Bayesian Analysis (2017) **12**, Number 2, pp. 335–350

Bayesian Functional Data Modeling for Heterogeneous Volatility

Bin Zhu[*‡] and David B. Dunson[†§]

Abstract. Although there are many methods for functional data analysis, less emphasis is put on characterizing variability among volatilities of individual functions. In particular, certain individuals exhibit erratic swings in their trajectory while other individuals have more stable trajectories. There is evidence of such volatility heterogeneity in blood pressure trajectories during pregnancy, for example, and reason to suspect that volatility is a biologically important feature. Most functional data analysis models implicitly assume similar or identical smoothness of the individual functions, and hence can lead to misleading inferences on volatility and an inadequate representation of the functions. We propose a novel class of functional data analysis models characterized using hierarchical stochastic differential equations. We model the derivatives of a mean function and deviation functions using Gaussian processes, while also allowing covariate dependence including on the volatilities of the deviation functions. Following a Bayesian approach to inference, a Markov chain Monte Carlo algorithm is used for posterior computation. The methods are tested on simulated data and applied to blood pressure trajectories during pregnancy.

Keywords: Bayesian functional data analysis, Gaussian process, state space model, stochastic differential equation, volatility heterogeneity.

1 Introduction

Multi-subject functional data arise frequently in many fields of research, including epidemiology, clinical trials and environmental health. Sequential observations are collected over time for multiple subjects, and can be treated as being generated from a smooth trajectory contaminated with noise. There are a rich variety of methods available for characterizing variability and covariate dependence in functional data ranging from hierarchical basis expansions to functional principal components analysis. In defining models for functional data and in interpreting variability in trajectories, either unexplained or due to covariates, the emphasis has been on differences in the level and local trends. Dynamic features, such as velocity, acceleration and especially volatility, are

*Tenure-Track Principal Investigator, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD 20852, e-mail: bin.zhu@nih.gov

†Arts & Sciences Distinguished Professor, Department of Statistical Science, Duke University, Durham, NC 27708, e-mail: dunson@stat.duke.edu

‡This work was supported by the Intramural Research Program of the National Cancer Institute, National Institutes of Health, Maryland, USA.

§This work was supported by Award Number R01ES017436 and R01ES17240 from the National Institute of Environmental Health Sciences, by funding from the National Institutes of Health (5P2O-RR020782-O3) and the U.S. Environmental Protection Agency (RD-83329301-0).

 DOI: 10.1214/16-BA1004

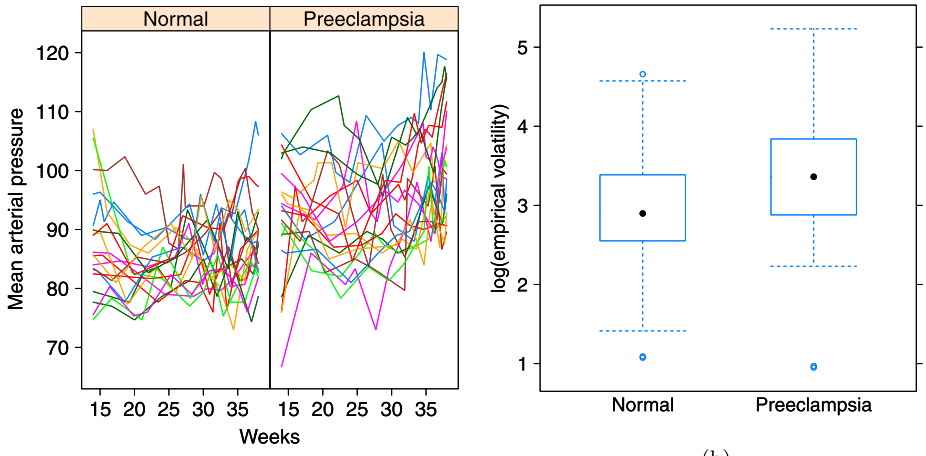
also important but have received less attention, with exceptions in the study of growth data (Ramsay and Silverman, 2002) and in finance (Heston, 1993; Jacquier et al., 2002; Shephard, 2005; Barndorff-Nielsen and Shephard, 2012; Horváth et al., 2014).

Analysis of functional data dynamics studies temporal changes in trajectories and effects of covariates on these changes. For example, Wang et al. (2008) used linear differential equations to model price velocity and acceleration in eBay auctions and explored the auction subpopulation effect. Müller and Yao (2010) modeled the velocity of online auction bids using empirical stochastic differential equations with time-varying coefficients and a smooth drift process. Zhu et al. (2011) inferred the rate functions of prostate-specific antigen profiles using a semiparametric stochastic velocity model, in which rate functions are regarded as realizations of Ornstein–Uhlenbeck processes dependent on covariates of interest.

This article investigates a different dynamic feature, the volatility, which measures the conditional variance of trajectory changes over an infinitesimal time interval. We propose a stochastic volatility regression model, with Gaussian process priors used for the group mean and subject specific deviation functions through stochastic differential equations. We further accommodate inference on covariate effects on volatility through allowing the diffusion term of stochastic differential equations for deviation functions to depend on covariates. Although volatility has been extensively studied through stochastic volatility models in finance, the setting, model specifications and data features are distinct from ours. Stochastic volatility models typically focus on a single volatility process which is time-varying and associated with a price process for high-frequency finance data. More relevant is the literature on multivariate stochastic volatility models; for recent references, refer to Loddo et al. (2011), Van Es and Spreij (2011), Müller et al. (2011), Ishihara and Omori (2012) and Durante et al. (2013).

This setting differs from ours in that the focus is on multivariate time series modeling instead of functional data analysis, with interest in the joint volatility dynamics over time for the different assets. In contrast, we are interested in studying variation across individuals in a time-constant subject-specific volatility; that is, certain subjects may have very smooth trajectories while other subjects have erratic trajectories. It is our conjecture that such volatility heterogeneity is common in biomedical settings, but is overlooked in analyzing data with models that implicitly prescribe a single level of smoothness for all subjects. As data are sparse and irregularly spaced in most studies, it is not surprising such behavior is overlooked. However, the volatility in a biomarker may be as important or more important than the overall level and trend in the biomarker. We provide motivation through the following longitudinal blood pressure data set.

The Healthy Pregnancy, Healthy Baby study (Miranda et al., 2009) collected longitudinal blood pressure measurements for pregnant women. Blood pressures are measured at irregularly spaced times during the second and third trimesters, with the number of measurements per subject varying from 9 to 19. We are interested in estimating subject-specific volatilities of blood pressure trajectories and in identifying covariates associated with the volatility. Figure 1(a) plots mean arterial pressure trajectories for twenty randomly selected normal women and women with preeclampsia, respectively. Clearly the mean arterial pressure trajectories among the preeclampsia group are more wiggly than



(a) (b)

Figure 1: (a) Mean arterial pressure trajectories for twenty randomly selected normal women and women with preeclampsia; (b) Log-transformed empirical volatilities for women in the normal group and preeclampsia group. Y_{ij} denotes blood pressure for the ith woman at time t_{ij}, and $U_{ij} = Y_{ij} - \bar{Y}_j$ is the deviation from the group mean blood pressure $\bar{Y}_j$. The empirical volatility measures the fluctuation of trajectories empirically and is defined as $\sum_{j=1}^{n_i-1} (U_{i,j+1} - U_{i,j})^2/\{n_i(t_{i,j+1} - t_{i,j})\}$ with n_i the number of observations for the ith woman.

the ones in the normal group, which is also implied by boxplots of log-transformed empirical volatilities in Figure 1(b). To explore volatility differences among various groups in addition to preeclampsia, we apply normal linear regression for log-transformed empirical volatilities with the covariates race, mother's age, obesity, preeclampsia, parity and smoking. The results suggest that preeclampsia and smoking are associated with empirical volatility, with p-values of 0.0005 and 0.002, respectively. This is a two-stage approach, which is useful as an exploratory tool but ignores measurement errors and uncertainty in volatility estimation.

Additionally, empirical volatilities in Figure 1(b) are heterogeneous even within the normal or preeclampsia group. This heterogeneity will be largely omitted when we apply functional data analysis methods with identical or similar smoothness for individual functions within a group. Consequently, the wiggly trajectories will be over-smoothed while the smooth trajectories will be under-smoothed. We can potentially estimate the individual trajectories separately but it is well known that borrowing of information will dramatically improve performance for sparse functional data. In addition, separate estimation does not allow for inferences on covariate effects and unexplained variability in volatility.

As for the clinical question addressed, the previous functional data analysis methods mainly focus on the shift of blood pressure level and ignore examining the volatility of

blood pressure, which measures haemodynamic stability and is crucial for cardiovascular health. For example, a recent study shows that blood pressure stability rather than blood pressure level is associated with increased survival among patients on hemodialysis (Raimann et al., 2012). For the Healthy Pregnancy, Healthy Baby study, we observe that preeclampsia is commonly accompanied by blood pressure over-swinging. The joint effect of high blood pressure level and large volatility may lead to adverse birth outcomes, such as low birth weight and preterm birth.

2 Stochastic Volatility Regression Model

2.1 The Model Specification

Suppose that $Y_i(t)$, $i = 1, 2, \ldots, m$, is the observation of the ith subject at time $t \in \mathcal{T}_i = \{t_{i,1}, t_{i,2}, \ldots, t_{i,n_i} < t_U\}$, with $\mathcal{T}_i$ the set of observation times before time t_U for the ith subject. We specify an observation equation for $Y_i(t)$ as

$$Y_i(t) = M_{k_i}(t) + U_i(t) + \varepsilon_i(t), \tag{1}$$

where $Y_i(t)$ is contaminated by a measurement error $\varepsilon_i(t)$ following a one-dimensional normal distribution with mean 0 and variance σ^2_ε. Assuming the ith subject belongs to the k_ith group, with $k_i \in \{1, 2, \ldots, g\}$, we include a k_ith group mean function $M_{k_i}(t) = E\{Y_i(t) \mid M_{k_i}(t)\}$ in the observation equation. The $U_i(t)$ characterizes the subject-specific deviation from the group-mean with $E\{U_i(t)\} = 0$.

The volatility of the ith subject is defined as the conditional variance of the $(q-1)$th order derivative of $U_i(t)$ over an infinitesimal time interval. We denote the volatility $\sigma^2_{U_i} = \lim_{h\to 0} h^{-1}E[\{D^{q-1}U_i(t+h) - D^{q-1}U_i(t)\}^2 \mid D^{q-1}U_i(t)]$ with differential operator $D^q = d^q/dt^q$. As volatility approaches zero, $U_i(t)$ would be a straight line. In contrast, increasing the value of volatility would lead to a more wiggly $U_i(t)$ with a larger magnitude of fluctuation around $M_{k_i}(t)$. We focus on the case when $\sigma^2_{U_i}$ is constant over time but varies across subjects and depends on covariates for smooth curves without spikes, in which observations per subject are sparse and insufficient to estimate volatility varying over t. In contrast, stochastic volatility models typically focus on a single or a few volatility processes in which volatility is time-varying but unrelated to covariates for high frequency financial data. Our motivation is drawn from the blood pressure data; in related applications it is of substantial interest to do inferences on variability among subjects in the bumpiness or erratic-ness of biomarkers collected over time.

We specify Gaussian process priors for $M_{k_i}(t)$ and $U_i(t)$ using stochastic differential equations, which incorporate the group and individual volatilities $\sigma^2_{M_{k_i}}$ and $\sigma^2_{U_i}$:

$$D^p M_{k_i}(t) = \sigma_{M_{k_i}} \dot{W}_{k_i}(t), \tag{2}$$

$$D^q U_i(t) = \sigma_{U_i} \dot{W}'_i(t), \tag{3}$$

where $p, q \in \mathbb{N} \geq 1$, $\sigma_{M_{k_i}}, \sigma_{U_i} \in \mathbb{R}^+$, and $\dot{W}_{k_i}(t), \dot{W}'_i(t)$ are independent Gaussian white noise processes with $E\{\dot{W}_{k_i}(t)\} = E\{\dot{W}'_i(t)\} = 0$ and covariance function

$E\{\dot{W}_{k_i}(t)\dot{W}_{k_i}(t')\} = E\{\dot{W}_i'(t)\dot{W}_i'(t')\} = \delta(t-t')$, a delta function. Equations (2) and (3) specify $M_{k_i}(t)$ and $U_i(t)$ as integrated Brownian motions, which have Bayesian connections to smoothing splines (Wahba, 1990; Gu, 2013). More details are given in Section 2.2. Drift terms can be included in equations (2) and (3) when domain-specific knowledge supports a particular curve pattern, such as convergent evolution for prostate-specific antigen profiles (Zhu et al., 2011). In this article, we are interested in investigating heterogeneous individual volatilities. Particularly, the volatility $\sigma^2_{U_i}$ in stochastic differential equation (3) is allowed to vary between subjects and with covariates through a simple Gaussian log linear model, $\log(\sigma^2_{U_i}) \sim N_1(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2)$, which can be extended easily to more complex specifications. In the current setting, the group volatility $\sigma_{M_{k_i}}$ would not depend on the same $\boldsymbol{x}_i$ and we has outlined the possible extension in Discussion section.

The mean and covariance functions of Gaussian process priors for $M_{k_i}(t)$ and $U_i(t)$ can be obtained by applying stochastic integration to stochastic differential equations (2) and (3) as detailed in the Supplementary Material (Zhu and Dunson, 2016), resulting in the following lemma.

Lemma 1. *$M_k(t)$, $k = 1, 2, \ldots, g$, and $U_i(t)$, $i = 1, 2, \ldots, n$, are the summations of mutually independent Gaussian processes written as $M_k(t) = M_{k0}(t) + M_{k1}(t)$ and $U_i(t) = U_{i0}(t) + U_{i1}(t)$ with corresponding mean functions $E\{M_{k0}(t)\} = E\{M_{k1}(t)\} = E\{U_{i0}(t)\} = E\{U_{i1}(t)\} = 0$ and covariance functions*

$$K_{M_{k0}}(s,t) = \sigma^2_{M_0}\mathcal{R}_{M_0}(s,t) = \sigma^2_{M_0}\sum_{l=0}^{p-1}\phi_l(s)\phi_l(t),$$

$$K_{M_{k1}}(s,t) = \sigma^2_{M_k}\mathcal{R}_{M_1}(s,t) = \sigma^2_{M_k}\int_{\mathcal{T}} G_p(s,u)G_p(t,u)du,$$

$$K_{U_{i0}}(s,t) = \sigma^2_{U_0}\mathcal{R}_{U_0}(s,t) = \sigma^2_{U_0}\sum_{l=0}^{q-1}\phi_l(s)\phi_l(t),$$

$$K_{U_{i1}}(s,t) = \sigma^2_{U_i}\mathcal{R}_{U_1}(s,t) = \sigma^2_{U_i}\int_{\mathcal{T}} G_q(s,u)G_q(t,u)du,$$

respectively, where $\phi_l(t) = t^l/l!$, $G_q(s,u) = (s-u)_+^{q-1}/(q-1)!$ and $s,t,u \in \mathcal{T} = [0, t_U]$. We denote $\boldsymbol{M}_{k_i0} = \{M_{k_i}(0), D^1M_{k_i}(0), \ldots, D^{p-1}M_{k_i}(0)\} \sim N_p(\boldsymbol{0}, \sigma^2_{M_0}\boldsymbol{I})$ and $\boldsymbol{U}_{i0} = \{U_i(0), D^1U_i(0), \ldots, D^{q-1}U_i(0)\} \sim N_q(\boldsymbol{0}, \sigma^2_{U_0}\boldsymbol{I})$ as the initial values of $M_{k_i}(t)$ and $U_i(t)$ and their derivatives up to orders $p-1$ and $q-1$ respectively.

Hence, we can represent the prior of $M_{k_i}(t)+U_i(t)$ as a hierarchical Gaussian process,

$$\begin{aligned} M_{k_i}(t) + U_i(t) \mid M_{k_i}(t) &\sim GP(M_{k_i}(t), K_{U_{i0}}(s,t) + K_{U_{i1}}(s,t)), \\ M_{k_i}(t) &\sim GP(0, K_{M_{k0}}(s,t) + K_{M_{k1}}(s,t)), \end{aligned}$$

where $GP(M(t), K(s,t))$ denotes a Gaussian process with mean function $M(t)$ and covariance function $K(s,t)$. Different from previous hierarchical Gaussian processes (Park and Choi, 2010), in which the covariance function is modeled as squared exponential, and is identical across subjects within a group, here $K_{U_{i0}}(s,t) + K_{U_{i1}}(s,t)$ is subject-specific and depends on covariates through $\sigma^2_{U_i}$.

To carry out Bayesian inference, we further specify the following prior distributions. In particular, $\boldsymbol{M}_{k_i 0} \sim N_p(\boldsymbol{0}, \sigma^2_{M_0}\boldsymbol{I})$ with $\sigma^2_{M_0} = 10^4$ in the below applications, $\boldsymbol{U}_{i0} \sim N_q(\boldsymbol{0}, \sigma^2_{U_0}\boldsymbol{I})$, $\sigma^2_\varepsilon \sim invGa(a,b)$, $\sigma^2_{M_k} \sim invGa(a,b)$ and $\sigma^2_{U_0} \sim invGa(a,b)$, where $invGa(a,b)$ denotes the inverse gamma distribution with shape parameter a and scale parameter b. We choose weakly informative priors, for example $a = b = 0.01$, to allow the data to dominate the inference of posteriors of σ^2_ε, $\sigma^2_{M_k}$ and $\sigma^2_{U_0}$, as illustrated by the MCMC steps 3a), 3b) and 3d) in Section 3. In practice, we have found the posterior distributions for these hyperparameters to be substantially more concentrated than the prior in simulations and applications, suggesting substantial Bayesian learning. Additionally, the $\boldsymbol{\beta}$ and σ^2 follow the independent Jeffreys' prior, $f(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$.

2.2 Double-Penalized Smoothing Spline

It is well known that the smoothing spline estimate is interpretable as a Bayes estimate under an integrated Wiener process prior (Wahba, 1990). By similar arguments, we can show that when the volatilities are given and $\sigma^2_{M_0}$ and $\sigma^2_{U_0}$ go to infinity, the posterior means of $M_k(t)$ and $U_i(t)$ are equivalent to the double penalized smoothing spline $\hat{M}_k(t) + \hat{U}_i(t)$, which is the minimizer of the double penalized sum-of-squares,

$$\begin{aligned}\text{DPSS} = &\sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} \{Y(t_{ij}) - M_{k_i}(t_{ij}) - U_i(t_{ij})\}^2 + \\ &\sum_{k=1}^{g} \lambda_{M_k} \int_{\mathcal{T}} \{D^p M_k(t)\}^2 \, dt + \sum_{i=1}^{m} \lambda_{U_i} \int_{\mathcal{T}} \{D^q U_i(t)\}^2 \, dt,\end{aligned} \tag{4}$$

where penalty terms $\int_{\mathcal{T}} \{D^p M_k(t)\}^2 dt$ and $\int_{\mathcal{T}} \{D^q U_i(t)\}^2 dt$ penalize the roughness of $M_k(t)$ and $U_i(t)$ respectively, where the smoothness and the fidelity to data are balanced by the smoothing parameters $\lambda_{M_k} = \sum_{i:k_i=k} \sigma^2_\varepsilon / (n_i \sigma^2_{M_k})$ and $\lambda_{U_i} = \sigma^2_\varepsilon / (n_i \sigma^2_{U_i})$, which balance the fidelity to the data and smoothness of $M_k(t)$ and $U_i(t)$ respectively. Expressions for $\hat{M}_k(t)$ and $\hat{U}_i(t)$, depending on λ_{M_k} and λ_{U_i}, can be obtained explicitly, as shown in the following theorem.

Theorem 2.1. *The smoothing splines $\hat{M}_k(t)$ and $\hat{U}_i(t)$ with $t \in \mathcal{T}$ minimize the double-penalized sum-of-squares* (4) *and have the forms*

$$\hat{M}_k(t) = \sum_{l=0}^{p-1} \mu_{kl} \phi_l(t) + \sum_{j=1}^{n} \nu_{kj} \mathcal{R}_{M_1}(t_j, t) = \boldsymbol{\mu}_k^{\mathrm{T}} \boldsymbol{\phi}_\mu(t) + \boldsymbol{\nu}_k^{\mathrm{T}} \boldsymbol{R}_{M_1}(t)$$

$$\hat{U}_i(t) = \sum_{l=0}^{q-1} \alpha_{il} \phi_l(t) + \sum_{j=1}^{n_i} \gamma_{ij} \mathcal{R}_{U_1}(t_{ij}, t) = \boldsymbol{\alpha}_i^{\mathrm{T}} \boldsymbol{\phi}_\alpha(t) + \boldsymbol{\gamma}_i^{\mathrm{T}} \boldsymbol{R}_{U_{i1}}(t)$$

where $\boldsymbol{\mu}_k = \{\mu_{k0}, \mu_{k1}, \ldots, \mu_{k(p-1)}\}^{\mathrm{T}}$, $\boldsymbol{\nu}_k = (\nu_{k1}, \nu_{k2}, \ldots, \nu_{kn})^{\mathrm{T}}$, $\boldsymbol{\alpha}_i = \{\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{i(q-1)}\}^{\mathrm{T}}$ *and* $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{in_i})^{\mathrm{T}}$ *are the coefficients for the bases*

$$\boldsymbol{\phi}_\mu(t) = \{\phi_0(t), \phi_1(t), \ldots, \phi_{p-1}(t)\}^{\mathrm{T}},$$
$$\boldsymbol{R}_{M_1}(t) = \{\mathcal{R}_{M_1}(t_1, t), \mathcal{R}_{M_1}(t_2, t), \ldots, \mathcal{R}_{M_1}(t_n, t)\}^{\mathrm{T}},$$

$$\boldsymbol{\phi}_\alpha(t) = \{\phi_0(t), \phi_1(t), \ldots, \phi_{q-1}(t)\}^{\mathrm{T}},$$
$$\boldsymbol{R}_{U_{i1}}(t) = \{\mathcal{R}_{U_1}(t_{i1}, t), \mathcal{R}_{U_1}(t_{i2}, t), \ldots, \mathcal{R}_{U_1}(t_{in_i}, t)\}^{\mathrm{T}},$$

with $t_j \in \mathcal{T}_m = \cup_{i=1}^m \mathcal{T}_i = \{t_j,\ j = 1, 2, \ldots, n\}$, *the index set of unique observation times among all* m *subjects.*

Given $\hat{M}_k(t)$ *and* $\hat{U}_i(t)$, *the double-penalized sum-of-squares* (4) *can be written as*

$$\begin{aligned}\mathrm{DPSS} = &\sum_{i=1}^{m} \frac{1}{n_i}(\boldsymbol{Y}_i - \boldsymbol{\Delta}_i\boldsymbol{\phi}_\mu\boldsymbol{\mu}_{k_i} - \boldsymbol{\Delta}_i\boldsymbol{R}_{M_1}\boldsymbol{\nu}_{k_i} - \boldsymbol{\phi}_{\alpha_i}\boldsymbol{\alpha}_i - \boldsymbol{R}_{U_{i1}}\boldsymbol{\gamma}_i)^{\mathrm{T}} \times \\ &(\boldsymbol{Y}_i - \boldsymbol{\Delta}_i\boldsymbol{\phi}_\mu\boldsymbol{\mu}_{k_i} - \boldsymbol{\Delta}_i\boldsymbol{R}_{M_1}\boldsymbol{\nu}_{k_i} - \boldsymbol{\phi}_{\alpha_i}\boldsymbol{\alpha}_i - \boldsymbol{R}_{U_{i1}}\boldsymbol{\gamma}_i) + \\ &\sum_{k=1}^{g} \lambda_{M_k}\boldsymbol{\nu}_k^{\mathrm{T}}\boldsymbol{R}_{M_1}\boldsymbol{\nu}_k + \sum_{i=1}^{m} \lambda_{U_i}\boldsymbol{\gamma}_i^{\mathrm{T}}\boldsymbol{R}_{U_{i1}}\boldsymbol{\gamma}_i,\end{aligned} \tag{5}$$

where

$$\boldsymbol{Y}_i = \{Y(t_{i1}), Y(t_{i2}), \ldots, Y(t_{in_i})\}^{\mathrm{T}}, \quad \boldsymbol{\Delta}_i = (\delta_{jj'})_{n_i \times n},$$
$$\boldsymbol{\phi}_\mu = \{\boldsymbol{\phi}_\mu(t_1), \boldsymbol{\phi}_\mu(t_2), \ldots, \boldsymbol{\phi}_\mu(t_n)\}^{\mathrm{T}}, \quad \boldsymbol{R}_{M_1} = \{\boldsymbol{R}_{M_1}(t_1), \boldsymbol{R}_{M_1}(t_2), \ldots, \boldsymbol{R}_{M_1}(t_n)\},$$
$$\boldsymbol{\phi}_{\alpha_i} = \{\boldsymbol{\phi}_\alpha(t_{i1}), \boldsymbol{\phi}_\alpha(t_{i2}), \ldots, \boldsymbol{\phi}_\alpha(t_{in_i})\}^{\mathrm{T}}, \quad \boldsymbol{R}_{U_{i1}} = \{\boldsymbol{R}_{U_{i1}}(t_{i1}), \boldsymbol{R}_{U_{i1}}(t_{i2}), \ldots, \boldsymbol{R}_{U_{i1}}(t_{in_i})\}$$

with $\delta_{jj'} = 1$ *if* i*th subject has an observation at time* $t_{ij} = t_{j'}$, $t_{ij} \in \mathcal{T}_i$, $t_{j'} \in \mathcal{T}_m$ *and* $\delta_{jj'} = 0$ *otherwise.*

The proofs of Theorem 2.1 and the following Corollary are included in Supplementary Material.

Corollary 1. *The* $\boldsymbol{\mu}_k$, $\boldsymbol{\nu}_k$, $\boldsymbol{\alpha}_i$ *and* $\boldsymbol{\gamma}_i$ *can be obtained through a backfitting algorithm or the Gauss–Seidel method, iterating the following two steps until convergence:*

(a) Given $\hat{\boldsymbol{\mu}}_k$ *and* $\hat{\boldsymbol{\nu}}_k$, *update* $\hat{\boldsymbol{\alpha}}_i = (\boldsymbol{\phi}_{\alpha_i}^{\mathrm{T}}\boldsymbol{S}_{U_i}^{-1}\boldsymbol{\phi}_{\alpha_i})^{-1}\boldsymbol{\phi}_{\alpha_i}^{\mathrm{T}}\boldsymbol{S}_{U_i}^{-1}\tilde{\boldsymbol{Y}}_i$ *and* $\hat{\boldsymbol{\gamma}}_i = \boldsymbol{S}_{U_i}^{-1}\{\boldsymbol{I} - \boldsymbol{\phi}_{\alpha_i}(\boldsymbol{\phi}_{\alpha_i}^{\mathrm{T}}\boldsymbol{S}_{U_i}^{-1}\boldsymbol{\phi}_{\alpha_i})^{-1}\boldsymbol{\phi}_{\alpha_i}^{\mathrm{T}}\boldsymbol{S}_{U_i}^{-1}\}\tilde{\boldsymbol{Y}}_i$, *where* $\boldsymbol{S}_{U_i} = \boldsymbol{R}_{U_i1} + n_i\lambda_{U_i}\boldsymbol{I}$ *and* $\tilde{\boldsymbol{Y}}_i = \boldsymbol{Y}_i - \boldsymbol{\Delta}_i\boldsymbol{\phi}_\mu\hat{\boldsymbol{\mu}}_{k_i} - \boldsymbol{\Delta}_i\boldsymbol{R}_{M_1}\hat{\boldsymbol{\nu}}_{k_i}$, $i = 1, 2, \ldots, m$;

(b) Given $\hat{\boldsymbol{\alpha}}_i$ *and* $\hat{\boldsymbol{\gamma}}_i$, *update* $\hat{\boldsymbol{\mu}}_k = (\boldsymbol{\phi}_\mu^{\mathrm{T}}\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{S}_{M_k}^{-1}\boldsymbol{\Delta}\boldsymbol{\phi}_\mu)^{-1}\boldsymbol{\phi}_\mu^{\mathrm{T}}\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{S}_{M_k}^{-1}\tilde{\boldsymbol{Y}}_k$ *and* $\hat{\boldsymbol{\nu}}_k = \boldsymbol{S}_{M_k}^{-1}\{\boldsymbol{I} - \boldsymbol{\Delta}\boldsymbol{\phi}_\mu(\boldsymbol{\phi}_\mu^{\mathrm{T}}\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{S}_{M_k}^{-1}\boldsymbol{\Delta}\boldsymbol{\phi}_\mu)^{-1}\boldsymbol{\phi}_\mu^{\mathrm{T}}\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{S}_{M_k}^{-1}\}\tilde{\boldsymbol{Y}}_k$, *where* $\boldsymbol{S}_{M_k} = \boldsymbol{\Delta}\boldsymbol{R}_{M_1} + \lambda_{M_k}\boldsymbol{I}$, $\tilde{\boldsymbol{Y}}_k = \sum_{i:k_i=k}\boldsymbol{\Delta}_i^{\mathrm{T}}(\boldsymbol{Y}_i - \boldsymbol{\phi}_{\alpha_i}\hat{\boldsymbol{\alpha}}_i - \boldsymbol{R}_{U_{i1}}\hat{\boldsymbol{\gamma}}_i)/n_i$ *and* $\boldsymbol{\Delta} = \sum_{i:k_i=k}\boldsymbol{\Delta}_i^{\mathrm{T}}\boldsymbol{\Delta}_i/n_i$, $k = 1, 2, \ldots, g$.

3 Posterior Computation

We prefer a fully Bayesian approach that allows for uncertainty in smoothing parameters through hyperpriors. In addition, it is unclear how to implement generalized cross validation (Chap. 4, Wahba, 1990) when λ_{U_i} depends on covariances, and when n is large, it is computational infeasible to invert the $n \times n$ matrix $\boldsymbol{S}_{M_k}$ involved in the backfitting algorithm. Instead, we propose an Markov chain Monte Carlo algorithm for posterior computation that solves these problems. The algorithm achieves computational efficiency by leveraging on the Markovian property of stochastic differential equations and samples $M_k(t)$ and $U_i(t)$ through the simulation smoother (Durbin and Koopman, 2002), which requires the following proposition.

Proposition 1. *Let $X(t)$ denote a $(r-1)$th-order integral Wiener process, defined by the stochastic differential equation $D^r X(t) = \dot{W}(t)$. Consequently, the $\boldsymbol{X}_j = \{X(t_j), D^1X(t_j), \ldots, D^{r-1}X(t_j)\}^{\mathrm{T}}$, $j = 1, 2, \ldots, n$, follows a state equation*

$$\boldsymbol{X}_{j+1} = \boldsymbol{G}_j \boldsymbol{X}_j + \boldsymbol{\omega}_j,$$

where $\boldsymbol{G}_j = \sum_{k=0}^{r} \delta_j^k \boldsymbol{C}^k / k!$ and $\boldsymbol{\omega}_j \sim N_r(\boldsymbol{0}, \boldsymbol{W}_j)$ with $\boldsymbol{C} = (c_{ll^{\mathrm{T}}})_{r\times r}$, $c_{ll'} = 1$ when $l' = l + 1$ and $c_{ll'} = 0$,otherwise, $\boldsymbol{W}_j = \int_0^{\delta_j} \exp\{\boldsymbol{C}(\delta_j - u)\}\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}} \exp\{\boldsymbol{C}^{\mathrm{T}}(\delta_j - u)\}du$, $\boldsymbol{D} = (0, 0, \ldots, 1)^{\mathrm{T}}$ and $\delta_j = t_{j+1} - t_j$.

The proof is in Supplementary Material. Finally, we outline the proposed Markov chain Monte Carlo algorithm as follows.

(1) Given $M_{k_i}(t_{ij})$, σ_ε^2 and $\sigma_{U_i}^2$, sample $U_i(t_{ij})$, $i = 1, 2, \ldots, m$, $j = 0, 1, \ldots, n_i$. Let $Y_{U_{ij}} = Y_i(t_{ij}) - M_{k_i}(t_{ij})$ and the stochastic volatility regression model for the ith subject can be expressed as the following state space model (Jones, 1993; Durbin and Koopman, 2001), from which we can draw samples of $U_i(t_{ij})$ and its derivatives using the simulation smoother.

$$\begin{aligned} Y_{U_{ij}} &= \boldsymbol{F}_{U_{ij}} \boldsymbol{U}_{ij} + \varepsilon_{U_{ij}}, \\ \boldsymbol{U}_{i(j+1)} &= \boldsymbol{G}_{U_{ij}} \boldsymbol{U}_{ij} + \sigma_{U_i} \boldsymbol{\omega}_{U_{ij}}, \end{aligned}$$

where $\boldsymbol{F}_{U_{ij}} = (1, 0, \ldots, 0)$, $\boldsymbol{U}_{ij} = \{U_i(t_{ij}), D^1U_i(t_{ij}), \ldots, D^{q-1}U_i(t_{ij})\}^{\mathrm{T}}$ and $\varepsilon_{U_{ij}} \overset{\text{i.i.d.}}{\sim} N_1(0, \sigma_\varepsilon^2)$, which denotes $\varepsilon_{U_{ij}}$ independently following an identical distribution $N_1(0, \sigma_\varepsilon^2)$. Similar to the $\boldsymbol{G}_j$, $\boldsymbol{\omega}_j$ and $\boldsymbol{W}_j$ in Proposition 1, the $\boldsymbol{G}_{U_{ij}}$, $\boldsymbol{\omega}_{U_{ij}}$ and $\boldsymbol{W}_{U_{ij}}$ follow the same specifications with $r = q$.

(2) Given $U_i(t_j)$, σ_ε^2 and $\sigma_{M_k}^2$, sample $M_k(t_j)$, $k = 1, 2, \ldots, g$, $j = 0, 1, \ldots, n$. Similarly, we rewrite the stochastic volatility regression model for the kth group as the following state space model and then sample $M_{k_i}(t_{ij})$ and its derivatives by the simulation smoother.

$$\begin{aligned} \boldsymbol{Y}_{M_{kj}} &= \boldsymbol{F}_{M_{kj}} \boldsymbol{M}_{kj} + \boldsymbol{\varepsilon}_{M_{kj}}, \\ \boldsymbol{M}_{k(j+1)} &= \boldsymbol{G}_{M_{kj}} \boldsymbol{M}_{kj} + \sigma_{M_k} \boldsymbol{\omega}_{M_{kj}}, \end{aligned}$$

where $\boldsymbol{Y}_{M_{kj}} = (Y^i_{M_{kj}})_{m\times 1}$, $\boldsymbol{M}_{kj} = \{M_k(t_j), D^1M_k(t_j), \ldots, D^{p-1}M_k(t_j)\}^{\mathrm{T}}$, $\boldsymbol{F}_{M_{kj}} = (F^{il}_{M_{kj}})_{m\times p}$ and $\boldsymbol{\varepsilon}_{M_{kj}} = \mathrm{diag}(\varepsilon^1_{M_{kj}}, \varepsilon^2_{M_{kj}}, \ldots, \varepsilon^m_{M_{kj}})$. When ith subject has an observation at time t_j and $k_i = k$, $Y^i_{M_{kj}} = Y_i(t_j) - U_i(t_j)$, $F^{i1}_{M_{kj}} = 1$ and $\varepsilon^i_{M_{kj}} \sim N_1(0, \sigma_\varepsilon^2)$. Otherwise, $Y^i_{M_{kj}} = F^{il}_{M_{kj}} = \varepsilon^i_{M_{kj}} = 0$. The $\boldsymbol{G}_{M_{kj}}$, $\boldsymbol{\omega}_{M_{kj}}$ and $\boldsymbol{W}_{M_{kj}}$ are given by Proposition 1 with $r = p$.

(3a) Given $M_{k_i}(t_{ij})$ and $U_i(t_{ij})$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n_i$, sample σ_ε^2 from $invGa(a + \sum_{i=1}^m n_i/2, b + \sum_{i=1}^m \sum_{j=1}^{n_i} \{Y_i(t_{ij}) - M_{k_i}(t_{ij}) - U_i(t_{ij})\}^2/2)$, the posterior distribution of σ_ε^2.

(3b) Given $\boldsymbol{U}_{i0}$, sample $\sigma_{U_0}^2$ from $invGa(a + mq/2, b + \sum_{i=0}^m \boldsymbol{U}_{i0}^{\mathrm{T}}\boldsymbol{U}_{i0}/2)$, the posterior distribution of $\sigma_{U_0}^2$.

(3c) Given $\boldsymbol{M}_{kj}$, sample $\sigma^2_{M_k}$ from the posterior distribution $invGa(a + np/2, b + \sum_{j=0}^{n-1}(\boldsymbol{M}_{k(j+1)} - \boldsymbol{G}_{M_{kj}}\boldsymbol{M}_{kj})^{\mathrm{T}}\boldsymbol{W}^{-1}_{M_{kj}}(\boldsymbol{M}_{k(j+1)} - \boldsymbol{G}_{M_{kj}}\boldsymbol{M}_{kj})/2)$.

(3d) Given $\boldsymbol{U}_{ij}$, $\boldsymbol{\beta}$ and σ^2, sample $\sigma^2_{U_i}$ using a Metropolis–Hasting algorithm. We choose $\sigma^2_{U_i} \sim invGa(a, b)$ as the proposal prior distribution and a proposal $\sigma^{2*}_{U_i}$ can be easily drawn from $invGa(a + n_i q/2, b + \sum_{j=0}^{n_i-1}(\boldsymbol{U}_{i(j+1)} - \boldsymbol{G}_{U_{ij}}\boldsymbol{U}_{ij})^{\mathrm{T}}\boldsymbol{W}^{-1}_{U_{ij}}(\boldsymbol{U}_{i(j+1)} - \boldsymbol{G}_{U_{ij}}\boldsymbol{U}_{ij})/2)$ the corresponding proposal posterior distribution. The $\sigma^{2*}_{U_i}$ will be accepted with the following probability and discarded otherwise with $\sigma^2_{U_i}$ unchanged,

$$\min\left\{\frac{f_{LN}(\sigma^{2*}_{U_i} \mid \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2)\prod_{j=0}^{n_i-1} f_{N_q}(\boldsymbol{U}_{i(j+1)} - \boldsymbol{G}_{U_{ij}}\boldsymbol{U}_{ij} \mid \boldsymbol{0}, \sigma^{2*}_{U_i}\boldsymbol{W}_{U_{ij}})f_{invGa}(\sigma^2_{U_i} \mid a_{U_i}, b_{U_i})}{f_{LN}(\sigma^2_{U_i} \mid \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2)\prod_{j=0}^{n_i-1} f_{N_q}(\boldsymbol{U}_{i(j+1)} - \boldsymbol{G}_{U_{ij}}\boldsymbol{U}_{ij} \mid \boldsymbol{0}, \sigma^2_{U_i}\boldsymbol{W}_{U_{ij}})f_{invGa}(\sigma^{2*}_{U_i} \mid a_{U_i}, b_{U_i})}, 1\right\},$$

where f_{LN}, f_{N_q} and f_{invGa} denote the log-normal, q-dimensional normal and inverse gamma probability density functions respectively with $a_{U_i} = a + n_i q/2$, $b_{U_i} = b + \sum_{j=0}^{n_i-1}(\boldsymbol{U}_{i(j+1)} - \boldsymbol{G}_{U_{ij}}\boldsymbol{U}_{ij})^{\mathrm{T}}\boldsymbol{W}^{-1}_{U_{ij}}(\boldsymbol{U}_{i(j+1)} - \boldsymbol{G}_{U_{ij}}\boldsymbol{U}_{ij})/2$.

(4) Given $\sigma^2_{U_i}$, sample $\boldsymbol{\beta}$ and σ^2. Let $\boldsymbol{Z} = (\log\sigma^2_{U_1}, \log\sigma^2_{U_2}, \ldots, \log\sigma^2_{U_m})^{\mathrm{T}}$, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}$ and $\hat{\sigma}^2 = (\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/(m-k)$. We draw τ from Chi-squared distribution with $m-k$ degrees of freedom and set $\sigma^2 = (m-k)\hat{\sigma}^2/\tau$ and then sample $\boldsymbol{\beta}$ from $N_m(\hat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1})$.

The algorithm coded in R can be downloaded at http://dceg.cancer.gov/tools/analysis/SVR/SVR_beta.zip.

4 Simulation

We carry out two simulation studies to evaluate the performance of the proposed method and compare it to alternative methods including natural cubic splines (Wahba, 1990), functional principal components analysis (Yao et al., 2005) and functional mixed effects models (Guo, 2002). The comparison focuses on performance in estimating the trajectory $M_{k_i}(t) + U_i(t)$, the volatility $\sigma^2_{U_i}$ and the coefficients $\boldsymbol{\beta}$. We considered both the cases with heterogeneous and homogeneous volatilities respectively, the later of which is in favor of the functional principal components analysis method and functional mixed effects models.

The first simulation study is designed to investigate the consequence of ignoring heterogeneity of volatilities. One hundred replicated datasets, each consisting of 100 trajectories, are sampled from the stochastic volatility regression model, in which the log-transformed volatilities are normally distributed. More precisely, we choose $\boldsymbol{\beta} = (0, 0.6, 2)^{\mathrm{T}}$ and $\boldsymbol{x}_i = (1, x_{i1}, x_{i2})^{\mathrm{T}}$ with x_{i1} and x_{i2} sampled from $x_{i1} \overset{\text{i.i.d.}}{\sim} Bin(1, 0.4)$ and $x_{i2} \overset{\text{i.i.d.}}{\sim} N_1(0, 0.25)$ respectively. Given $\boldsymbol{\beta}$ and $\boldsymbol{x}_i$, volatilities $\sigma^2_{U_i}$'s are drawn from $\log(\sigma^2_{U_i}) \sim N_1(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, 1)$. Along with $\sigma^2_{M_1} = \sigma^2_{M_2} = 10$, $\sigma^2_\varepsilon = 1$, $p = 2$ and $q = 1$, $M_1(t)$, $M_2(t)$, $U_i(t)$ and $\varepsilon_i(t)$ are sampled at $t \in \{0.2, 0.4, \ldots, 4\}$ from equations (2) and (3) and the distribution of measurement error $\varepsilon_i(t)$. Twenty percent of samples are removed

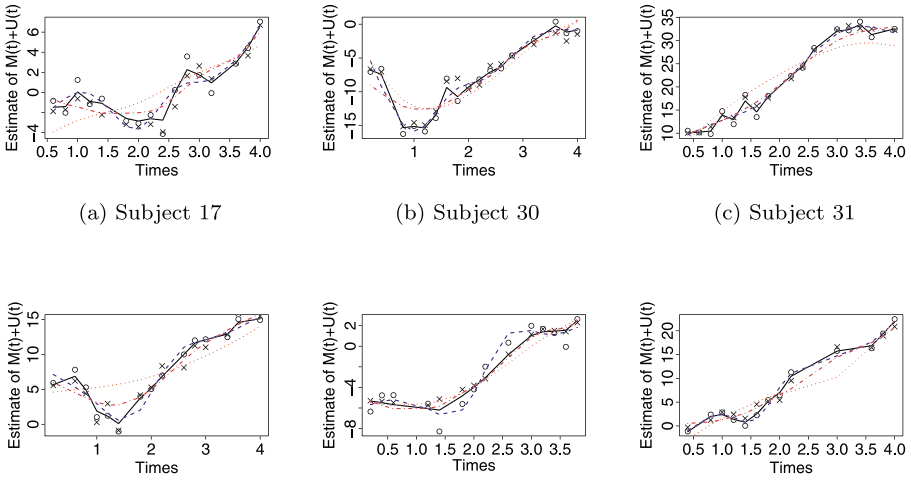
completely at random, resulting in an average of 16 unequally spaced observations per subject. The ith subject is randomly assigned to one of the two groups with equal probability and $Y_i(t)$ is obtained from observation equation (1).

We first apply stochastic volatility regression using the proposed Markov chain Monte Carlo algorithm, with 15,000 iterations and keeping every 5th of the last 10,000 samples for posterior analysis. It takes about 80 minutes on a personal computer with 2.33 Gigahertz Intel(R) Xeon(R) central processing unit. Posterior means are chosen as the estimates of $M_{k_i}(t)+U_i(t)$, $\sigma^2_{U_i}$ and $\boldsymbol{\beta}$. The trajectories $M_{k_i}(t)+U_i(t)$'s are estimated by natural cubic splines for one subject at a time, by functional principal components analysis and by functional mixed effects models for subjects within each group, taking about 1 minute, 2 minutes and 50 minutes on the same personal computer, respectively. For natural cubic spline, functional principal components analysis and functional mixed effects models, we may also estimate covariate effects on volatility through a two-stage method: estimating empirical volatility by $\sum_{j=1}^{n_i-1}(\hat{U}_{i,j+1}-\hat{U}_{i,j})^2/\{n_i(t_{i,j+1}-t_{i,j})\}$ in the first stage with $\hat{U}_{i,j}$ the estimate of $U_i(t)$ at time t_{ij}, and in the second stage, empirical volatilities are regressed on covariates to obtain the estimate of $\boldsymbol{\beta}$.

For each simulated dataset, we calculate average squared error for the trajectory $\text{ASE}(M+U)=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\{\hat{M}_{k_i}(t_{ij})+\hat{U}_i(t_{ij})-M_{k_i}(t_{ij})-U_i(t_{ij})\}^2/(mn_i)$, average squared error for log volatility $\text{ASE}\{log(\sigma^2_U)\}=\sum_{i=1}^{m}\{log(\hat{\sigma}^2_{U_i})-log(\sigma^2_{U_i})\}^2/m$, and squared errors for coefficient estimates $\text{SE}(\beta_l)=(\hat{\beta}_l-\beta_l)^2$, $l=0,1,2$. Table 1 reports means of $\text{ASE}(M+U)$, $\text{ASE}\{log(\sigma^2_U)\}$ and $\text{SE}(\beta_l)$ across 100 replicate datasets. Means of average squared errors and means of squared errors by natural cubic spline and functional principal components analysis approaches are significantly inflated, for example, being doubled and tripled in $\text{MASE}(M+U)$ respectively, while $\text{MASE}(M+U)$ of functional mixed effects models is slightly larger than for stochastic volatility regression. We randomly select a data set for close examination. We calculate the individual average squared error of the trajectory $\sum_{j=1}^{n_i}\{\hat{M}_{k_i}(t_{ij})+\hat{U}_i(t_{ij})-M_{k_i}(t_{ij})-U_i(t_{ij})\}^2/n_i$, and select the subjects with the largest individual average squared errors for natural cubic splines and functional principal components.

Figure 2 shows estimates of the trajectory for six subjects. The figure illustrates that, by treating one trajectory at a time, natural cubic splines lead to over fitting, e.g. Figure 2(d) and 2(e), with both over and under estimated volatilities. Functional principal components analysis instead faces problems in not adapting to the different volatility levels, for example, subjects with high volatility are over smoothed, e.g. Figure 2(b) and 2(d). Functional mixed effects models does not borrow smoothness information across subjects. Hence, it fits some curves well (e.g. Figure 2(c)) but over smooths other curves (e.g. Figure 2(b)). Although this simulation is based on the proposed model, it nonetheless illustrates the importance of adaptation of varying smoothness while borrowing smoothness information across subjects.

Our second simulation study assumes constant volatilities across subjects, with the set-up otherwise identical to the first study. The observations are generated from $Y_i(t)=10\{t+sin(t)\}+0.6\alpha_{1i}cos(\pi t/10)+0.2\alpha_{2i}sin(\pi t/10)+\varepsilon_i(t)$ for subjects in the first group and from $Y_i(t)=10\{t+cos(t)\}+0.5\alpha_{1i}cos(\pi t/10)+0.3\alpha_{2i}sin(\pi t/10)+\varepsilon_i(t)$ for the



(a) Subject 17 (b) Subject 30 (c) Subject 31

(d) Subject 39 (e) Subject 44 (f) Subject 86

Figure 2: The plots of observation ($\circ$) and trajectory at time t_{ij} ($\times$), as well as estimates of trajectory $M_{k_i}(t)+U_i(t)$ by stochastic volatility regression (—), natural cubic splines ($-\ -\ -$), functional principal components analysis ($\cdots$) and functional mixed effects models ($-\cdot-\cdot-$), for six subjects in one simulated dataset with the largest individual average squared errors $\sum_{j=1}^{n_i}\{\hat{M}_{k_i}(t_{ij})+\hat{U}_i(t_{ij})-M_{k_i}(t_{ij})-U_i(t_{ij})\}^2/n_i$.

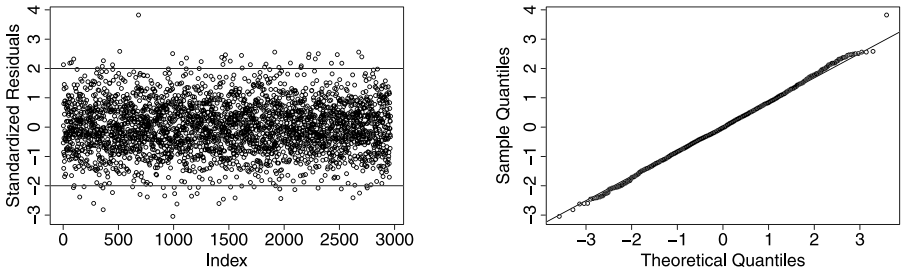
			Case I			Case II
Method	$M+U$	$\log(\sigma_U^2)$	β_0	β_1	β_2	$M+U$
Stochastic volatility regression	0.345	0.614	0.043	0.081	0.075	1.122
Natural cubic spline	0.609	1.297	0.089	0.165	1.724	1.477
Functional PCA	1.099	2.966	1.144	0.185	1.969	1.185
Functional mixed effects models	0.576	2.220	0.647	0.180	1.839	1.112

Table 1: The mean of squared errors or average square errors of the estimates of trajectory, volatility and covariate effect across 100 replicate datasets for stochastic volatility regression, natural cubic spline, functional principal components analysis (PCA) and functional mixed effects models.

ones in the second group, with $\alpha_{1i} \overset{\text{i.i.d.}}{\sim} N_1(0,4)$, $\alpha_{2i} \overset{\text{i.i.d.}}{\sim} N_1(0,1)$ and $\varepsilon_i(t) \overset{\text{i.i.d.}}{\sim} N_1(0,1)$. As illustrated in Table 1, stochastic volatility regression, similar to functional principal components analysis and functional mixed effects models, has similar performance with lower errors than natural cubic splines. This suggests that stochastic volatility regression can also adapt to the homogeneous case.

5 Applications

It is a standard practice to monitor blood pressure of pregnant woman. However, fluctuations in pregnancy and the associated factors are largely unstudied. We apply the



(a) (b)

Figure 3: The scatter plot and quantile-quantile plot of standardized residuals.

Parameter	Mean	Mode	Standard deviation	95% highest posterior density interval
σ^2_{ε}	17.807	17.818	0.694	[16.389, 19.106]
$\sigma^2_{M_1}$	0.236	0.187	0.181	[0.040, 0.556]
$\sigma^2_{M_2}$	0.204	0.162	0.148	[0.042, 0.472]
$\sigma^2_{U_0}$	46.729	46.412	4.776	[38.263, 56.619]
σ^2	0.734	0.741	0.333	[0.082, 1.295]

Table 2: Blood pressure data: Posterior summary of parameters in the stochastic volatility regression model.

proposed stochastic volatility regression approach to analyze longitudinal blood pressure measurements in the Healthy Pregnancy, Healthy Baby study, aiming to investigate the stability of blood pressure trajectories and identify the associated factors. The data consist of 106 non-Hispanic white and 176 non-Hispanic black women whose first blood pressure measurement is collected before the 16th week of gestation and the last one no earlier than the 37th week of gestation. Most subjects have 9 (35.10% of them), 10 (29.28%) or 11 (14.98%) measurements spaced at irregular times. The covariates we focused on include race as non-Hispanic white versus non-Hispanic black and indicators of advanced maternal age, obesity, preeclampsia, previous pregnancy, and smoking. The analysis was implemented as in the simulation studies. Trace plots and autocorrelation plots suggest rapid convergence and mixing. The posterior means of standardized residuals $\varepsilon_i(t)$'s are plotted in Figure 3a. Most of points locate within two standard deviations from the mean zero. In addition, a QQ-plot in Figure 3b of the empirical quantiles of the posterior means of the standardized residuals shows close agreement with a diagonal line. These diagnostics suggest that the proposed model fits the data well. Posterior summaries of selected parameters are presented in Table 2.

The panels (a) and (b) of Figure 4 show posterior means and 95% credible intervals of the average blood pressure for non-Hispanic white and non-Hispanic black groups, respectively, which share a common pattern: decreasing till the late stage of the second trimester during 20 to 25 weeks and then increasing toward the pre-pregnancy level. Within ethnic group, there is significant heterogeneity among women in the stability of the blood pressure trajectory. As Figure 4(c) indicates, posterior means of volatility vary

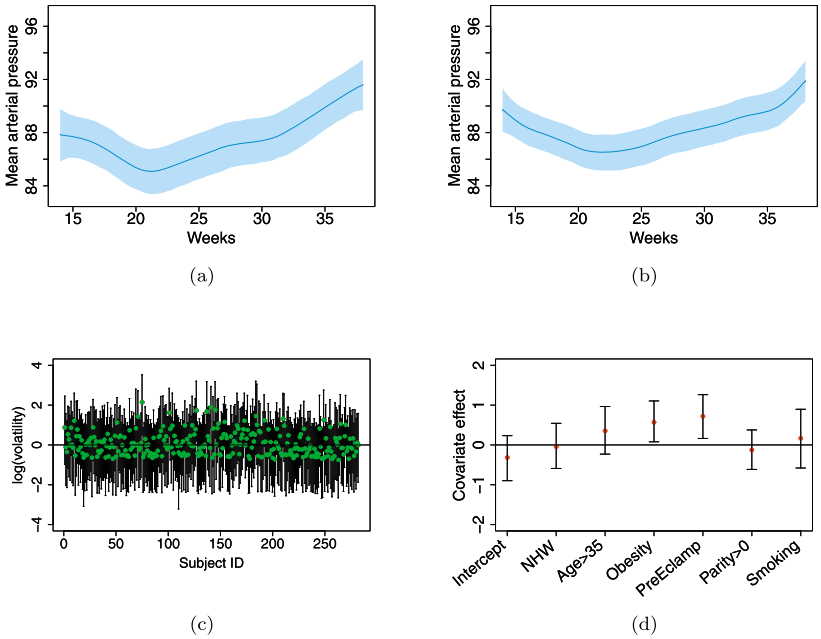


Figure 4: The posterior means and 95% highest posterior density credible intervals for (a) the blood pressure during the 2nd and 3rd trimesters for non-Hispanic white group; (b) the blood pressure during the 2nd and 3rd trimesters for non-Hispanic black group; (c) the volatility in the logarithmic scale; (d) covariate effects.

from -0.5 to 2 in the logarithmic scale, suggesting some women have stable trajectories parallel to the group mean, while other women have erratic trajectories.

Most interesting, we find that obesity and preeclampsia are associated with blood pressure volatility, with their 95% credible intervals not covering zero in Figure 4(d). This implies that pregnant women with obesity and/or preeclampsia are more likely to demonstrate irregular patterns of blood pressure relative to their ethnic group. We further examine the characteristics of women with extreme volatilities. Among the eight women presenting with the largest volatilities, most of them are non-Hispanic black with obesity and preeclampsia, do not smoke and give birth to a baby for the first time; half of them are younger than 35. For the eight subjects with the smallest volatilities, they are surprising homogeneous, with all but one being non-Hispanic white without obesity and preeclampsia, younger than 35, not smoking and giving birth to a baby before.

6 Discussion

We have proposed a Bayesian model to investigate functional data volatility and its association with covariates. As an important dynamic feature, volatility measures the

stability of the biological process. The analysis of volatility not only reveals its heterogeneity among subjects but also its dependence on the covariates of interest. Complementing current FDA methods, which mainly focus on trends in each trajectory and (perhaps) derivatives, the proposed method initiates the exploration of stability of functional data. As illustrated with the blood pressure data, our view is that substantial new insights can be obtained in a rich variety of biomedical applications by studying volatility.

The proposed model utilizes Markovian property and adopts a state space model approach to achieve computational efficiency, which requires $\mathcal{O}(m^2n)$ for calculating the matrix inverse with m subjects and n observations per subject. In contrast, the linear mixed model requires $\mathcal{O}(m^3n^3)$ for the same calculation. The current algorithm however would face the challenge of handling large number of subjects, the solution of which warrants further investigation.

The proposed stochastic volatility regression model is closely related to the functional mixed effects models (Guo, 2002) but with different specifications. Both models incorporate population-average and subject-specific curves, whose smoothness properties are the same in functional mixed effects models but could be different in stochastic volatility regression model with unequal p and q. Moreover, although both models allow smoothness parameters of subject-specific curves to vary, they are dependent on covariates in stochastic volatility regression model but not in the functional mixed effects models.

The proposed model can be extended in multiple different directions. For example, we may substitute single group mean function in equation (1) by a weighted sum of several mean functions with covariates. Limited by the sparse observations in the HPHB study, we assume the volatility time-constant, so that each subject has their own distinct volatility controlling the "erraticness" of their function. Given denser measurements, we may allow volatility to vary across time and subjects. It is also of interest to avoid normality assumptions in modeling the population distribution of volatility and in developing methods that scale to high-dimensional covariates.

Supplementary Material

Supplementary material of "Bayesian functional data modeling for heterogeneous volatility" (DOI: 10.1214/16-BA1004SUPP; .pdf).

References

Barndorff-Nielsen, O. and Shephard, N. (2012). *Financial volatility in continuous time*. Cambridge: Cambridge University Press. 336

Durante, D., Scarpa, B., and Dunson, D. (2014). "Locally adaptive factor processes for multivariate time series." *Journal of Machine Learning Research*, 15(1): 1493–1522. 336

Durbin, J. and Koopman, S. (2002). "A simple and efficient simulation smoother for state space time series analysis." *Biometrika*, 89(3): 603–616. MR1929166. doi: http://dx.doi.org/10.1093/biomet/89.3.603. 341

Durbin, J. and Koopman, S.J. (2001). *Time series analysis by state space methods*. Oxford: Oxford University Press. MR1856951. 342

Gu, C. (2013). *Smoothing spline ANOVA models*, volume 297. Springer. MR3025869. doi: http://dx.doi.org/10.1007/978-1-4614-5369-7. 339

Guo, W. (2002). "Functional mixed effects models." *Biometrics*, 58(1): 121–128. MR1891050. doi: http://dx.doi.org/10.1111/j.0006-341X.2002.00121.x. 343, 348

Heston, S. (1993). "A closed-form solution for options with stochastic volatility with applications to bond and currency options." *Review of Financial Studies*, 6(2): 327–343. doi: http://dx.doi.org/10.1093/rfs/6.2.327. 336

Horváth, L., Kokoszka, P., and Rice, G. (2014). "Testing stationarity of functional time series." *Journal of Econometrics*, 179(1): 66–82. MR3153649. doi: http://dx.doi.org/10.1016/j.jeconom.2013.11.002. 336

Ishihara, T. and Omori, Y. (2012). "Efficient Bayesian estimation of a multivariate stochastic volatility model with cross leverage and heavy-tailed errors." *Computational Statistics & Data Analysis*, 56(11): 3674–3689. MR2943920. doi: http://dx.doi.org/10.1016/j.csda.2010.07.015. 336

Jacquier, E., Polson, N., and Rossi, P. (2002). "Bayesian analysis of stochastic volatility models." *Journal of Business and Economic Statistics*, 20(1): 69–87. MR1940631. doi: http://dx.doi.org/10.1198/073500102753410408. 336

Jones, R.H. (1993). *Longitudinal data with serial correlation: a state-space approach*. New York: Chapman & Hall/CRC. MR1293123. doi: http://dx.doi.org/10.1007/978-1-4899-4489-4. 342

Loddo, A., Ni, S., and Sun, D. (2011). "Selection of multivariate stochastic volatility models via Bayesian stochastic search." *Journal of Business & Economic Statistics*, 29(3): 342–355. MR2848508. doi: http://dx.doi.org/10.1198/jbes.2010.08197. 336

Miranda, M.L., Maxson, P., and Edwards, S. (2009). "Environmental contributions to disparities in pregnancy outcomes." *Epidemiologic Reviews*, 31(1): 67. 336

Müller, H., Sen, R., and Stadtmüller, U. (2011). "Functional data analysis for volatility." *Journal of Econometrics*, 165(2): 233–245. doi: http://dx.doi.org/10.1016/j.jeconom.2011.08.002. 336

Müller, H.G. and Yao, F. (2010). "Empirical dynamics for longitudinal data." *The Annals of Statistics*, 38(6): 3458–3486. MR2766859. doi: http://dx.doi.org/10.1214/09-AOS786. 336

Park, S. and Choi, S. (2010). "Hierarchical Gaussian process regression." In *Asian Conference on Machine Learning*, 95–110. 339

Raimann, J., Usvyat, L., Thijssen, S., Kotanko, P., Rogus, J., Lacson, E., and Levin, N. (2012). "Blood pressure stability in hemodialysis patients confers a survival advantage: results from a large retrospective cohort study." *Kidney International*, 81(6): 548–558. doi: http://dx.doi.org/10.1038/ki.2011.426. 338

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*, volume 77. Springer. MR1910407. doi: http://dx.doi.org/10.1007/b98886. 336

Shephard, N. (2005). *Stochastic volatility: selected readings*. Oxford: Oxford University Press. MR2203295. 336

Van Es, B. and Spreij, P. (2011). "Estimation of a multivariate stochastic volatility density by kernel deconvolution." *Journal of Multivariate Analysis*, 102(3): 683–697. MR2755024. doi: http://dx.doi.org/10.1016/j.jmva.2010.12.003. 336

Wahba, G. (1990). *Spline models for observational data*, volume 59. Philadelphia: Society for Industrial Mathematics. MR1045442. doi: http://dx.doi.org/10.1137/1.9781611970128. 339, 340, 341, 343

Wang, S., Jank, W., Shmueli, G., and Smith, P. (2008). "Modeling price dynamics in eBay auctions using differential equations." *Journal of the American Statistical Association*, 103(483): 1100–1118. MR2528829. doi: http://dx.doi.org/10.1198/016214508000000670. 336

Yao, F., Müller, H., and Wang, J. (2005). "Functional linear regression analysis for longitudinal data." *The Annals of Statistics*, 33(6): 2873–2903. MR2253106. doi: http://dx.doi.org/10.1214/009053605000000660. 343

Zhu, B. and Dunson B. (2016). Supplementary material of "Bayesian functional data modeling for heterogeneous volatility." *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/16-BA1004SUPP. 339

Zhu, B., Taylor, J., and Song, P. (2011). "Semiparametric stochastic modeling of the rate function in longitudinal studies." *Journal of the American Statistical Association*, 106(496): 1485–1495. MR2896851. doi: http://dx.doi.org/10.1198/jasa.2011.tm09294. 336, 339