B_W \mathbf{w}_i^{\mathrm{T}}, \mathrm{diag}(\sigma^2_{1,Z,}, \sigma^2_{2,Z}) \otimes P^{-1}\} \qquad \text{for } i = 1, \ldots, I, \\
\sigma^2_{m,Z} &\sim \mathrm{IG}(a_{m,Z}, b_{m,Z}) \qquad \text{for } m = 1, 2, \\
B_{W,k} &\sim N\{0, \mathrm{diag}(\sigma^2_{1,W_k}, \sigma^2_{2,W_k}) \otimes P^{-1}\}, \\
\sigma^2_{m,W_k} &\sim \mathrm{IG}(a_{m,W_k}, b_{m,W_k}) \qquad \text{for } m = 1, 2 \text{ and } k = 1, \ldots, p.
\end{aligned}\right\} \tag{5}$$

This model extends the univariate outcome model (4), but the Gibbs sampler and variational Bayes approximations can be directly modified for bivariate data. Similarly, the method for setting initial values and choosing hyperparameters that was given in Section 2.3 can be adapted to model (5).

This bivariate model is motivated by the data structure, in which X- and Y-position curves are observed concurrently for all motions—in that sense, the bivariate model is more faithful to the observed data than fitting separate univariate models. A bivariate model also explicitly models correlation in X- and Y-position error curves. In our application, this correlation may provide insight into sensory or visual feedback in reaching motions, or into the biomechanical processes that are involved. Nonetheless, it is possible to fit a bivariate model by using two separate univariate models, especially if X- and Y-errors are uncorrelated or if this correlation is not scientifically meaningful.

3. Simulations

We demonstrate the performance of our method by using a simulation in which generated data mimic the motivating application; all code needed to reproduce these simulations is available from `http://wileyonlinelibrary.com/journal/rss-datasets`. Our simulations consider three groups: control subjects' dominant hand, the affected dominant hand of moderately affected stroke patients and the affected dominant hand of severely affected stroke patients. We therefore created a three-level categorical predictor with groups 1, 2 and 3 referring to controls, moderately affected subjects and severely affected subjects respectively.

Curves are observed on a common grid of length $D = 25$. Data are generated from the univariate outcome model $\mathbf{y}_{ij} = \mathbf{w}_i\boldsymbol{\beta} + \mathbf{b}_i + \boldsymbol{\epsilon}_{ij}$ where $\mathbf{w}_i$ is a length 3 binary vector indicating group membership for subject i, $\boldsymbol{\beta}$ is a $3 \times D$ matrix whose rows are group average curves, $\mathbf{b}_i$ is a length D random effect and $\boldsymbol{\epsilon}_{ij}$ is a residual vector. For each simulated data set, the predictors $\mathbf{w}_i$ are sampled from a multinomial distribution with probabilities set to the proportions of each group in the motivating data, random effects $\mathbf{b}_i$ are drawn from an $N(0, \Sigma^{\mathbf{b}})$ distribution and residuals $\boldsymbol{\epsilon}_{ij}$ are drawn from an $N(0, \Sigma^{\epsilon})$ distribution.

The quantities $\boldsymbol{\beta}$, $\Sigma^{\mathbf{b}}$ and Σ^{ϵ} are chosen to resemble our motivating data. First, we focus on a subset of the full data set that consists of the observed Y-position curves from reaches to the target at 180°. In this subset, we find group level average curves for each of the groups of interest; these become the coefficient functions in the rows of $\boldsymbol{\beta}$. For each subject in our subset, we find the subject level average curve and subtract the corresponding group level mean; then we calculate the covariance $\Sigma^{\mathbf{b}}$ of these curves. Finally, for each curve in our subset we subtract the subject level mean; then we calculate the covariance Σ^{ϵ} of these curves. The number of subjects I is set to

(a) 60,
(b) 120 or
(c) 180.

In all cases, we fix the number of observations per subject to be $J_i = 5$. To illustrate the simulation design, Fig. 2(a) shows the coefficient functions. Fig. 2(b) shows a complete simulated data set with $I = 60$ and highlights data for three subjects.

For each sample size we generate 100 data sets. Parameters are estimated by using the Gibbs sampler and variational Bayes algorithm that were described in Sections 2.1 and 2.2 respectively, with hyperparameters and initial values chosen as in Section 2.3. For the Gibbs sampler, we used chains of length 5000 and discarded the first 1000 as burn-in. To provide a frame of

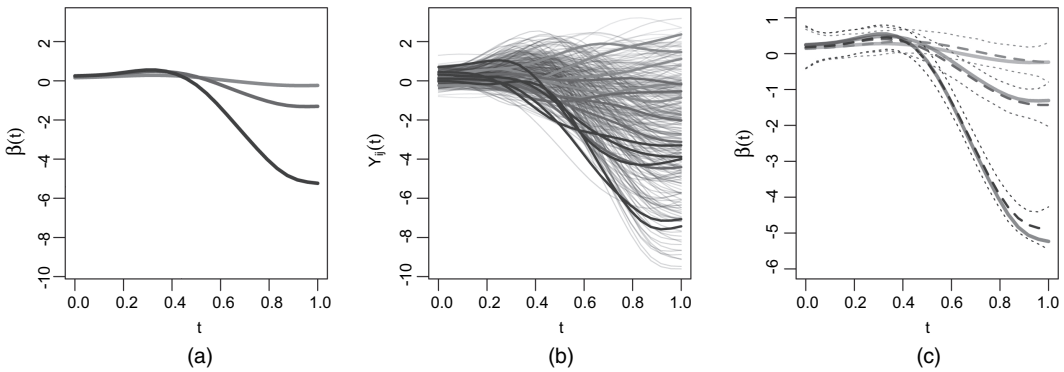

Fig. 2. (a) Three coefficient functions used to simulate data (, $\beta_0(t)$; , $\beta_1(t)$; , $\beta_2(t)$), (b) completed simulated data sets, with three subjects (one from each group) highlighted and (c) true coefficient functions, as well as their estimates and credible intervals derived from the data set in (b)

reference for our methods, we compare with the `pffrGLS()` function in the `refund` R package. This extends the penalized function-on-function regression model assuming independent errors implemented in `pffr()` (Scheipl *et al.*, 2015); generalized least squares is used to account for residual correlation in `pffrGLS()` and a mixed model framework is used for parameter estimation. In a process similar to the generalized least squares method described by Reiss *et al.* (2010) for cross-sectional function-on-scalar regression, we first fit the model assuming independence and use the residual curves to estimate the covariance matrix for use in `pffrGLS()`. Although it is not described anywhere at the time of writing, to the best of our knowledge `pffrGLS()` represents the current state of the art in function-on-scalar regression with subject level random effects.

Fig. 3(a) show the integrated mean-squared error $\text{IMSE} = \int \{\hat{\beta}(t) - \beta(t)\}^2 \, \mathrm{d}t$ for each coefficient function, estimation method and sample size. IMSE is indistinguishable for the Gibbs sampler and variational Bayes approaches, indicating that for posterior means the variational Bayes approximation is reasonable. As expected, IMSE decreases as the sample size increases. Both approaches are comparable with or outperform the mixed model approach, sometimes substantially. Fig. 3(b) shows the computation time for each sample size and approach (simulations were executed in parallel on a compute cluster with Intel Xeon central processsor units running at 2.30 GHz; memory usage was 2, 4 and 6 Gbytes for $I = 60$, 120, 180 respectively). Not surprisingly, the variational Bayes algorithm is substantially faster than the complete Gibbs sampler. However, there are also meaningful improvements in computation time comparing the variational Bayes algorithm with the mixed model: for $i = 180$, the median computation time for the variational Bayes approach was roughly 15 s, whereas the median computation time for the mixed model was nearly 2 h. This discrepancy in computation time is in part due to the need to fit two models (one using `pffr()` and one using `pffrGLS()`) for the mixed model; also, the code for the variational Bayes algorithm is tailored to the model at hand whereas `pffrGLS()` uses the more general `mgcv` package for estimating parameters. However, we also note that our implementation of `pffrGLS()` used five (rather than 10) basis functions. For $I = 180$, using 10 basis functions in `pffrGLS()` required 20 000 s—roughly 3.5 times longer than the Gibbs sampler.

Table 1 presents the average coverage probability of 95% pointwise confidence intervals constructed by using the Gibbs sampler, the variational Bayes algorithm and the frequentist mixed model for each coefficient function and sample size. For both proposed Bayesian approaches,

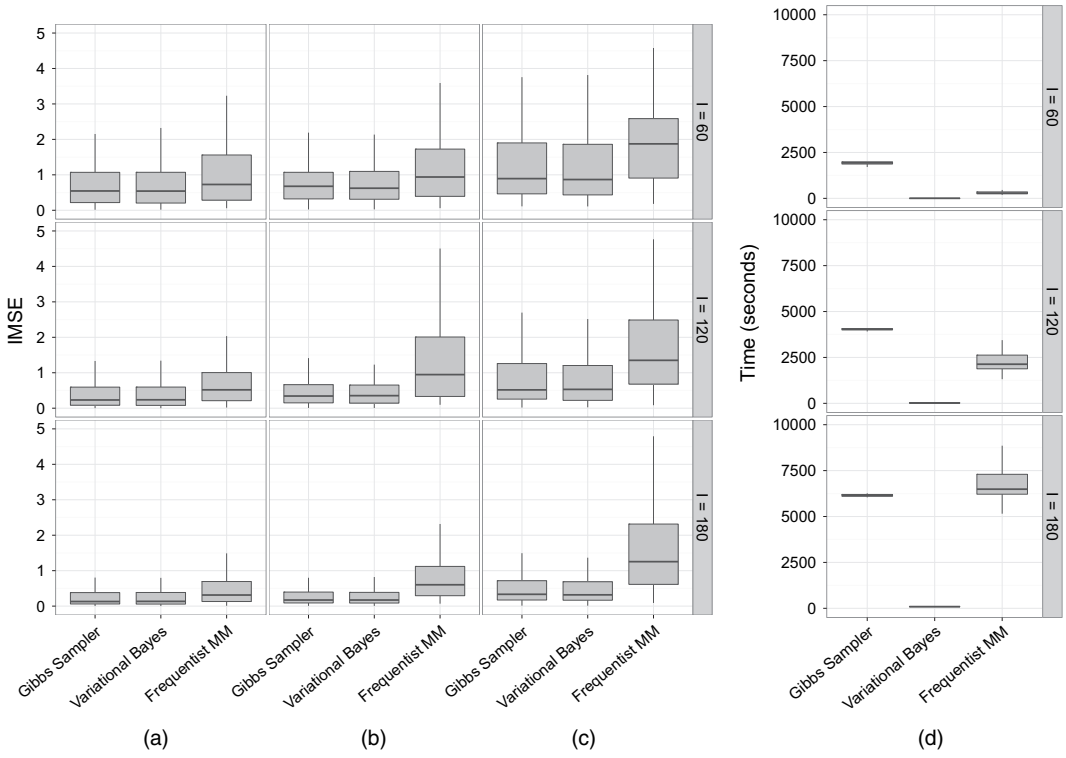

Fig. 3. Simulation results: (a)–(c) IMSE, defined as IMSE = $\int\{\hat{\beta}(t) - \beta(t)\}\mathrm{d}t$, for each coefficient function, sample size and estimation technique; (d) computation time for each sample size and estimation method

Table 1. Average coverage of 95% credible intervals constructed by using the Gibbs sampler, the variational Bayes algorithm and pffr()

I	*Coverages for the following methods:*								
	Gibbs sampler			*Variational Bayes*			*Frequentist mixed model*		
	$\beta_0(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_0(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_0(t)$	$\beta_1(t)$	$\beta_2(t)$
60	0.98	0.98	0.98	0.97	0.97	0.96	0.43	0.47	0.43
120	0.98	0.99	0.98	0.97	0.99	0.96	0.37	0.34	0.34
180	0.97	0.99	0.97	0.96	0.98	0.96	0.40	0.32	0.27

the coverage slightly exceeds the nominal level and is often between 0.96 and 0.98. As expected, the coverage improves as the sample size increases. For the mixed model the coverage is well below nominal levels for all coefficients and sample sizes; it is possible that the code is still under development and that future iterations may provide better inference. Although the coverage for the variational Bayes approximation is reasonable in our simulations, we do not necessarily recommend basing inference in practice on this approach owing to the difficulty in verifying the assumptions regarding the factorization of the posterior distribution. Rather, we favour the

variational algorithm as a fast method for model building and we base inference on a full Gibbs sampler.

4. Application

We now apply the developed methods to the motivating data that were described in Section 1.1. In our data set affected patients exhibit arm paresis, which is a weakness or motor control deficit affecting either the dominant or non-dominant arm, due to a unilateral stroke. Patients experienced stroke more than 6 months before data collection, meaning that observed motor control deficits are not due to short-term effects but rather are chronic in nature. To quantify the severity of arm impairment we use the upper extremity portion of the Fugl-Meyer motor assessment, which is a well-known and widely used clinical assessment of motor impairment. Fugl-Meyer scores were assessed for the affected arm only and, for upper extremity testing, scores range from 0 to 66 with 66 indicating healthy function. Controls were not scored and were assigned a Fugl-Meyer score of 66. Kinematic data collected for the left hand were reflected through the Y-axis and thus are in the same intrinsic joint space as data for the right hand (i.e. motions to the target at 180° reach across the body and involve both the shoulder and the elbow).

Our focus is the effect of the severity of arm impairment on control of visually guided reaching, where impairment is quantified by using the Fugl-Meyer score. In addition to severity of impairment, we control for important covariates in our regression modelling. We adjust for target direction (with eight possible targets, treated as a categorical predictor) hand used (dominant and non-dominant), whether the arm is affected by stroke (affected and unaffected) and, potentially, interactions between these variables. Interactions of severity of impairment and other covariates are possible and are likely for target direction: the effect of stroke may be greatest for the more biomechanically difficult targets that involve co-ordination of multiple joints.

Our data analysis proceeds in two parts and focuses on estimation of the bivariate model (5). First, we use the variational Bayes algorithm that was developed in Section 2.2 to explore several possible models that include different combinations of target, hand used, affectedness and severity of impairment as well as potential interactions. In all models, subject level random effects are estimated for each target and hand; these effects are *a priori* assumed to be independent. The computational efficiency of the variational Bayes algorithm is crucial at this stage, allowing the fast evaluation and comparison of models. The assumptions that underlie the variational algorithm make it unsuitable for inference in our real data application, and comparisons are made on the basis of percentage variance explained. Therefore, after identifying a plausible final model, we estimate all model parameters by using the complete Gibbs sampler that was described in Section 2.1 and we base inference for the effect of stroke on this analysis.

4.1. Exploratory analyses using variational Bayes methods

In what follows, we are interested in estimated fixed effects by using a variety of structures for the population mean. We select hyperparameters as described in Section 2.3 and estimate models by using the variational approximation that was described in Section 2.2. The computation time was under 20 min for each model that we consider; the importance of fast computation in the model building stage cannot be understated, since it allows the consideration and refinement of many candidate models.

As a reference for the more complex structures that follow, we began with a model that uses only target direction as a predictor. This model addresses directional variation only, but the eight fixed effects account for roughly 90% of observed variance in the outcome. Following this,

Table 2. Description and comparison of models considered†

Model	*Fixed effects*	*Number of fixed effects*	*Relative PVE for fixed effects (%)*
0	Tar	8	Reference
1	FM+Tar	9	0.1
2	FM×Tar	16	4.6
3	FM^2×Tar	32	5.2
4	FM×Tar×Hand	32	8.4
5	FM×Tar×Aff	32	8.8
6	FM^2×Tar×Aff	64	10.1
7	FM×Tar×Hand×Aff	64	11.9

†Fixed effects structure is described in the second column, where 'Tar' represents the target direction (as a categorical variable), 'Hand' represents hand used (dominant or non-dominant), 'Aff' indicates an affected hand, 'FM' is the continuous Fugl-Meyer score, '+' indicates additive effects and '×' indicates interactions. The number of fixed effects induced by the model structure is given in the third column. The fourth column provides the percentage of outcome variance explained by the model relative to a model with only target as a covariate (defined in equation (6)).

several models that included the Fugl-Meyer score, hand used and affectedness as predictors were considered. Table 2 provides the fixed effects that were used in each of the models we consider, the number of fixed effects for each model and the percentage of outcome variance explained by fixed effects. Percentage variance explained is given relative to the target-only model by using

$$\text{relative PVE} = 100\left[1 - \frac{\text{var}\{\text{vec}(Y - \hat{Y}_m)\}}{\text{var}\{\text{vec}(Y - \hat{Y}_0)\}}\right] \tag{6}$$

for models $m \in 1, \ldots, 7$, where Y is the matrix of observed trajectories and $\hat{Y}_m$ is the matrix of estimated trajectories based on fixed effects in model m.

The poor performance of model 1 compared with model 2 indicates the importance of interaction between target and of severity of impairment due to the target-specific direction of the effect of stroke and to differing levels of biomechanical difficulty. Models 3, 4 and 5 build on model 2 by adding a quadratic effect of the Fugl-Meyer score and interactions with hand used and affectedness respectively. The relatively small improvement of the quadratic model may support an assumption of linearity, and the comparability of models 4 and 5 may be because, in our data set, the affected hand was more likely to be the dominant hand. Models 6 and 7 build on model 5 by allowing a quadratic effect of the Fugl-Meyer score and an interaction with the hand used. Again, the improvement from using a quadratic term is comparable with the improvement from the interaction. For all models, the fixed and random effects together explain roughly 50% of outcome variance; the remaining 50% is residual variance around subject level means. This partitioning of variance usefully quantifies the extent to which motor control is explainable by covariates, subject-specific deviations and trajectory level variation. In Section 4.2 we discuss inference for model 7.

Fig. 4 compares the estimated fixed effects from models 2, 5 and 7. For each model (in rows), we show the estimated mean trajectories in an affected dominant hand for Fugl-Meyer scores 66, 51, 36 and 21 (in columns). In model 2, the effect of increasing severity of stroke is assumed to be the same in both the affected and the unaffected hand. This is unlikely given

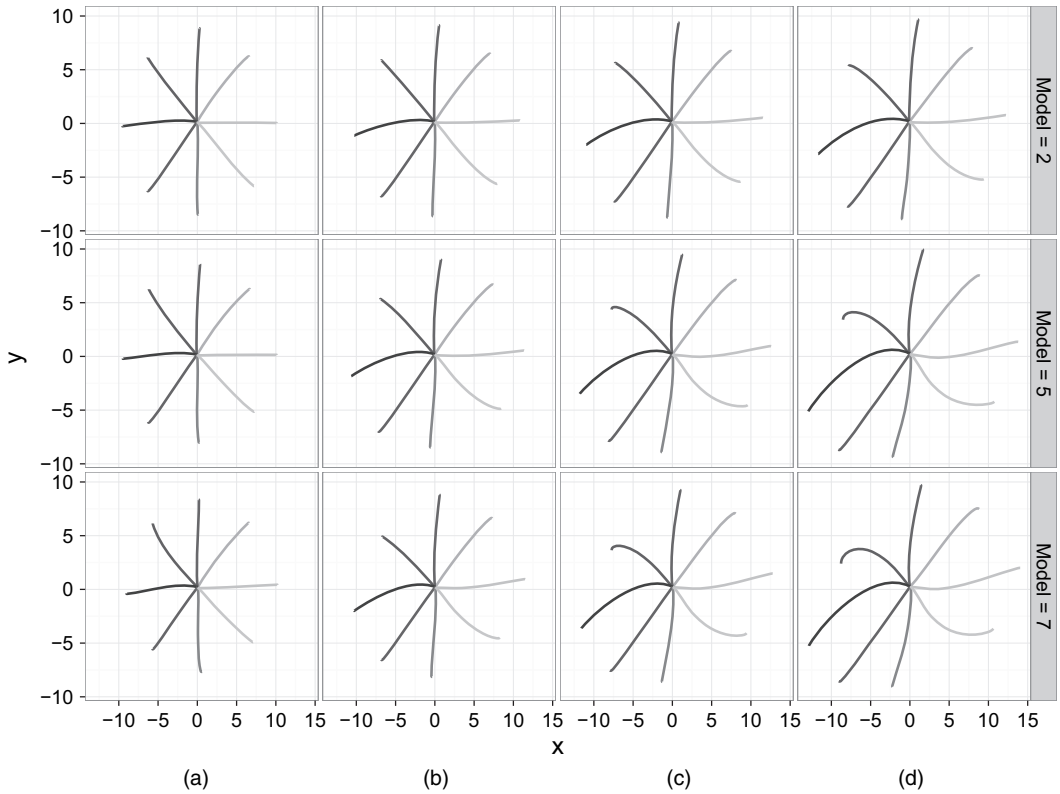

Fig. 4. Estimated motions for an affected dominant arm with Fugl-Meyer scores (a) 66, (b) 51, (c) 36 and (d) 21 under models 2, 5 and 7: an interactive version of this figure is available from the first author's Web site

that our data set consists of patients with unilateral stroke. Model 5 estimates separate effects of severity of stroke for the affected and unaffected arm. Comparing models 2 and 5 for an affected arm in Fig. 4, model 5 indicates large effects of increasing severity of stroke; for an unaffected arm (not shown) model 5 indicates a small or no effect. Model 7 additionally separates the affected dominant from affected non-dominant hands, with the scientific interpretation that a stroke of the same severity could affect these limbs differently. The differences between models 5 and 7 are subtle for affected dominant arms, but more noticeable for unaffected and non-dominant arms. An interactive version of Fig. 4 is available from the first author's Web site.

Estimates of subject level effects are shown in Fig. 5 for two subjects (separately by row) overlayed on the observed trajectories. Fixed effects estimates based on model 7 are shown in bold full curves and subject level estimates including random effects are shown in bold broken curves. In Figs 5(a)–5(c) is a control subject's dominant hand; fixed effects and random-effects estimates differ only slightly, indicating relatively little subject deviation from the population mean. In Figs 5(d)–5(f) is a severely affected (Fugl-Meyer score 28) subject's affected dominant hand. Here, fixed effects are noticeably curved for several targets, indicating a systematic effect of stroke. Subject level estimates differ from the fixed effects in some cases (particularly for targets at 0° and 180°), illustrating the idiosyncratic effects of stroke in this patient. Data for these patients are shown in Fig. 1.

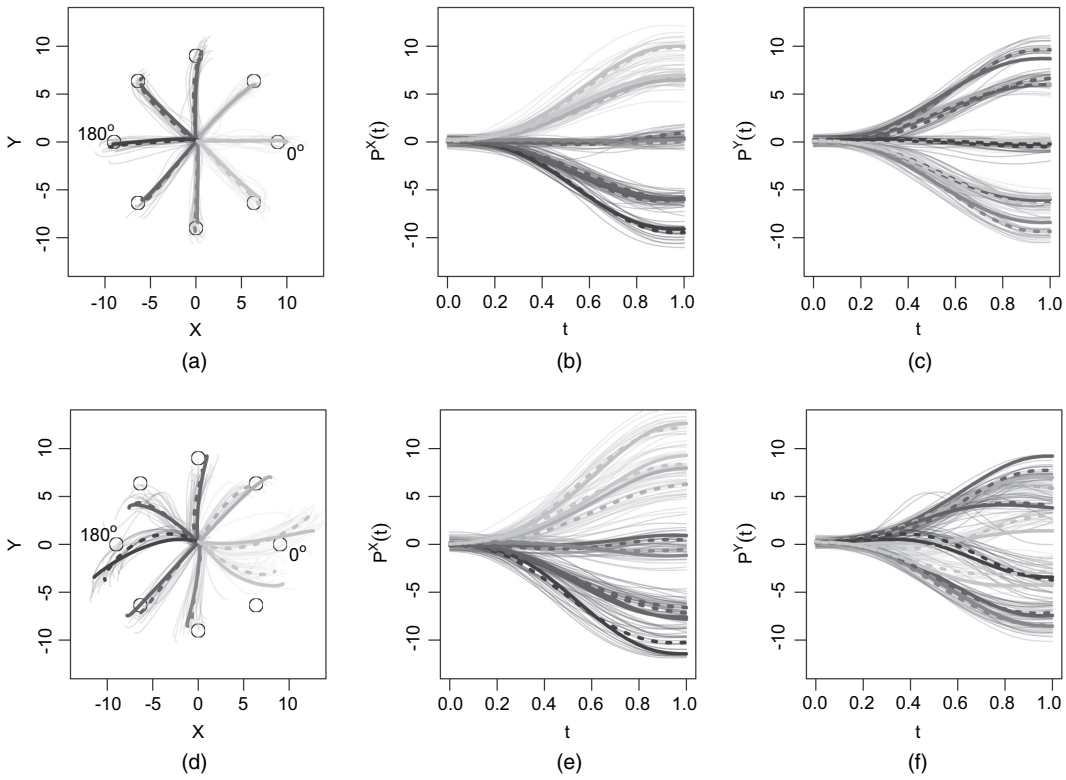

Fig. 5. Observed data (——) overlayed with estimated fixed (——) and random (– – –) effects (data for two subjects are shown: in (a)–(c) the dominant hand of a control subject, and in (d)–(f) the affected dominant hand of a severe stroke patient; data for these subjects appear in Fig. 1)

4.2. Full Bayesian analysis

After exploring several candidate fixed effects structures, we fit model 7 by using a fully Bayesian analysis to explore inferential properties of estimated coefficients. In particular, we are interested in the target- and hand-specific systematic effects of the Fugl-Meyer score as a continuous covariate. For our final model, we used five chains with random starting values, setting the chain length to 2000 iterations and discarding the first 500 as burn-in. Hyperparameters and initial values were chosen as described in Section 2.3. Analyses assessing the sensitivity to hyperparameter values appear in the Web-based supplementary materials. Computations took 2.5 days per chain on Intel Xeon central processor units running at 2.30 GHz with 8 Gbytes memory, which emphasizes the importance of a fast approximation for data exploration and model building.

Fig. 6 shows the estimated effect of a 10-unit decrease in Fugl-Meyer score in a dominant hand affected by stroke for all target directions. The top and bottom rows show the marginal effect on the X- and Y-position curves respectively. The panels show the posterior mean as a bold curve, and a sample from the posterior as light grey curves. These results show clear significant effects for all targets, verifying that increasing severity of stroke has a systematic effect on motor control. For motions to the target at 0°, increasing severity of stroke leads to overreach (increases in the X-direction) as well as a vertical shift (increase in the Y-position). Other targets can be interpreted similarly. The largest effects are generally in directions that require multijoint co-ordination and are thus more biomechanically difficult.

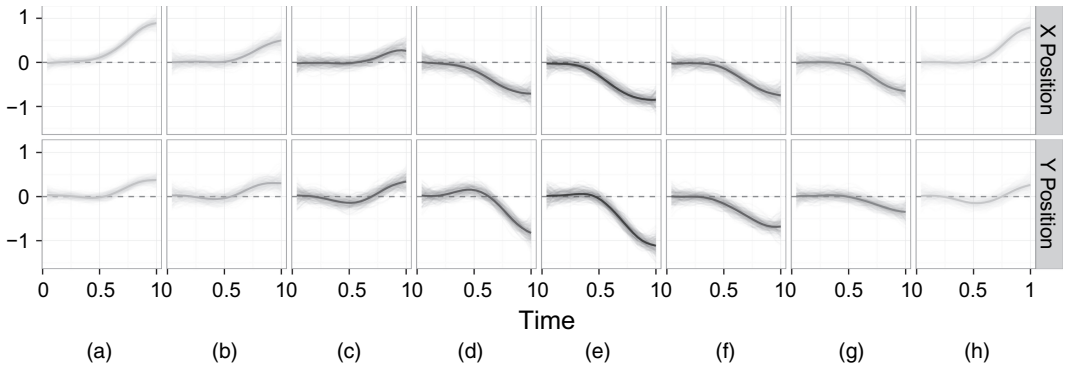

Fig. 6. Estimated effect of a 10-unit decrease in Fugl-Meyer score in an affected dominant hand (———, posterior means; ———, posterior sample): for (a) target 0, (b) target 45, (c) target 90, (d) target 135, (e) target 180, (f) target 225, (g) target 270 and (h) target 315, for the X- and Y-positions

Fig. 7 illustrates the quality and convergence of our Markov chains. Figs 7(a)–7(c) show plots for a representative spline coefficient in a fixed effect function $\beta_k(t)$; Figs 7(d)–7(f) show the variance component $\sigma_{W,k}$ that controls the degree of penalization in this function. Figs 7(a) and 7(d) show the five posterior chains for the parameter of interest, started from randomly chosen values with the burn-in period shaded; this shows that chains converge quickly and there is little sensitivity to the choice of starting value. An auto-correlation function is shown in Figs 7(b) and 7(e) and indicates low auto-correlation. Figs 7(c) and 7(f) show the convergence criterion of Gelman and Rubin (1992) as a function of iteration number. Values near 1 indicate convergence, which is typically attained after only a few hundred iterations. Finally, in Fig. 7(g) we show the posterior mean residual covariance surface to illustrate the correlation within X- and Y-position residuals and the correlation between them.

5. Concluding remarks

This paper has focused on the development of a regression framework for the analysis of kinematic data used to assess motor control in stroke patients. Our model allows flexible mean structures, subject level random effects, bivariate outcomes and correlated errors. We develop a hierarchical Bayesian estimation framework; crucially, a fast and accurate variational Bayes approximation to the full Gibbs sampler allows extensive data exploration and model building before estimation with the full Bayesian approach. Implementations of both approaches and complete simulation code are publicly available.

Variational approximations are often quite inaccurate for the construction of credible intervals because of the factorization assumptions and use of parametric densities to approximate the posterior distribution. In that sense, the good inferential performance that was found in simulations is perhaps surprising. In part this performance is due to the hierarchical recentring in model (4), in which observed curves are centred on subject effects, subject effects are centred on fixed effects and fixed effects are centred on zero. This formulation helps to decrease the posterior correlation between fixed and random effects and improves the quality of the approximation (and, importantly, the mixing for the Gibbs sampler). A direct implementation of model (3), which centres observed curves on the combination of fixed and random effects and centres both fixed and random effects near zero, had similar accuracy of estimation but much poorer inference. Despite these results, we recommend caution when using the variational

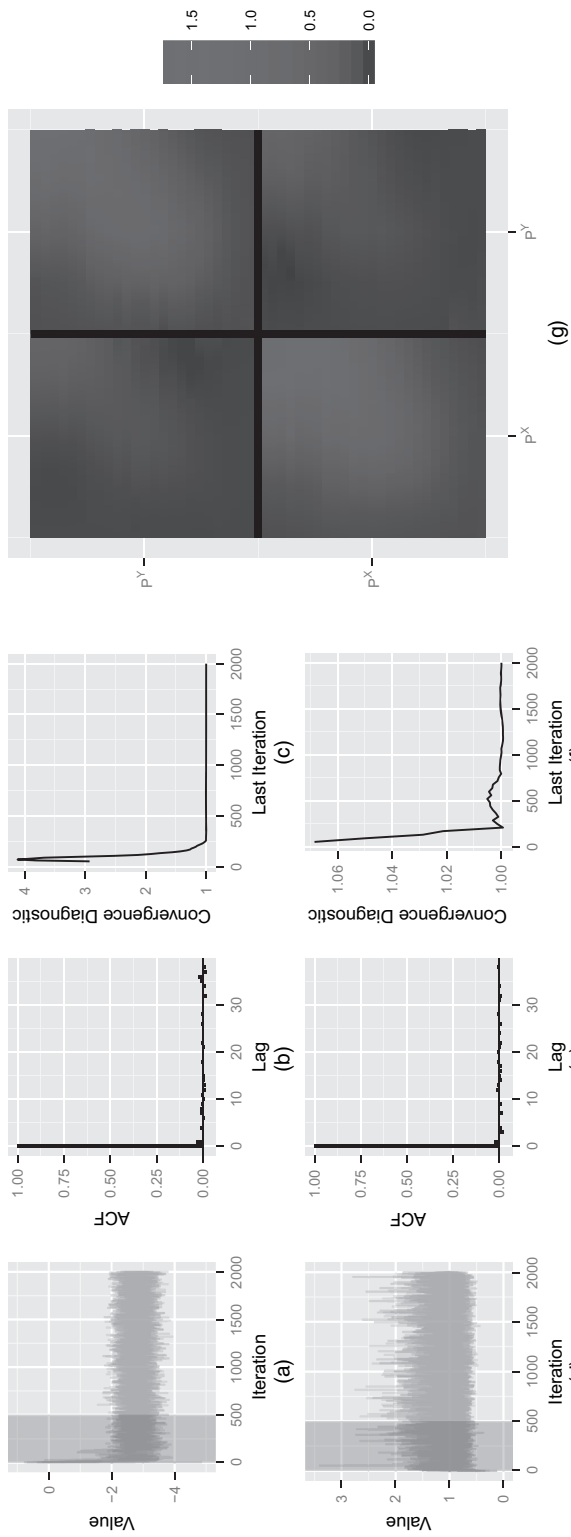

Fig. 7. (a)–(f) Diagnostic and convergence criteria for representative parameters, and (g) residual covariance surface: (a)–(c) spline coefficient in a fixed effect function; (d)–(f) variance component associated with the coefficient function; (a), (d) five chains with random starting values; (b), (e) auto-correlation function; (c), (f) Gelman and Rubin's (1992) diagnostic criterion

approximation for inference, and we prefer this method for model building, cross-validation of results, bootstrapping or other computationally intensive procedures. Finally, we note that the good performance of the variational approximation suggests that other approximate methods might be suitable for this model and should be explored carefully.

The application of our developed methodology to the motivating data yields novel insights into the effect of arm impairment on control of visually guided reaching. We demonstrate consistent systematic effects of stroke on reaching trajectories by using the Fugl-Meyer score as a continuous covariate that are direction dependent. Our final model indicates that roughly 10% of variability in observed trajectories is due to systematic effects of severity of impairment; subject-specific idiosyncrasies account for an additional 40%. Although not of primary concern here, our application also allows comparisons of dominant and non-dominant hand among controls, as well as consideration of systematic effects in the unaffected hand following stroke. Additional work will examine the replicability of these findings in larger studies and quantify possible overfitting.

Future work may take several directions. In statistical methodology, additional flexibility in the mean structure, for instance by allowing non-linear effects of covariates, could broaden the applicability of the model. Parameterizing the residual correlation structure as a function of severity of impairment would more accurately reflect the disease process. Implementing and testing our methods for the case that curves are observed on a sparse grid would broaden the class of problems to which these methods can be applied. New general purpose programming languages for Bayesian analysis (including sampling and optimization) will facilitate the implementation of more classes of prior distributions, potentially leading to better models and reducing the reliance on convenient priors (Stan Development Team, 2013). In the applied setting, extension to three-dimensional kinematics will be necessary as experiments allow more complex reaching motions. Longitudinal experiments to explore treatment effects and to describe the natural history of recovery are under way; accompanying methods will be needed to account for within-subject correlations over time.

Acknowledgements

We thank John Krakauer for his scientific insight and guidance, and Johnny Liang, Sophia Ryan and Sylvia Huang for their assistance in data collection. The first author's research was supported in part by award R01HL123407 from the National Heart, Lung, and Blood Institute and by award R21EB018917 from the National Institute of Biomedical Imaging and Bioengineering.

References

Baladandayuthapani, V., Ji, Y., Talluri, R., Nieto-Barajas, L. E. and Morris, J. S. (2010) Bayesian random segmentation models to identify shared copy number aberrations for array CGH data. *J. Am. Statist. Ass.*, **105**, 1358–1375.

Baladandayuthapani, V., Mallick, B., Young Hong, M., Lupton, J., Turner, N. and Carroll, R. J. (2007) Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics*, **64**, 64–73.

Bishop, C. M. (2006) *Pattern Recognition and Machine Learning*. New York: Springer.

Broderick, J. (2004) Stroke therapy in the year 2025: burden, breakthroughs, and barriers to progress. *Stroke*, **35**, 205–211.

Brumback, B. and Rice, J. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Am. Statist. Ass.*, **93**, 961–976.

Coderre, A. M., Zeid, A. A., Dukelow, S. P., Demmer, M. J., Moore, K. D., Demers, M. J., Bretzke, H., Herter, T. M., Glasgow, J. I., Norman, K. E., Bagg, S. D. and Scott, S. H. (2010) Assessment of upper-limb sensorimotor function of subacute stroke patients using visually guided reaching. *Neurorehab. Neurl Repr*, **24**, 528–541.

Crainiceanu, C., Reiss, P., Goldsmith, J., Huang, L., Huo, L. and Scheipl, F. (2012) refund: regression with functional data. *R Package Version 0.1-6*. (Available from `http://CRAN.R-project.org/package=refund`.)

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S. and Punjabi, N. M. (2009) Multilevel functional principal component analysis. *Ann. Appl. Statist.*, **4**, 458–488.

Eilers, P. H. C. and Marx, B. D. (1996) Flexible smoothing with B-splines and penalties. *Statist. Sci.*, **11**, 89–121.

Fugl-Meyer, A., Jääskö, L., Leyman, I., Olsson, S. and Steglind, S. (1974) The post-stroke hemiplegic patient: 1, A method for evaluation of physical performance. *Scand. J. Rehab. Med.*, **7**, 13–31.

Gelfand, A., Sahu, S. K. and Carlin, B. (1995) Efficient parameterizations for generalized linear mixed models. *Biometrika*, **82**, 479–488.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Baysn Anal.*, **1**, 515–533.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statist. Sci.*, **7**, 457–472.

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Borden, W. B., Bravata, D. M., Dai, S., Ford, E. S., Fox, C. S., Franco, S., Fullerton, H. J., Gillespie, C., Hailpern, S. M., Heit, J. A., Howard, V. J., Huffman, M. D., Kissela, B. M., Kittner, S. J., Lackland, D. T., Lichtman, J. H., Lisabeth, L. D., Magid, D., Marcus, G. M., Marelli, A., Matchar, D. B., McGuire, D. K., Mohler, E. R., Moy, C. S., Mussolino, M. E., Nichol, G., Paynter, N. P., Schreiner, P. J., Sorlie, P. D., Stein, J., Turan, T. N., Virani, S. S., Wong, N. D., Woo, D. and Turner, M. B. on behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee (2013) Heart disease and stroke statistics—2013 update: a report from the American Heart Association. *Circulation*, **127**, e6–e245.

Goldsmith, J., Greven, S. and Crainiceanu, C. M. (2013) Corrected confidence bands for functional data using principal components. *Biometrics*, **69**, 41–51.

Goldsmith, J., Wand, M. P. and Crainiceanu, C. M. (2011) Functional regression via variational Bayes. *Electron. J. Statist.*, **5**, 572–602.

Greven, S., Crainiceanu, C. M., Caffo, B. and Reich, D. (2010) Longitudinal functional principal component analysis. *Electron. J. Statist.*, **4**, 1022–1054.

Guo, W. (2002) Functional mixed effects models. *Biometrics*, **58**, 121–128.

Huang, V., Ryan, S., Kane, L., Huang, S., Berard, J., Kitago, T., Mazzoni, P. and Krakauer, J. (2012) 3D robotic training in chronic stroke improves motor control but not motor function. Society for Neuroscience, New Orleans.

Jordan, M. I. (2004) Graphical models. *Statist. Sci.*, **19**, 140–155.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999) An introduction to variational methods for graphical models. *Mach. Learn.*, **37**, 183–233.

Kitago, T., Liang, J., Huang, V. S., Hayes, S., Simon, P., Tenteromano, L., Lazar, R. M., Marshall, R. S., Mazzoni, P., Lennihan, L. and Krakauer, J. W. (2013) Improvement after constraint-induced movement therapy recovery of normal motor control or task-specific compensation? *Neurorehab. Neurl Repr*, **27**, 99–109.

Kwakkel, G., Kollen, B. J., van der Grond, J. and Prevo, A. J. H. (2003) Probability of regaining dexterity in the flaccid upper limb impact of severity of paresis and time since onset in acute stroke. *Stroke*, **34**, 2181–2186.

Lang, C. E., Wagner, J. M., Edwards, D. F., Sahrmann, S. A. and Dromerick, A. W. (2006) Recovery of grasp versus reach in people with hemiparesis poststroke. *Neurorehab. Neurl Repr*, **20**, 444–454.

Levin, M. F. (1996) Interjoint coordination during pointing movements is disrupted in spastic hemiparesis. *Brain*, **119**, 281–293.

van der Linde, A. (2008) Variational Bayesian functional PCA. *Computnl Statist. Data Anal.*, **53**, 517–533.

McLean, M. W., Scheipl, F., Hooker, G., Greven, S. and Ruppert, D. (2013) Bayesian functional generalized additive models for sparsely observed covariates. To be published.

Montagna, S., Tokdar, S. T., Neelson, B. and Dunson, D. B. (2012) Bayesian latent factor regression for functional and longitudinal data. *Biometrics*, **69**, 1064–1073.

Morris, J. S. and Carroll, R. J. (2006) Wavelet-based functional mixed models. *J. R. Statist. Soc.* B, **68**, 179–199.

Morris, J. S., Vannucci, M., Brown, P. J. and Carroll, R. J. (2003) Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *J. Am. Statist. Ass.*, **98**, 573–583.

Ormerod, J. and Wand, M. P. (2010) Explaining variational approximations. *Am. Statistn*, **64**, 140–153.

Ormerod, J. and Wand, M. P. (2012) Gaussian variational approximation inference for generalized linear mixed models. *Am. Statistn*, **21**, 2–17.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis*. New York: Springer.

Reiss, P. T., Huang, L. and Mennes, M. (2010) Fast function-on-scalar regression with penalized basis expansions. *Int. J. Biostatist.*, **6**, article 28.

Ruppert, D. (2002) Selecting the number of knots for penalized splines. *J. Computnl Graph. Statist.*, **11**, 735–757.

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

Scheipl, F., Staicu, A.-M. and Greven, S. (2015) Functional additive mixed models. *J. Computnl Graph. Statist*., to be published.

Staicu, A.-M., Crainiceanu, C. and Carroll, R. (2010) Fast methods for spatially correlated multilevel functional data. *Biostatistics*, **11**, 177–194.
Stan Development Team (2013) *Stan Modeling Language User's Guide and Reference Manual, Version 1.3*. (Available from `http://mc-stan.org/`.)
Titterington, D. M. (2004) Bayesian methods for neural networks and related models. *Statist. Sci.*, **19**, 128–139.
Yang, R. and Berger, J. O. (1994) Estimation of a covariance matrix using the reference prior. *Ann. Statist.*, **22**, 1195–1211.
Yao, F., Müller, H. and Wang, J. (2005) Functional data analysis for sparse longitudinal data. *J. Am. Statist. Ass.*, **100**, 577–590.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:

'Appendices to: Assessing effects of stroke on control using hierarchical function-on-scalar regression'.