Electronic Journal of Statistics
Vol. 11 (2017) 4258–4296
ISSN: 1935-7524
DOI: 10.1214/17-EJS1324

A variational Bayes approach to a semiparametric regression using Gaussian process priors

Victor M. H. Ong, David K. Mensah, David J. Nott

Department of Statistics and Applied Probability
National University of Singapore, Singapore

Seongil Jo

Department of Statistics (Institute of Applied Statistics)
Chonbuk National University, Republic of Korea

Beomjo Park and Taeryon Choi*

Department of Statistics, Korea University, Republic of Korea
e-mail: trchoi@korea.ac.kr

Abstract: This paper presents a variational Bayes approach to a semiparametric regression model that consists of parametric and nonparametric components. The assumed univariate nonparametric component is represented with a cosine series based on a spectral analysis of Gaussian process priors. Here, we develop fast variational methods for fitting the semiparametric regression model that reduce the computation time by an order of magnitude over Markov chain Monte Carlo methods. Further, we explore the possible use of the variational lower bound and variational information criteria for model choice of a parametric regression model against a semiparametric alternative. In addition, variational methods are developed for estimating univariate shape-restricted regression functions that are monotonic, monotonic convex or monotonic concave. Since these variational methods are approximate, we explore some of the trade-offs involved in using them in terms of speed, accuracy and automation of the implementation in comparison with Markov chain Monte Carlo methods and discuss their potential and limitations.

MSC 2010 subject classifications: Primary 62G08; secondary 62F15.
Keywords and phrases: Cosine series, Gaussian process, model selection, shape restricted regression, variational Bayes.

Received August 2016.

1. Introduction

This paper develops a mean field variational Bayes approximation algorithm for a semiparametric regression model, known as a partial linear model, that consists of parametric and nonparametric components. The nonparametric component is represented with a cosine series based on a spectral analysis of Gaussian

*Corresponding author.

process priors. Specifically, the semiparametric regression model is given by

$$Y_i = \boldsymbol{w}_i^\top \boldsymbol{\beta} + f(x_i) + \epsilon_i,\ i = 1, \ldots, n, \tag{1.1}$$

where $\boldsymbol{w}_i^\top \boldsymbol{\beta}$ is referred to as the *parametric component*, $\boldsymbol{w}_i$ and $\boldsymbol{\beta}$ are $p+1$ dimensional vectors of known covariates and coefficients respectively, and $f(\cdot)$ is an unknown function of x that is univariate and defined on the interval $[0, 1]$, the *nonparametric component*. The error terms $\{\epsilon_i\}$ are a random sample from a normal distribution with mean 0 and an unknown variance σ^2. For modeling the nonparametric component $f(x)$, a Gaussian process is used for the unknown f, $f(x) = Z(x)$, where Z is a second-order Gaussian process with mean function equal to zero and covariance function $\nu(s, t) = \mathrm{Cov}(Z(s), Z(t))$, $s, t \in [0, 1]$.

Gaussian processes provide a natural way to specify prior distributions on the space of functions for nonparametric regression (O'Hagan, 1978), and are also widely used for machine learning applications (e.g., Rasmussen and Williams (2006)). One of the main practical drawbacks in the application of Gaussian process regression (hereafter GPR) is the computational burden in fitting these models when the number of data points increases, due to the need for large dense matrix calculations and associated storage requirements. An alternative approach that avoids these problems is to linearize the covariance function and to use a computationally efficient basis representation via the spectral representation of covariance functions. For example, Paciorek (2007), Lázaro-Gredilla et al. (2010) and Tan et al. (2016) considered using the spectral representation of a stationary covariance function based on Bochner's theorem (see, e.g., Grenander (1981) and Cressie and Wikle (2011)).

On the other hand, Lenk (1999) and Lenk and Choi (2017) exploited the spectral representation via the Karhunen-Loéve expansion and Mercer's theorem (see, e.g., Grenander (1981) and Adler and Taylor (2007)),

$$Z(x) = \sum_{j=0}^{\infty} \theta_j \varphi_j(x) \tag{1.2}$$

where $\varphi_j(x)$, $j \geq 0$, form an orthonormal basis on $[0, 1]$. In particular, the cosine functions, $\varphi_0(x) = 1$ and $\varphi_j(x) = \sqrt{2}\cos(\pi j x)$, $j \geq 1$ are used as an orthonormal basis with unknown spectral coefficients to be estimated, $\theta_j = \int_0^1 Z(x)\varphi_j(x)dx$. In addition to Gaussian process priors, Bayesian inference has been considered for the semiparametric regression model using spline smoothing (e.g., Zhao and Lian (2014), Hu, Zhao and Lian (2015) and Waldmann and Kneib (2015)) and wavelets (e.g., Ko, Qu and Vannucci (2009) and Wand and Ormerod (2011)), for instance.

In Lenk and Choi (2017) the Bayesian semiparametric regression framework using Gaussian process priors in (1.2) was used, referred to as *Bayesian spectral analysis regression* (BSAR), and Markov chain Monte Carlo (MCMC) methods were developed. In particular, they proposed a Bayesian method to estimate shape-restricted regression functions by assuming that the derivatives of the functions are squares of Gaussian processes. In regression models it is often the

case that subject matter knowledge imposes shape restrictions on the unknown regression functions, which can yield fitted models that are more interpretable and have improved performance compared to those without restrictions. The proposed model based on BSAR in Lenk and Choi (2017) was able to successfully deal with shape restrictions for the regression functions that are monotonic, monotonic convex or concave. Lenk (1999) considered related methods in the case without shape restriction. Lenk and Choi (2017) showed that their method is flexible for handling different kinds of shape restrictions and that it enjoys good performance compared to other Bayesian methods in the literature, for example, using spline smoothing (Shively, Sager and Walker, 2009; Meyer, Hackstadt and Hoeting, 2011), Bernstein polynomials (Curtis and Ghosh, 2011), and Gaussian processes (Lin and Dunson, 2014; Wang and Berger, 2016). The comparisons in Lenk and Choi (2017) are restricted to the univariate setting.

However, the approach of Lenk and Choi (2017) has a disadvantage for handling large data sets, mainly in requiring lengthy computation times for MCMC, especially in regression models with shape constraints. This is despite the fact that it is based on carefully designed MCMC algorithms resulting in methodology which is often faster than the alternative methods mentioned above, all of which are based on generic MCMC methods. Further, an `R` package is available for practitioners, using compiled `Fortran` code to maximize computational efficiency(Jo et al., 2017), but alternative numerical methods still need to be developed for real-time applications or large data sets. Variational Bayes (VB) methods are known to be fast deterministic alternatives to Markov chain Monte Carlo (MCMC) for Bayesian computation, facilitating approximate posterior inference for the parameters in complex statistical models (see, e.g., Waterhouse, Mackay and Robinson (1996), Jordan et al. (1999) and Attias (2000) for early developments of the method and Titterington (2004), Jordan (2004) and Ormerod and Wand (2010) for nontechnical overviews). In the nonparametric and semiparametric regression context, variational approximation schemes have found increasing use; for instance, real-time semiparametric regression (Wand and Ormerod, 2011), truncated power splines for partially linear additive models with variable selection (Zhao and Lian, 2014), penalized splines for mean and quantile regression in geoadditive latent Gaussian regression (Waldmann and Kneib, 2015), and sparse spectrum Gaussian process regression (Tan et al., 2016).

The objective of the current study is to develop fast variational Bayes computation methods for the semiparametric regression model of (1.1) using Gaussian process priors, which reduce computation time by an order of magnitude over the MCMC methods of Lenk and Choi (2017). Specifically, we provide variational Bayes approximation methods for spectral representations of one-dimensional Gaussian processes via the cosine basis expansion of (1.2). Further, we explore the possible use of the variational lower bound for model choice of a parametric regression model against a semiparametric alternative. In addition, we develop a variational Bayes approximation scheme to solve the computational challenges associated with MCMC with shape restrictions in a univariate nonlinear regression function, which is more challenging than the regression model

without shape restriction because of the non-conjugacy of many factors associated with those shape restrictions. To the best of our knowledge, there exists no variational Bayes approximation methods in the literature for shape-restricted regression models, and thus, our work is the first variational Bayes approach to the semiparametric regression with shape restrictions. This new approach is limited to one-dimensional Gaussian processes and shape constraints of monotonicity and convexity in the current work, but broadens the applicability of variational Bayes approximation in the context of Gaussian process regression modeling.

The rest of the paper is organized as follows. Section 2 provides a brief overview of variational Bayes approximation methods and reviews the basic model structure and the hierarchical prior specification proposed in Lenk and Choi (2017). Then, we develop a variational Bayes algorithm for fitting the unrestricted model, that is, BSAR without shape restriction. In Section 3, the shape restricted models, that is, BSAR with shape restrictions, are considered with monotonicity and convexity, and appropriate variational Bayes approximation schemes are developed. Section 4 illustrates the empirical performance of the proposed variational Bayes methods with simulation studies and real applications. Since these variational methods are approximate, we explore some of the trade-offs involved in using them in terms of speed, accuracy and automation of the implementation, in comparison with MCMC methods as well as other existing variational Bayes approximations for semiparametric regression in the literature. In Section 5, we discuss the potential and limitations of the methodology along with concluding remarks.

2. A variational Bayes approximation for a Bayesian spectral analysis regression model

2.1. An overview of variational Bayes methods

Consider a general Bayesian model with parameter vector $\boldsymbol{\delta}$, its prior density function $p(\boldsymbol{\delta})$, observation vector $\boldsymbol{y}$, and its assumed probability density function $p(\boldsymbol{y}|\boldsymbol{\delta})$. We assume that $\boldsymbol{y}$ and $\boldsymbol{\delta}$ are continuous for simplicity. Then, the posterior density function is given by

$$p(\boldsymbol{\delta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\delta})p(\boldsymbol{y}|\boldsymbol{\delta})}{p(\boldsymbol{y})},\ p(\boldsymbol{y}) = \int p(\boldsymbol{\delta})p(\boldsymbol{y}|\boldsymbol{\delta})d\boldsymbol{\delta},$$

where $p(\boldsymbol{y})$ is a marginal probability density function of $\boldsymbol{y}$.

For Bayesian inference with the posterior density $p(\boldsymbol{\delta}|\boldsymbol{y})$, which is often mathematically intractable, variational approximation methods (e.g. Jordan (2004), Titterington (2004), and Ormerod and Wand (2010)) can be employed. In these variational approximations, the posterior density $p(\boldsymbol{\delta}|\boldsymbol{y})$ is approximated by a density $q(\boldsymbol{\delta})$ from some tractable family, and $q(\boldsymbol{\delta})$ is chosen optimal in terms of minimization of the Kullback-Leibler (KL) divergence between $q(\boldsymbol{\delta})$ and $p(\boldsymbol{\delta}|\boldsymbol{y})$.

It is easy to see that

$$\log p(\boldsymbol{y}) = \mathcal{L}(q) + \int \log \frac{q(\boldsymbol{\delta})}{p(\boldsymbol{\delta}|\boldsymbol{y})} q(\boldsymbol{\delta}) d\boldsymbol{\delta} \tag{2.1}$$

where $\mathcal{L}(q) = \int \log \frac{p(\boldsymbol{\delta})p(\boldsymbol{y}|\boldsymbol{\delta})}{q(\boldsymbol{\delta})} q(\boldsymbol{\delta}) d\boldsymbol{\delta}$ is the *variational lower bound* (because it forms a lower bound on $\log p(\boldsymbol{y})$), and the second term in (2.1) is the KL divergence between $q(\boldsymbol{\delta})$ and $p(\boldsymbol{\delta}|\boldsymbol{y})$. The fact that $\mathcal{L}(q)$ is a lower bound clearly follows from (2.1) and the non-negativity of the KL divergence. Clearly maximizing $\mathcal{L}(q)$ with respect to $q(\cdot)$ is equivalent to minimizing the KL divergence term in (2.1).

The term variational Bayes(VB) is often used to denote variational inference when some kind of product restriction is made on the approximating distribution $q(\cdot)$ but where this distribution is otherwise arbitrary. This approach is also sometimes known as mean field variational Bayes (MFVB). By a product restriction we mean that we partition the parameter vector $\boldsymbol{\delta}$ into blocks, $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_k)$, and consider a density function $q(\cdot)$ that is assumed to factorize as $q(\boldsymbol{\delta}) = \prod_j q_j(\boldsymbol{\delta}_j)$. In variational Bayes a coordinate ascent approach is used to maximize $\mathcal{L}(q)$ by updating each term in $q(\boldsymbol{\delta}) = \prod_j q(\boldsymbol{\delta}_j)$ in turn with all other terms fixed. The update for $q_j(\boldsymbol{\delta}_j)$ takes the form

$$q_j(\boldsymbol{\delta}_j) \propto \exp\left(\mathrm{E}_{-j}(\log p(\boldsymbol{\delta}) p(\boldsymbol{y}|\boldsymbol{\delta}))\right) \tag{2.2}$$

where $\mathrm{E}_{-j}(\cdot)$ denotes expectation with respect to $\prod_{i \neq j} q_i(\boldsymbol{\delta}_i)$. If all the conditional distributions have a conjugate-exponential structure, then q_j takes the parametric form of an exponential family, and the variational update procedures are conveniently performed (Ghahramani and Beal, 2001).

A general algorithmic implementation of the procedure in this setting is given by the variational message passing algorithm of Winn and Bishop (2005). When there are nonconjugate factors in the model, one way to proceed is to use a generalization of variational message passing, namely the nonconjugate variational message passing (NCVMP) algorithm (Knowles and Minka, 2011; Wand, 2014). In NCVMP, for a factor $q_j(\boldsymbol{\delta}_j)$ having an intractable mean field update, it is assumed to have the parametric form of a natural exponential family

$$q_j(\boldsymbol{\delta}_j|\boldsymbol{\rho}_j) = \exp\left(\boldsymbol{\rho}_j^\top S_j(\boldsymbol{\delta}_j) - h_j(\boldsymbol{\rho}_j)\right), \tag{2.3}$$

where $\boldsymbol{\rho}_j$ are the vector of the natural parameters, $S_j(\boldsymbol{\delta}_j)$ are sufficient statistics of $\boldsymbol{\rho}_j$, and $h_j(\boldsymbol{\rho}_j)$ is a normalizing factor. A fixed-point updating procedure can then be derived, which reduces to the variational message passing update in the conjugate-exponential case. See Knowles and Minka (2011) and Wand (2014) for further details.

2.2. Bayesian spectral analysis regression (BSAR)

As briefly discussed in Section 1, the Bayesian spectral analysis regression (BSAR) model (Lenk, 1999; Lenk and Choi, 2017) expresses the Gaussian process as an infinite series expansion (1.2) and uses the cosine basis function on

$[0,1]$ as a choice of orthonormal system for the unknown nonparametric function f. In the semiparametric regression model of (1.1), the parametric term $\boldsymbol{w}_i^\top \boldsymbol{\beta}$ includes an intercept β_0, confounded with θ_0, and the basis function $\varphi_0(x)$ is dropped in the representation of f.

The infinite series in (1.2) is approximated by a finite sum $Z_J(x)$:

$$f(x) = Z(x) \approx Z_J(x) = \sum_{j=1}^{J} \theta_j \varphi_j(x), \tag{2.4}$$

where J denotes the truncation point. The mean integrated squared error between Z and Z_J decreases in J and can be made as small as desired because the sum of the variance is assumed to be finite, $\sum_{j=0}^{\infty} \nu_j^2 < \infty$, where $\nu_j^2 = \int_0^1 \int_0^1 \nu(s,t)\varphi_j(s)\varphi_j(t)dsdt$. Note that if the prior distribution of θ_j is inherited from Z by the spectral representation, then the choice of J does not considerably affect the accuracy of estimating f for sufficiently large J (Lenk (1999) and Lenk and Choi (2017)).

Using the approximation in (2.4), the BSAR model is expressed as $y_i = \boldsymbol{w}_i^\top \boldsymbol{\beta} + \sum_{j=1}^{J} \theta_j \varphi_j(x_i) + \epsilon_i,\ i = 1, \ldots, n$, and written in matrix notation,

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{\theta}_J^\top \boldsymbol{\varphi}_J + \boldsymbol{\epsilon}, \tag{2.5}$$

where

$$\begin{aligned} \boldsymbol{y} &= (y_1, \ldots, y_n)^\top \text{ and } \boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^\top \\ \boldsymbol{x} &= (x_1, \ldots, x_n)^\top \text{ and } \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top \\ \boldsymbol{\theta}_J &= (\theta_1, \ldots, \theta_J)^\top \text{ and } \boldsymbol{\varphi}_J = (\varphi_{ji})_{J\times n},\ \varphi_{ji} = \varphi_j(x_i). \end{aligned}$$

Then, based on the BSAR model structure in (2.5), the following hierarchical prior specification is considered for θ_j, $j \geq 1$, where a conditionally independent scale-invariant distribution is assigned to each of θ_j,

$$\begin{aligned} \theta_j|\sigma, \tau, \gamma &\sim \mathrm{N}\left(0, \sigma^2\tau^2 \exp[-j\gamma]\right) \text{ for } j \geq 1 \text{ and } \gamma > 0, \\ \tau^2 &\sim \mathrm{IG}\left(\frac{r_{0,\tau}}{2}, \frac{s_{0,\tau}}{2}\right) \text{ and } \gamma \sim \mathrm{Exp}(w_0). \end{aligned} \tag{2.6}$$

The prior probability that θ_j is in a neighborhood of zero increases with j and γ, and it decays to zero exponentially fast as indicated in (2.6). Further, we consider hyper priors on τ and γ in a hierarchical fashion (2.6), which allows the data to select the optimal smoothness given the data and structure of the model (Lenk and Choi, 2017). The prior specification is completed with a conjugate prior distribution for $\boldsymbol{\beta}$, which is also scale-invariant, $\boldsymbol{\beta} \sim N(\mu_\beta^0, \sigma^2\Sigma_\beta^0)$, and for σ^2, $\sigma^2 \sim \mathrm{IG}\left(\frac{r_{0,\sigma}}{2}, \frac{s_{0,\sigma}}{2}\right)$. All the remaining hyperparameters are assumed to be known.

2.3. A variational Bayes approximation for BSAR

In this subsection, we provide a variational Bayes approximation for the semiparametric regression model, BSAR, without any shape restriction on f. To be

specific, a variational Bayes approximation algorithm, `Algorithm 1`, is given based on the model structure of (2.5). The joint posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\theta}_J, \sigma^2, \tau^2, \psi)$ is approximated by a variational approximation with the product form of

$$q(\boldsymbol{\beta}, \boldsymbol{\theta}_J, \sigma^2, \tau^2, \psi) = q_1(\boldsymbol{\beta}) q_2(\boldsymbol{\theta}_J) q_3(\sigma^2) q_4(\tau^2) q_5(\psi), \quad |\psi| = \gamma. \tag{2.7}$$

In (2.7), we have introduced a new hyperparameter ψ instead of γ by the reparametrization of $|\psi| = \gamma$, and the corresponding prior distribution of ψ is given as the double exponential distribution, $\psi \sim \mathrm{DE}(0, w_0)$, with a density function $p(\psi) = 0.5 w_0 \exp(-w_0|\psi|)$, $-\infty < \psi < \infty$. Note that such a reparametrization in terms of ψ causes ψ only to be identifiable up to a sign change, but this is not a problem in practice as the variational optimization will lock on to one of the equivalent local modes. The reparametrization allows us to use a normal distribution for the variational approximation to the posterior distribution of ψ in the corresponding NCVMP variational updates. Although one could apply the NCVMP update directly to the parametrization of γ with, say, a gamma distribution, some unacceptable restrictions on the variational parameters are necessary for the existence of all the moments in the variational lower bound so that we avoid this approach here.

We use mean field variational updates for all the factors except for $q_5(\psi)$. That is, the mean field updates are based on the commonly used conjugate distributions for q_1–q_4; $q_1(\boldsymbol{\beta})$ is a normal distribution, parametrized as $N(\mu^q_\beta, \Sigma^q_\beta)$, $q_2(\boldsymbol{\theta}_J)$ is also a normal distribution, denoted as $N(\mu^q_\theta, \Sigma^q_\theta)$, $q_3(\sigma^2)$ is an inverse gamma distribution, denoted as $IG(r_{q,\sigma}/2, s_{q,\sigma}/2)$, and $q_4(\tau^2)$ is an inverse gamma denoted as $IG(r_{q,\tau}/2, s_{q,\tau}/2)$. These mean field updates are given in the Appendix. Here and in the Appendix we use the following notation. If $f(x) \propto g(x)$ for two functions $f(\cdot)$ and $g(\cdot)$, we write $\log f(x) \doteq \log g(x)$ to show that $\log f(x)$ and $\log g(x)$ differ by an additive constant not depending on x. Note that all the expectations in the Appendix, denoted by E_{-k}, $k = 1, 2, 3, 4$ are with respect to the marginal variational density for the parameters except for the parameters in the kth block under consideration.

Since the update for ψ is a non-standard one, we give some details here. For updating $q_5(\psi)$ we use an NCVMP update and assume $q_5(\psi)$ normal, $N(\mu^q_\psi, {\sigma^q_\psi}^2)$. In the derivation of NCVMP updates for ψ below, all the expectations denoted as E_5 are with respect to the full variational density. For ψ, applying the general procedures of the NCVMP algorithm, (Knowles and Minka, 2011) to $q_5(\psi)$, we first compute $S_k (\equiv S_k(\mu^q_\psi, {\sigma^q_\psi}^2))$, $k = 1, 2$ as given below:

$$\begin{aligned} S_1 &= \mathrm{E}_5(\log p(\psi)) \doteq -w_0 \mathrm{E}_5(|\psi|) \\ &= -w_0 \left\{ \sigma^q_\psi \sqrt{\frac{2}{\pi}} \exp\left(-\frac{{\mu^q_\psi}^2}{2{\sigma^q_\psi}^2} \right) + \mu^q_\psi \left(1 - 2\Phi\left(-\frac{\mu^q_\psi}{\sigma^q_\psi} \right) \right) \right\}, \end{aligned}$$

$$\begin{aligned} S_2 &= \mathrm{E}_5(\log p(\boldsymbol{\theta}_J|\sigma^2,\tau^2,\psi)) \\ &\doteq -\frac{1}{2}\mathrm{E}_5\left(\frac{1}{\sigma^2}\right)\mathrm{E}_5\left(\frac{1}{\tau^2}\right)\sum_{j=1}^{\top}(\Sigma^q_{\theta,jj}+{\mu^q_{\theta,j}}^2)Q_j(\mu^q_\psi,{\sigma^q_\psi}^2) \\ &\quad +\frac{1}{2}\mathrm{E}_5(|\psi|)\frac{J(J+1)}{2} \\ &\doteq -\frac{1}{2}\frac{r_{q,\sigma}}{s_{q,\sigma}}\frac{r_{q,\tau}}{s_{q,\tau}}\sum_{j=1}^{J}(\Sigma^q_{\theta,jj}+{\mu^q_{\theta,j}}^2)Q_j(\mu^q_\psi,{\sigma^q_\psi}^2)-\frac{J(J+1)}{4w_0}S_1. \end{aligned}$$

where the expression for $Q_j(\cdot)$ is given in the Appendix and μ^q_ψ and σ^q_ψ are then updated by

$${\sigma^q_\psi}^2 \leftarrow -\frac{1}{2}\left\{\frac{\partial S_1}{\partial {\sigma^q_\psi}^2}+\frac{\partial S_2}{\partial {\sigma^q_\psi}^2}\right\}^{-1}, \quad \mu^q_\psi \leftarrow \mu^q_\psi+{\sigma^q_\psi}^2\left\{\frac{\partial S_1}{\partial \mu^q_\psi}+\frac{\partial S_2}{\partial \mu^q_\psi}\right\}.$$

In addition to deriving updates for the variational factors, we also require an expression for the variational lower bound, $\mathcal{L}(q)$ in (2.1), which is specifically given by

$$\begin{aligned} \mathrm{E}(\log p(\boldsymbol{y},\boldsymbol{\delta})) &= \mathrm{E}(\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta}_J,\sigma^2))+\mathrm{E}(\log p(\boldsymbol{\beta}|\sigma^2)) \\ &\quad +\mathrm{E}(\log p(\boldsymbol{\theta}_J|\sigma^2,\tau^2,\psi)) \\ &\quad +\mathrm{E}(\log p(\psi))+\mathrm{E}(\log p(\sigma^2))+\mathrm{E}(\log p(\tau^2)), \qquad (2.8) \\ \mathrm{E}(\log q(\boldsymbol{\delta})) &= \mathrm{E}(\log q_1(\boldsymbol{\beta}))+\mathrm{E}(\log q_2(\boldsymbol{\theta}_J))+\mathrm{E}(\log q_3(\sigma^2)) \\ &\quad +\mathrm{E}(\log q_4(\tau^2))+\mathrm{E}(\log q_5(\psi)). \qquad (2.9) \end{aligned}$$

The terms in the expressions above are given in the Appendix. The variational lower bound $\mathcal{L}(q)$ is used for two purposes, one being for defining the stopping rule in the variational Bayes approximation algorithm and the other being for model selection for choosing between two competing models. For the use of the stopping rule, the variational Bayesian algorithm, as given in `Algorithm 1` below, is terminated when the increase of the lower bound of the log-likelihood (2.1) is negligible. Further, we use the lower bound of the log-likelihood as an approximation of the marginal likelihood for testing the adequacy of semiparametric regression model in the empirical analysis presented in Section 4.

Based on all the above development, we now provide the following variational Bayes approximation algorithm for BSAR without restriction, `Algorithm 1`, describing the updates of all the variational parameters ϑ in the approximate distribution of $(\boldsymbol{\beta},\boldsymbol{\theta}_J,\sigma^2,\tau^2,\psi)$. Here, $\vartheta=\left\{\mu^q_\beta,\Sigma^q_\beta,\mu^q_\psi,{\sigma^q_\psi}^2,r_{q,\tau},s_{q,\tau},r_{q,\sigma},s_{q,\sigma},\mu^q_\theta,\Sigma^q_\theta\right\}$ denotes the set of variational parameters in the approximate distribution of $(\boldsymbol{\beta},\boldsymbol{\theta}_J,\sigma^2,\tau^2,\psi)$.

In implementing `Algorithm 1`, note that the update for Σ^q_θ often results in a numerically singular matrix because the shrinkage spectral coefficients $\boldsymbol{\theta}_J$ essentially degenerate at zero. Thus, for numerical stability, we set such coefficients exactly to zero in implementing the scheme numerically, which effectively corresponds to a change of the truncation point J.

Algorithm 1

Input: Data $\boldsymbol{y}$, tolerance tol, prior parameters;

Output: Optimized variational parameters μ_β^q, Σ_β^q, μ_ψ^q, ${\sigma_\psi^q}^2$, $r_{q,\tau}, s_{q,\tau}$, $r_{q,\sigma}$, $s_{q,\sigma}$, μ_θ^q, Σ_θ^q and the corresponding lower bound value $\mathcal{L}(q)$.

Initialize ϑ:

$$\mu_\psi^q \leftarrow 0,\ {\sigma_\psi^q}^2 \leftarrow 0,\ r_{q,\sigma} \leftarrow r_{0,\sigma} + J + p + n,\ r_{q,\tau} \leftarrow r_{0,\tau} + J,$$
$$s_{q,\sigma} \leftarrow,\ s_{0,\sigma},\ s_{q,\tau} \leftarrow s_{0,\tau},\ \mu_\beta^q \leftarrow \mu_\beta^0,\ L_{old} = -\infty,\ dif = tol + 1;$$

While $dif > tol$ **do**

$$\Sigma_\theta^q \leftarrow \left(\frac{r_{q,\sigma}}{s_{q,\sigma}}\boldsymbol{\varphi}_J^\top\boldsymbol{\varphi}_J + \frac{r_{q,\sigma}}{s_{q,\sigma}}\frac{r_{q,\tau}}{s_{q\tau}}\mathrm{diag}\left(\mathrm{E}(\Gamma^{-1})\right)\right)^{-1},\ \mu_\theta^q \leftarrow \frac{r_{q,\sigma}}{s_{q,\sigma}}\Sigma_\theta^q\boldsymbol{\varphi}_J^\top(\boldsymbol{y} - \boldsymbol{W}\mu_\beta^q),$$
$$s_{q,\sigma} \leftarrow s_{0,\sigma} + \frac{r_{q,\tau}}{s_{q,\tau}}\mathrm{tr}\left(\left(\Sigma_\theta^q + \mu_\theta^q\mu_\theta^{q\top}\right)\mathrm{diag}(\mathrm{E}(\Gamma^{-1}))\right) + \mathrm{tr}(\boldsymbol{W}^\top\boldsymbol{W}\Sigma_\beta^q) +$$
$$\mathrm{tr}(\boldsymbol{\varphi}_J^\top\boldsymbol{\varphi}_J\Sigma_\theta^q) + \mathrm{tr}({\Sigma_\beta^0}^{-1}\Sigma_\beta^q) + (\boldsymbol{y} - \boldsymbol{W}\mu_\beta^q - \boldsymbol{\varphi}_J\mu_\theta^q)^\top(\boldsymbol{y} - \boldsymbol{W}\mu_\beta^q - \boldsymbol{\varphi}_J\mu_\theta^q) +$$
$$(\mu_\beta^q - \mu_\beta^0)^\top{\Sigma_\beta^0}^{-1}(\mu_\beta^q - \mu_\beta^0),$$
$$s_{q,\tau} \leftarrow s_{0,\tau} + \frac{r_{q,\sigma}}{s_{q,\sigma}}\mathrm{tr}\left(\left(\Sigma_\theta^q + \mu_\theta^q\mu_\theta^{q\top}\right)\mathrm{diag}(\mathrm{E}(\Gamma^{-1}))\right),$$
$$\Sigma_\beta^q \leftarrow \frac{s_{q,\sigma}}{r_{q,\sigma}}\left(\boldsymbol{W}^\top\boldsymbol{W} + {\Sigma_\beta^0}^{-1}\right)^{-1},\ \mu_\beta^q \leftarrow \frac{r_{q,\sigma}}{s_{q,\sigma}}\Sigma_\beta^q\left({\Sigma_\beta^0}^{-1}\mu_\beta^0 + \boldsymbol{W}^\top(\boldsymbol{y} - \boldsymbol{\varphi}_J\mu_\theta^q)\right),$$
$${\sigma_\psi^q}^2 \leftarrow -\frac{1}{2}\left\{\frac{\partial S_1}{\partial {\sigma_\psi^q}^2} + \frac{\partial S_2}{\partial {\sigma_\psi^q}^2}\right\}^{-1},\quad \text{where } S_j = S_j(\mu_\psi^q, {\sigma_\psi^q}^2),\ j = 1, 2,$$
$$\mu_\psi^q \leftarrow \mu_\psi^q + {\sigma_\psi^q}^2\left\{\frac{\partial S_1}{\partial \mu_\psi^q} + \frac{\partial S_2}{\partial \mu_\psi^q}\right\},$$
$$L_{new} = \mathcal{L}(q),\ dif \leftarrow L_{new} - L_{old},$$
$$L_{old} \leftarrow L_{new};$$

end

3. Variational Bayes approximations for the shape-restricted models

In this section, we consider the shape restricted regression models, that is, BSAR with shape restrictions of monotonicity and concavity, and develop appropriate variational approximation schemes for them. As discussed in Section 1, there have been several methods proposed on Bayesian shape-restricted regression, all of which use MCMC methods (see, e.g, Shively, Sager and Walker (2009), Meyer, Hackstadt and Hoeting (2011), Curtis and Ghosh (2011), Lenk and Choi (2017), and the references therein), and no results have been discussed in the context of variational approximation.

Here, we focus on the shape restricted regression models of Lenk and Choi (2017), BSAR with shape restrictions, in which the derivatives of the regression functions are modelled in terms of squares of Gaussian processes for shape constraints, based on their spectral representations. That is, the proposed approach of Lenk and Choi (2017) enforces shape restrictions on the ℓth derivative of f

as the square of a Gaussian process $Z(x)$ in (1.2),

$$f^{(\ell)}(x) = \delta Z^2(x) \tag{3.1}$$

where $\delta \in \{-1, 1\}$ and ℓ are given by the user. For example, when ℓ is 1 and δ is 1, f is non-decreasing, and f is a non-decreasing and convex function when ℓ is 2 and δ is 1. The proposed method of Lenk and Choi (2017) based on the characterization of (3.1) was shown to be flexible for handling different kinds of shape restrictions and to have good performance compared to other Bayesian methods in the literature. We provide fast and efficient variational Bayes approximation methods for monotonic, monotonic convex or monotonic concave regression models based on the framework of (3.1).

3.1. A Variational Bayes approximation for the monotone function

We first consider the shape-restricted model with monotone regression functions and develop its variational approximation algorithm. The derivative representation in (3.1) for the monotone function with $\ell = 1$ is rewritten in terms of the regression function $f(\cdot)$ by integration

$$f(x) = \delta \left[\int_0^x Z^2(s)ds - \int_0^1 \int_0^x Z^2(s)ds\, dx \right] \tag{3.2}$$

where δ is 1 for a non-decreasing function and -1 for a non-increasing function, and the last term is chosen to satisfy the mean-centering condition of $f(\cdot)$ (Lenk and Choi, 2017). Then, using the spectral representation of $Z(x)$ in (1.2), $f(x)$ is expanded as

$$\begin{aligned} f(x) &= \delta \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \theta_j \theta_k \varphi^a_{j,k}(x) \qquad (3.3) \\ \varphi^a_{j,k}(x) &= \int_0^x \varphi_j(s)\bar{\varphi}_k(s)ds - \int_0^1 \int_0^s \varphi_j(t)\bar{\varphi}_k(t)dt\, ds \text{ for } j,k \geq 0, \end{aligned}$$

where $\varphi^a_{j,k}(x)$ using the cosine basis are specifically given as (Lenk and Choi, 2017):

$$\begin{aligned} \varphi^a_{0,0}(x) &= x - 0.5 \\ \varphi^a_{0,j}(x) &= \varphi^a_{j,0}(x) = \frac{\sqrt{2}}{\pi j} \sin(\pi j x) - \frac{\sqrt{2}}{(\pi j)^2} \left[1 - \cos(\pi j)\right] \text{ for } j \geq 1, \\ \varphi^a_{j,j}(x) &= \frac{\sin(2\pi j x)}{2\pi j} + x - 0.5 \text{ for } j \geq 1, \\ \varphi^a_{j,k}(x) &= \frac{\sin\left[\pi(j+k)x\right]}{\pi(j+k)} + \frac{\sin\left[\pi(j-k)x\right]}{\pi(j-k)} \\ &\quad - \frac{1 - \cos[\pi(j+k)]}{[\pi(j+k)]^2} - \frac{1 - \cos[\pi(j-k)]}{[\pi(j-k)]^2} \\ &\quad \text{for } j \neq k \text{ and } j,k \geq 1. \end{aligned}$$

Thus, the semiparametric model (1.1) with monotone restriction is written in matrix notation as

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{\beta} + \delta\boldsymbol{\theta}_J^\top \boldsymbol{\varphi}_J^a(x)\boldsymbol{\theta}_J + \boldsymbol{\epsilon}, \tag{3.4}$$

where $\boldsymbol{\theta}_J = (\theta_0, \ldots, \theta_J)^\top$ is the $J+1$ vector of spectral coefficients, and $\boldsymbol{\varphi}_J^a(x)$ is a $J+1 \times J+1$ matrix with (j,k) entry $\varphi_{j,k}^a(x)$. The parameters in the model are the same as in the case with the unrestricted model $(\boldsymbol{\beta}, \boldsymbol{\theta}_J, \sigma^2, \tau^2, \psi)$. Thus, we adopt the same hierarchical prior specification of (2.6) in the case with BSAR without shape restrictions in Section 2 except for the prior distributions of θ_j, $j \geq 0$,

$$\theta_0|\sigma \sim N(0, \sigma\sigma_0^2), \text{ and } \theta_j|\sigma, \tau, \gamma \sim N(0, \sigma\tau^2 \exp[-j\gamma]) \tag{3.5}$$

That is, in the prior on θ_j, $j \geq 0$, to ensure scale-invariant prior specification as discussed in Lenk and Choi (2017), σ rather than σ^2 appears in the variance, in contrast to the BSAR without restrictions. Note that we do not consider the identifiability condition $\theta_0 \geq 0$ of Lenk and Choi (2017) in the variational Bayes approximation scheme. The optimization in the variational approximation in general locks on to one of the two equivalent modes obtained by switching the signs of all elements of $\boldsymbol{\theta}_J$.

In the variational Bayes approximation to the joint posterior distribution for the regression model with monotone restriction in (3.4), we use mean field updates for $\boldsymbol{\beta}, \sigma^2$ and τ^2 and an NCVMP update for ψ as before, but an NCVMP update with a normal factor for $\boldsymbol{\theta}_J$, in contrast to the case of BSAR without shape restrictions, because of the non-conjugacy for $\boldsymbol{\theta}_J$ with the characterization of the squared Gaussian processes in (3.1) and the scale-invariant prior specification of σ in (3.5).

The assumed form of the variational approximation in terms of the blocks $(\boldsymbol{\beta}, \boldsymbol{\theta}_J, \tau^2, \psi)$ is similar to BSAR without restrictions. $q_1(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$ is parametrized as $N(\mu_\beta^q, \Sigma_\beta^q)$, the factor $q_2(\boldsymbol{\theta}_J)$ for $\boldsymbol{\theta}_J$ is parametrized as $N(\mu_\theta^q, \Sigma_\theta^q)$, the factor $q_4(\tau^2)$ for τ^2 is inverse gamma, $IG(\frac{r_{q,\tau}}{2}, \frac{s_{q,\tau}}{2})$, and the factor $q_5(\psi)$ for ψ is parametrized as $N(\mu_\psi^q, {\sigma_\psi^q}^2)$. Note that the factor $q_3(\sigma^2)$ for the mean field update is not an inverse gamma but takes a different form because of the prior specification in (3.5) as mentioned before, with details given below. The mean field updates for β and τ^2 are described in the Appendix. Again we describe the non-standard non-conjugate updates in some detail and again we note that the expectations denoted by E_{-k} are with respect to the marginal variational density for the parameters except for the parameters in the kth block under consideration, and the expectations denoted by E_k are with respect to the full variational density.

We provide the details of the updating procedures as follows:

- For $\boldsymbol{\theta}_J$, the mean field update does not take the form of a standard distribution for monotone restrictions, and we use a multivariate normal approximation with the parameters updated by the NCVMP algorithm (Knowles and Minka, 2011; Wand, 2014). Specifically, define $S_k(\mu_\theta^q, \Sigma_\theta^q)$,

$k = 1, 2$,

$$\begin{aligned} S_1(\mu_\theta^q, \Sigma_\theta^q) &= \mathrm{E}_2\left(\log p(\boldsymbol{\theta}_J|\sigma^2, \tau^2, \psi)\right) \\ &\doteq -\frac{1}{2}\mathrm{E}_2\left(\frac{1}{\sigma}\right)\mathrm{tr}\left(\left(\Sigma_\theta^q + \mu_\theta^q \mu_\theta^{q\top}\right)\mathrm{diag}(\mathrm{E}_2(\Upsilon^{-1}))\right), \\ S_2(\mu_\theta^q, \Sigma_\theta^q) &= \mathrm{E}_2(\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}_J, \sigma^2)) \\ &\doteq -\frac{1}{2}\mathrm{E}_2\left(\frac{1}{\sigma^2}\right)\sum_i \Big\{(y_i - \boldsymbol{w}_i^\top \mu_\beta^q - \delta\mathrm{tr}(\Sigma_\theta^q \boldsymbol{\varphi}_J^a(x_i)) \\ &\quad -\delta\mu_\theta^{q\top}\boldsymbol{\varphi}_J^a(x_i)\mu_\theta^q)^2 + 2\mathrm{tr}(\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q \boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q) \\ &\quad +4\mu_\theta^{q\top}\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q \boldsymbol{\varphi}_J^a(x_i)\mu_\theta^q\Big\}, \end{aligned}$$

where

$$\begin{aligned} \Upsilon &= \left(\sigma_0^2, \tau^2\exp(-\gamma), \ldots, \tau^2\exp(-J\gamma)\right)^\top, \\ \mathrm{E}_2(\Upsilon^{-1}) &= (1/\sigma_0^2, r_{q,\tau}/s_{q,\tau}\mathrm{E}_2(\Gamma^{-1})). \end{aligned}$$

Then it follows from Wand (2014) that the NCVMP update takes the form

$$\Sigma_\theta^q \leftarrow -\frac{1}{2}\left\{\sum_{a=1}^2 \frac{\partial S_a(\mu_\theta^q, \Sigma_\theta^q))}{\partial \Sigma_\theta^q}\right\}^{-1}, \quad \mu_\theta^q \leftarrow \mu_\theta^q + \Sigma_\theta^q\left\{\sum_{a=1}^2 \frac{\partial S_a(\mu_\theta^q, \Sigma_\theta^q))}{\partial \mu_\theta^q}\right\}.$$

Using standard rules of matrix differential calculus, we obtain the NCVMP update for $\boldsymbol{\theta}_J$ given in Algorithm 2.

- For σ^2, the mean-field update is given as

$$\begin{aligned} \log q_3(\sigma^2) \doteq \mathrm{E}_{-3}(&\log p(\sigma^2) + \log p(\boldsymbol{\theta}_J|\sigma^2, \tau^2, \psi) + \log p(\boldsymbol{\beta}|\sigma^2) \\ &+ \log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}_J, \sigma^2)), \end{aligned}$$

where

$$\begin{aligned} \mathrm{E}_{-3}(\log p(\sigma^2)) &\doteq -\left(\frac{r_{0,\sigma}}{2} + 1\right)\log\sigma^2 - \frac{s_{0,\sigma}}{2\sigma^2}, \\ \mathrm{E}_{-3}(\log p(\boldsymbol{\theta}_J|\sigma^2, \tau^2, \psi) &\doteq -\frac{(J+1)}{2}\log\sigma - \frac{1}{2\sigma}\mathrm{E}_{-3}(\boldsymbol{\theta}_J^\top \mathrm{diag}(\Upsilon^{-1})\boldsymbol{\theta}_J) \\ &\doteq -\frac{(J+1)}{2}\log\sigma \\ &\quad -\frac{1}{2\sigma}\mathrm{tr}\left\{\left(\Sigma_\theta^q + \mu_\theta^q\mu_\theta^{q\top}\right)\mathrm{diag}(\mathrm{E}_{-3}(\Upsilon^{-1}))\right\}, \\ \mathrm{E}_{-3}(\log p(\boldsymbol{\beta}|\sigma^2)) &\doteq -\frac{p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\mathrm{E}((\boldsymbol{\beta} - \mu_\beta^0)^\top \Sigma_\beta^{0\,-1}(\boldsymbol{\beta} - \mu_\beta^0)) \\ &\doteq -\frac{p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left\{(\mu_\beta^q - \mu_\beta^0)^\top \Sigma_\beta^{0\,-1}(\mu_\beta^q - \mu_\beta^0)\right. \\ &\quad \left.+\mathrm{tr}(\Sigma_\beta^{0\,-1}\Sigma_\beta^q)\right\}, \end{aligned}$$

$$\begin{aligned}\mathrm{E}_{-3}(\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta}_J,\sigma^2)) \;\doteq\; & -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_i \Big\{(y_i - \boldsymbol{w}_i^\top \mu_\beta^q \\ & -\delta\mathrm{tr}(\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q) - \delta {\mu_\theta^q}^\top \boldsymbol{\varphi}_J^a(x_i)\mu_\theta^q)^2 \\ & +\boldsymbol{w}_i^\top \Sigma_\beta^q \boldsymbol{w}_i + 2\mathrm{tr}(\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q) \\ & +4{\mu_\theta^q}^\top \boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q\boldsymbol{\varphi}_J^a(x_i)\mu_\theta^q\Big\}.\end{aligned}$$

Hence

$$\begin{aligned}\log q(\sigma^2) \doteq & -\left\{\frac{r_{0,\sigma}+n+p+\frac{J+1}{2}}{2}+1\right\}\log\sigma^2 \\ & -\frac{1}{2\sigma^2}\Big(s_{0,\sigma} + (\mu_\beta^q-\mu_\beta^0)^\top {\Sigma_\beta^0}^{-1}(\mu_\beta^q-\mu_\beta^0) + \mathrm{tr}({\Sigma_\beta^0}^{-1}\Sigma_\beta^q) \\ & +\sum_{i=1}^n\Big\{(y_i - \boldsymbol{w}_i^\top\mu_\beta^q - \delta\mathrm{tr}(\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q) - \delta{\mu_\theta^q}^\top\boldsymbol{\varphi}_J^a(x_i)\mu_\theta^q)^2 \\ & +\boldsymbol{w}_i^\top\Sigma_\beta^q\boldsymbol{w}_i + 2\mathrm{tr}(\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q). \\ & +4{\mu_\theta^q}^\top\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q\boldsymbol{\varphi}_J^a(x_i)\mu_\theta^q\Big\}\Big) \\ & -\frac{1}{2\sigma}\mathrm{tr}\Big\{(\Sigma_\theta^q+\mu_\theta^q{\mu_\theta^q}^\top)\mathrm{diag}(\mathrm{E}_{-3}(\Upsilon^{-1}))\Big\}.\end{aligned}$$

Thus,

$$q(\sigma^2) \propto \left(\frac{1}{\sigma}\right)^{2a}\exp\left(\frac{b}{\sigma}-\frac{c}{\sigma^2}\right), \tag{3.6}$$

where

$$\begin{aligned}a &= \frac{r_{0,\sigma}+n+p+(J+1)/2}{2}+1, \\ b &= -\frac{1}{2}\mathrm{tr}\Big\{(\Sigma_\theta^q+\mu_\theta^q{\mu_\theta^q}^\top)\mathrm{diag}(\mathrm{E}_{-3}(\Upsilon^{-1}))\Big\}, \\ c &= \frac{1}{2}\Big(s_{0,\sigma} + 2\mathrm{tr}({\Sigma_\beta^0}^{-1}\Sigma_\beta^q) + (\mu_\beta^q-\mu_\beta^0)^\top{\Sigma_\beta^0}^{-1}(\mu_\beta^q-\mu_\beta^0) \\ &\quad +\sum_{i=1}^n\Big[(y_i-\boldsymbol{w}_i^\top\mu_\beta^q - \delta\mathrm{tr}(\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q) - \delta{\mu_\theta^q}^\top\boldsymbol{\varphi}_J^a(x_i)\mu_\theta^q)^2 \\ &\quad +2\mathrm{tr}(\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q) + 4{\mu_\theta^q}^\top\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q\boldsymbol{\varphi}_J^a(x_i)\mu_\theta^q + \boldsymbol{w}_i^\top\Sigma_\beta^q\boldsymbol{w}_i\Big]\Big).\end{aligned}$$

- The NCVMP update for $q_5(\psi)$ takes the same form as in Section 2.3, except that the term $r_{q,\sigma}/s_{q,\sigma}$ in $S_2(\mu_\psi^q, {\sigma_\psi^q}^2)$ is replaced by $\mathrm{E}_5(1/\sigma)$.

Details of the lower bound calculation are given in the Appendix. Based on all the above computations and notation, we provide the variational Bayes algorithm for BSAR with monotone restriction in `Algorithm 2`.

Algorithm 2

Input: Data $\boldsymbol{y}$, tolerance tol, prior parameters;

Output: Optimized variational parameters μ_β^q, Σ_β^q, μ_ψ^q, $\sigma_\psi^{q\,2}$, $r_{q,\tau}, s_{q,\tau}, r_{q,\sigma}, s_{q,\sigma}$, μ_θ^q, Σ_θ^q and the corresponding lower bound value $\mathcal{L}(q)$.

Initialize ϑ:

$$\mu_\psi^q \leftarrow 0,\ \sigma_\psi^{q\,2} \leftarrow 0,\ r_{q,\tau} \leftarrow r_{0,\tau} + J,\ s_{q,\tau} \leftarrow s_{0,\tau},\ \mu_\beta^q \leftarrow \mu_\beta^0,$$
$$\mathrm{E}\left(\frac{1}{\sigma}\right) \leftarrow \sqrt{\frac{a_{0,\sigma}}{b_{0,\sigma}}},\ \mathrm{E}\left(\frac{1}{\sigma^2}\right) \leftarrow \frac{a_{0,\sigma}}{b_{0,\sigma}},\ L_{new} = -\infty,\ dif = tol + 1;$$

While $dif > tol$ **do**

$$\Sigma_\theta^q \leftarrow -\frac{1}{2}\left\{-\frac{1}{2}\mathrm{E}\left(\frac{1}{\sigma}\right)\mathrm{diag}(\mathrm{E}(\Upsilon^{-1}))\right.$$
$$-\frac{1}{2}E\left(\frac{1}{\sigma^2}\right)\sum_i\left[6\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q\psi(x_i) + 4\boldsymbol{\varphi}_J^a(x_i)\mu_\theta^q\mu_\theta^{q\top}\boldsymbol{\varphi}_J^a(x_i)\right.$$
$$\left.\left.-2\delta(y_i - \boldsymbol{w}_i^\top\mu_\beta^q - \delta\mu_\theta^{q\top}\boldsymbol{\varphi}_J^a(x_i)\mu_\theta^q)\boldsymbol{\varphi}_J^a(x_i)\right]\right\}^{-1},$$
$$\mu_\theta^q \leftarrow \mu_\theta^q + \Sigma_\theta^q\left\{-\mathrm{E}\left(\frac{1}{\sigma}\right)\mathrm{diag}(\mathrm{E}(\Upsilon^{-1}))\mu_\theta^q - \frac{1}{2}\mathrm{E}\left(\frac{1}{\sigma^2}\right)\left[8\sum_i\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q\boldsymbol{\varphi}_J^a(x_i)\mu_\theta^q\right.\right.$$
$$\left.\left.-4\sum_i(y_i - \boldsymbol{w}_i^\top\mu_\beta^q - \delta\mathrm{tr}(\Sigma_\theta^q\boldsymbol{\varphi}_J^a(x_i)) - \delta\mu_\theta^{q\top}\boldsymbol{\varphi}_J^a(x_i)\mu_\theta^q)\boldsymbol{\varphi}_J^a(x_i)\mu_\theta^q\right]\right\},$$
$$s_{q,\tau} \leftarrow s_{0,\tau} + \mathrm{E}\left(\frac{1}{\sigma}\right)\mathrm{tr}\left(\left(\Sigma_\theta^{q*} + \mu_\theta^{q*}\mu_\theta^{q*\top}\right)\mathrm{diag}(\mathrm{E}(\Gamma^{-1}))\right),$$
$$\Sigma_\beta^q \leftarrow \mathrm{E}\left(\frac{1}{\sigma^2}\right)^{-1}\left(\boldsymbol{W}^\top\boldsymbol{W} + \Sigma_\beta^{0\,-1}\right)^{-1},$$
$$\mu_\beta^q \leftarrow \Sigma_\beta^q\mathrm{E}\left(\frac{1}{\sigma^2}\right)\left(\Sigma_\beta^{0\,-1}\mu_\beta^0 + \sum_{i=1}^n\boldsymbol{w}_i(y_i - \delta\mathrm{tr}(\boldsymbol{\varphi}_J^a(x_i)\Sigma_\theta^q - \delta\mu_\theta^{q\top}\boldsymbol{\varphi}_J^a(x_i)\mu_\theta^q)\right),$$
$$\sigma_\psi^{q\,2} \leftarrow -\frac{1}{2}\left\{\frac{\partial S_1}{\partial\sigma_\psi^{q\,2}} + \frac{\partial Q_2}{\partial\sigma_\psi^{q\,2}}\right\}^{-1},\ \mu_\psi^q \leftarrow \mu_\psi^q + \sigma_\psi^{q\,2}\left\{\frac{\partial S_1}{\partial\mu_\psi^q} + \frac{\partial Q_2}{\partial\mu_\psi^q}\right\},$$
$$L_{new} = \mathcal{L}(q),\ dif \leftarrow L_{new} - L_{old},$$
$$L_{old} \leftarrow L_{new};$$

end

Note once again that in `Algorithm 2`, μ_θ^{q*} is μ_θ^q with the first component removed, Σ_θ^{q*} is Σ_θ^q with the first row and column removed, and $Q_2 = Q_2(\mu_\psi^q, \sigma_\psi^{q\,2})$ is the same as $S_2(\mu_\psi^q, \sigma_\psi^{q\,2})$ with $r_{q,\sigma}/s_{q,\sigma}$ replaced by $\mathrm{E}_{-2}\left(\frac{1}{\sigma}\right)$. How to compute this last expectation is discussed in the Appendix. As in `Algorithm 1`, the update for Σ_θ^q often results in a numerically singular matrix, and we set coefficients with a prior degenerate on zero exactly to zero in implementation, which as we mentioned earlier effectively corresponds to a change of the truncation point J.

3.2. Variational Bayes approximations for the convex/concave function

We next consider variational Bayes approximation methods for convex or concave regression functions. When a twice-differentiable function is assumed to be either monotonic convex or monotonic concave, then the first and second derivatives of the function have the same sign. Then, from the shape-restricted representation of (3.1), the second derivative of f is modeled as the square of a Gaussian process $Z(x)$. That is, when ℓ is 2, f is a non-decreasing and convex function when $\delta = 1$ or non-increasing and concave function when $\delta = -1$, and f is represented as (Lenk and Choi, 2017):

$$f(x) = \delta \left[\int_0^x \int_0^s Z^2(t)dt\, ds - \int_0^1 \int_0^x \int_0^s Z^2(t)dt\, ds\, dx \right] + \alpha(x - 0.5). \quad (3.7)$$

Notice that if we take the first and second derivatives of $f(x)$ in (3.7), we get

$$f'(x) = \delta \int_0^x Z^2(s)ds + \alpha \text{ and } f''(x) = \delta Z^2(x),$$

and that $\delta\alpha \geq 0$ ensures monotonicity. Similar to the monotone restriction, the spectral representation of $f(x)$ with monotone convexity or concavity in (3.7) becomes

$$\begin{aligned} f(x) &= \delta \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \theta_j \theta_k \varphi^b_{j,k}(x) + \alpha(x - 0.5) \quad (3.8) \\ \varphi^b_{j,k}(x) &= \int_0^x \int_0^s \varphi_j(t)\varphi_k(t)dt\, ds - \int_0^1 \int_0^x \int_0^s \varphi_j(t)\varphi_k(t)dt\, ds\, dx. \end{aligned}$$

Then, the resulting basis, $\varphi^b_{j,k}$, is obtained as (Lenk and Choi, 2017):

$$\begin{aligned} \varphi^b_{0,0}(x) &= \frac{3x^2 - 1}{6} \\ \varphi^b_{0,j}(x) &= \varphi^b_{j,0}(x) = -\frac{\sqrt{2}}{(\pi j)^2} \cos(\pi j x) \text{ for } j \geq 1 \\ \varphi^b_{j,j}(x) &= -\frac{\cos(2\pi j x)}{(2\pi j)^2} + \frac{3x^2 - 1}{6} \text{ for } j \geq 1 \\ \varphi^b_{j,k}(x) &= -\frac{\cos[\pi(j + k)x]}{[\pi(j + k)]^2} - \frac{\cos[\pi(j - k)x]}{[\pi(j - k)]^2}, \quad \text{for } j \neq k \text{ and } j, k \geq 1. \end{aligned}$$

In order to develop the variational Bayes approximation method, instead of using the spectral representation of $f(x)$ in (3.8), we replace α with $\delta\alpha^2$ as an equivalent representation of $f(x)$,

$$f(x) = \delta \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \theta_j \theta_k \varphi^b_{j,k}(x) + \delta\alpha^2(x - 0.5), \quad (3.9)$$

in which the monotonicity and the convexity of $f(x)$ are controlled by a single parameter δ. Then, we have the following model structure for monotonic convex/concave regression functions, in which the variational Bayes approximation method is explored:

$$y_i = \boldsymbol{w}_i^\top \boldsymbol{\beta} + \delta(\boldsymbol{\theta}_J^\alpha)^\top \boldsymbol{\varphi}_J^{b,\alpha}(x_i)\boldsymbol{\theta}_J^\alpha + \epsilon_i \tag{3.10}$$

where $\boldsymbol{\theta}_J^\alpha = (\alpha, \theta_0, \theta_1, ..., \theta_J)^\top$ and $\boldsymbol{\varphi}_J^{b,\alpha}$ a $(J+2)\times(J+2)$ matrix with

$$\begin{aligned}
\varphi_{0,0}^{b,\alpha}(x) &= x - 0.5 \\
\varphi_{0,j}^{b,\alpha}(x) &= \varphi_{j,0}^{b,\alpha}(x) = 0 \text{ for } j \geq 1 \\
\varphi_{j,k}^{b,\alpha}(x) &= \varphi_{j-1,k-1}^{b}(x) \text{ for } j \geq 1.
\end{aligned}$$

By reformulating the model (3.10) with α^2 instead of α, we use a normal distribution for the variational approximation to the posterior distribution of α, based on a normal prior $\alpha \sim N(0, \sigma\sigma_{0,\alpha}^2)$, instead of a truncated normal prior considered in Lenk and Choi (2017). For the remaining parameters, we adopt the same priors as in the monotone case in Section 3.1. Thus, the unknown parameters in the model (3.10) are $(\boldsymbol{\beta}, \boldsymbol{\theta}_J^\alpha, \sigma^2, \tau^2, \psi)$. Similar to the monotone case, we use mean field updates for $\boldsymbol{\beta}$, σ^2, and τ^2, and NVCMP updates with normal factors for ψ and $\boldsymbol{\theta}_J^\alpha$. That is, $q_1(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$ is parametrized as $N(\mu_\beta^q, \Sigma_\beta^q)$, the factor $q_3(\sigma^2)$ has the same form as for the monotone case, the factor $q_4(\tau^2)$ for τ^2 is an inverse gamma, $IG(\frac{r_{q,\tau}}{2}, \frac{s_{q,\tau}}{2})$, and the factor $q_5(\psi)$ for ψ is parametrized as $N(\mu_\psi^q, \sigma_\psi^{q\,2})$. The factor $q_2(\boldsymbol{\theta}_J^\alpha)$ for $\boldsymbol{\theta}_J^\alpha$ is parameterized as $N(\mu_{\alpha,\theta}^q, \Sigma_{\alpha,\theta}^q)$. Note that we use the same notations for all the expectations as those used in Section 3.1. In addition, note that by replacing Σ_θ^q with $\Sigma_{\alpha,\theta}^q$, μ_θ^q with $\mu_{\alpha,\theta}^q$, $\boldsymbol{\varphi}_J^a(x_i)$ with $\boldsymbol{\varphi}_J^{b,\alpha}(x_i)$, and Υ with $\Upsilon^\alpha = (\sigma_{0,\alpha}^2, \sigma_0^2, \tau^2\exp(-\gamma), \ldots, \tau^2\exp(-J\gamma))^\top$, the updates for $\boldsymbol{\beta}$, $\boldsymbol{\theta}_J^\alpha$, τ^2 and ψ follow the same form as the variational updates derived in Section 3.1. Thus, the remaining ones are about the updating procedure for σ^2 and the variational lower bound, whose details are as follows:

- For σ^2, the update is

$$\begin{aligned}
\log q(\sigma^2) \doteq\ & \mathrm{E}(\log p(\sigma^2) + \log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}_J^\alpha, \sigma^2) + \log p(\boldsymbol{\theta}_J^\alpha|\sigma^2, \tau^2, \psi)) \\
& + \log p(\boldsymbol{\beta}|\sigma^2),
\end{aligned}$$

  where

$$\begin{aligned}
\mathrm{E}(\log p(\boldsymbol{\theta}_J^\alpha|\sigma^2, \tau^2, \psi) &\doteq -\frac{(J+2)}{2}\log\sigma - \frac{1}{2\sigma}\mathrm{E}((\boldsymbol{\theta}_J^\alpha)^\top \mathrm{diag}((\Upsilon^\alpha)^{-1})\boldsymbol{\theta}_J^\alpha) \\
&\doteq -\frac{(J+2)}{2}\log\sigma \\
&\quad - \frac{1}{2\sigma}\mathrm{tr}\left\{\left(\Sigma_{\alpha,\theta}^q + \mu_{\alpha,\theta}^q \mu_{\alpha,\theta}^{q\,\top}\right)\mathrm{diag}(\mathrm{E}((\Upsilon^\alpha)^{-1}))\right\},
\end{aligned}$$

  and $\mathrm{E}(\log p(\sigma^2))$, $\mathrm{E}(\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}_J^\alpha, \sigma^2))$, and $\mathrm{E}(\log p(\boldsymbol{\beta}|\sigma^2))$ are the same as in the monotone case after replacing $\boldsymbol{\varphi}_J^a(x_i)$ with $\boldsymbol{\varphi}_J^{b,\alpha}(x_i)$, Υ with Υ^α,

Σ_θ^q with $\Sigma_{\alpha,\theta}^q$, and μ_θ^q with $\mu_{\alpha,\theta}^q$. Therefore, we have

$$\log q(\sigma^2) \doteq 2a \log \frac{1}{\sigma} + \frac{b}{\sigma} - \frac{c}{\sigma^2},$$

where

$$a = \frac{r_{0,\sigma} + n + p + (J+2)/2}{2} + 1,$$
$$b = -\frac{1}{2}\mathrm{tr}\left\{(\Sigma_{\alpha,\theta}^q + \mu_{\alpha,\theta}^q {\mu_{\alpha,\theta}^q}^\top)\mathrm{diag}(E((\Upsilon^\alpha)^{-1}))\right\}$$

and

$$\begin{aligned} c =& \frac{1}{2}\Bigg(s_{0,\sigma} + 2\mathrm{tr}({\Sigma_\beta^0}^{-1}\Sigma_\beta^q) + (\mu_\beta^q - \mu_\beta^0)^\top {\Sigma_\beta^0}^{-1}(\mu_\beta^q - \mu_\beta^0) \\ &+ \sum_{i=1}^n \Big[(y_i - \boldsymbol{w}_i^\top \mu_\beta^q - \delta\mathrm{tr}(\boldsymbol{\varphi}_J^{b,\alpha}(x_i)\Sigma_{\alpha,\theta}^q) - \delta {\mu_{\alpha,\theta}^q}^\top \boldsymbol{\varphi}_J^{b,\alpha}(x_i)\mu_{\alpha,\theta}^q)^2 \\ &+ 2\mathrm{tr}(\boldsymbol{\varphi}_J^{b,\alpha}(x_i)\Sigma_{\alpha,\theta}^q \boldsymbol{\varphi}_J^{b,\alpha}(x_i)\Sigma_{\alpha,\theta}^q) \\ &+ 4{\mu_{\alpha,\theta}^q}^\top \boldsymbol{\varphi}_J^{b,\alpha}(x_i)\Sigma_{\alpha,\theta}^q \boldsymbol{\varphi}_J^{b,\alpha}(x_i)\mu_{\alpha,\theta}^q + \boldsymbol{w}_i^\top \Sigma_\beta^q \boldsymbol{w}_i \Big] \Bigg). \end{aligned}$$

To derive the variational lower bound, the computations are the same as in the monotone case if we replace Σ_θ^q with $\Sigma_{\alpha,\theta}^q$, μ_θ^q with $\mu_{\alpha,\theta}^q$, $\boldsymbol{\varphi}_J^a(x_i)$ with $\boldsymbol{\varphi}_J^{b,\alpha}(x_i)$, and Υ with $\Upsilon^\alpha = (\sigma_{0,\alpha}^2, \sigma_0^2, \tau^2\exp(-\gamma), \ldots, \tau^2\exp(-J\gamma))^\top$, except for the terms $\mathrm{E}(\log p(\boldsymbol{\theta}_J^\alpha|\sigma^2,\tau^2,\psi))$ and $\mathrm{E}(\log q(\boldsymbol{\theta}_J^\alpha))$, which are given as

$$\begin{aligned} \mathrm{E}(\log p(\boldsymbol{\theta}_J^\alpha|\sigma^2,\tau^2,\psi)) =& -\frac{(J+2)}{2}\mathrm{E}(\log 2\pi\sigma) - \frac{1}{2}\log\sigma_0^2 - \frac{1}{2}\log\sigma_{0,\alpha}^2 \\ &- \frac{J}{2}\left\{\log(s_{q,\tau}/2) - \psi(r_{q,\tau}/2)\right\} \\ &+ \frac{J(J+1)}{4}\left\{\sigma_\psi^q\sqrt{\frac{2}{\pi}}\exp\left(-\frac{{\mu_\psi^q}^2}{2{\sigma_\psi^q}^2}\right)\right. \\ &\left. + \mu_\psi^q\left(1 - 2\Phi\left(-\frac{\mu_\psi^q}{{\sigma_\psi^q}^2}\right)\right)\right\} \\ &- \frac{1}{2}\mathrm{E}\left(\frac{1}{\sigma}\right)\mathrm{tr}\left\{\left(\Sigma_{\alpha,\theta}^q + \mu_{\alpha,\theta}^q {\mu_{\alpha,\theta}^q}^\top\right)\mathrm{diag}(\mathrm{E}((\Upsilon^\alpha)^{-1}))\right\}, \\ \mathrm{E}(\log q(\boldsymbol{\theta}_J^\alpha)) =& -\frac{J+2}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_{\alpha,\theta}^q| - \frac{J+2}{2}. \end{aligned}$$

Hence, it follows that the variational Bayes algorithm for BSAR with monotonic convex restriction is the same as `Algorithm 2` but with the replacements described above.

4. Empirical analysis

4.1. Curve fitting with VB approximations

In this section, we compare the performance of the VB approximations of the BSAR we proposed in Section 3 with some existing methods, based on simulation studies. Specifically, we first consider fitting univariate nonparametric regression models and compare the proposed VB approximation method of BSAR without restrictions, referred to as the VBU (Variational Bayes for Unrestricted model), with the three variational approximation approaches, VB-SSGP of Tan et al. (2016), VB-Spline of Zhao and Lian (2014) and VB-Pspline of Waldmann and Kneib (2015).

Our numerical implementation of all the VB approximations including VBU is written in `R`. In the implementation, the tolerance value *tol* in VBU is set to be 0.0001, and the hyperparameters for the priors in the VBU are set as $r_{0,\sigma} = 2(2 + m_{0,\sigma}^2/\nu_{0,\sigma})$, $r_{0,\tau} = 2(2 + m_{0,\tau}^2/\nu_{0,\tau})$, $s_{0,\sigma} = m_{0,\sigma}(\nu_{0,\sigma} - 2)$, $s_{0,\tau} = m_{0,\tau}(\nu_{0,\tau} - 2)$, $m_{0,\sigma} = 1$, $\nu_{0,\sigma} = 1000$, $m_{0,\tau} = 1$, $\nu_{0,\tau} = 100$, $\omega_0 = 2$, $\mu_\beta^0 = 0$, $\Sigma_\beta^0 = 100$, and $\sigma_0^2 = 100^2$. The speed of convergence of the variational approach is known to be sensitive to the starting values chosen for μ_ψ^q and μ_θ^q, and we choose $\mu_\psi^q = 1$ and $\mu_\theta^q = (1, 0, ..., 0)^T$ as our starting value for the VBU. For numerical implementations of other methods, `R` codes were obtained for VB-spline from the authors of Zhao and Lian (2014) and for VB-SSGP from the authors of Tan et al. (2016) by personal communication, and for the VB-Pspline method of Waldmann and Kneib (2015) the accompanying `R` package, `VA`, is used.

We simulate 50 datasets with two different sample sizes $n = 100$ and 200 based on the regression model $y = f(x) + \epsilon$ and use the root mean integrated squared error ($RMISE$) between the true function f and the posterior mean $\widehat{f}$ for performance evaluation. We consider $N = 50$ simulated datasets and by writing $\widehat{f}_j(\cdot)$ for the posterior mean obtained from dataset j, we define

$$RMISE_j(\widehat{f}_j, f) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\{f(x_i^{(j)}) - \widehat{f}_j(x_i^{(j)})\right\}^2}, j = 1, ..., N,$$

where $x_i^{(j)}$ is the ith value of covariate x in dataset j. To compare different methods, we consider the $RMISE_j(\widehat{f}_j, f)$ values averaged over the different datasets $j = 1, \cdots, N$. The values of x are equally spaced on 0 to 1, and the same values are used for each dataset with the same sample size.

In the first simulation study, we consider the following nonlinear regression models:

$$\begin{aligned}
(f_1)\quad y &= \sin(2(4x-2)) + 2\exp((-16^2)(x-0.5)^2) + \epsilon,\\
(f_2)\quad y &= 2 - 5x + \exp\{5(x-0.6)\} + \epsilon,\\
(f_3)\quad y &= x + \cos(4x) + \epsilon,\\
(f_4)\quad y &= 10\frac{\exp[15(x-0.4)]}{\exp[15(x-0.4)]+1} + \epsilon,
\end{aligned}$$

where $\epsilon \sim N(0,1)$. Figure 1 displays the simulated data and the true mean curve, respectively. The average RMISE values for the four methods, VBU, VB-SSGP, VB-spline and VB-Pspline are summarized in Table 1 with the standard errors (s.e.) and computing time (time) in seconds within parentheses.

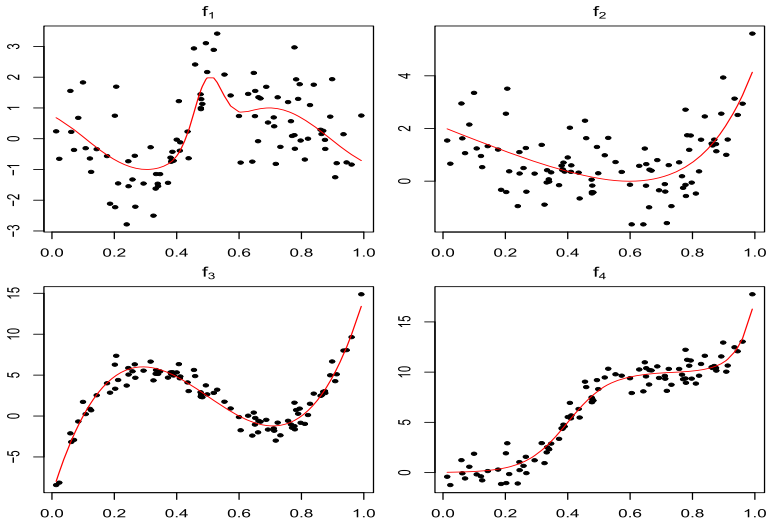

FIG 1. *Simulated data (black circle) and the true mean curves (red solid line) for* f_1–f_4

It appears that no method dominates and that the four methods have equivalent performance based on the standard errors. VBU has the best average RMISE in two functions f_1 and f_4 while other methods have the best average RMISE in f_2 and f_3. Overall, the simulation results indicate that the VBU is competitive with other variational methods in terms of RMISE as well as computing time. In terms of computational speed, VBU and VB-Pspline are the two best variational methods, which have the shortest computing times in all cases.

In the second simulation study, we consider the following monotone regression models:

$$\begin{aligned}
\texttt{Sigmoid} &: \quad Y = 5\exp(10x-5)/\left[1+\exp(10x-5)\right]+\epsilon,\\
\texttt{Sinusoid} &: \quad Y = 2\pi x+\sin(2\pi x)+\epsilon,\\
\texttt{Expo} &: \quad Y = \exp(6x-3)+\epsilon,\\
\texttt{LogX} &: \quad Y = \log(1+10x)+\epsilon,\\
\texttt{Const} &: \quad Y = \epsilon,
\end{aligned}$$

where $\epsilon \sim N(0,1)$. Figure 2 displays the simulated data and the true mean curve, respectively. Since there are no existing VB approximation methods for shape-restricted regression models, we compare the performance of the proposed VB approximation method of BSAR with shape restrictions, referred to as the VBM (variational Bayes for the monotone model) with the BSARM for monotone

TABLE 1
Average RMISE and computing time (seconds) for nonlinear functions over 50 repetitions

Function	n	VBU	VB-SSGP	VB-Spline	VB-Pspline
f_1	100	0.33	0.4	0.42	0.35
	(s.e./time)	(0.060/0.020)	(0.122/0.308)	(0.047/2.922)	(0.066/0.017)
	200	0.26	0.36	0.38	0.27
	(s.e./time)	(0.034/0.022)	(0.047/0.757)	(0.027/3.617)	(0.039/0.017)
f_2	100	0.30	0.27	0.22	0.23
	(s.e./time)	(0.06/0.019)	(0.06/0.294)	(0.06/3.142)	(0.07/0.020)
	200	0.2162	0.2073	0.1647	0.1690
	(s.e./time)	(0.0367/0.020)	(0.0361/0.728)	(0.0380/3.773)	(0.0403/0.020)
f_3	100	0.23	0.16	0.17	0.22
	(s.e./time)	(0.061/0.017)	(0.052/0.274)	(0.050/1.629)	(0.043/0.029)
	200	0.17	0.13	0.13	0.17
	(s.e./time)	(0.048/0.016)	(0.042/0.615)	(0.042/2.247)	(0.044/0.027)
f_4	100	0.24	0.30	0.29	0.25
	(s.e./time)	(0.059/0.014)	(0.076/1.116)	(0.050/1.418)	(0.056/0.009)
	200	0.18	0.22	0.24	0.18
	(s.e./time)	(0.048/0.013)	(0.055/1.961)	(0.037/1.751)	(0.050/0.012)

regression model of Lenk and Choi (2017) and Bayesian regression splines with monotone restrictions, BRSM of Meyer, Hackstadt and Hoeting (2011), with the latter two methods implemented using MCMC. Our numerical implementation of the VBM approach is written in `R`, and the entire setup, including initial values and the tolerance value, are the same as in VBU. For numerical implementations of other methods, an `R` package, `bsamGP`(Jo et al., 2017) is used for BSARM, and BRSM is implemented by the `R` code available from the author's website as given in Meyer, Hackstadt and Hoeting (2011).

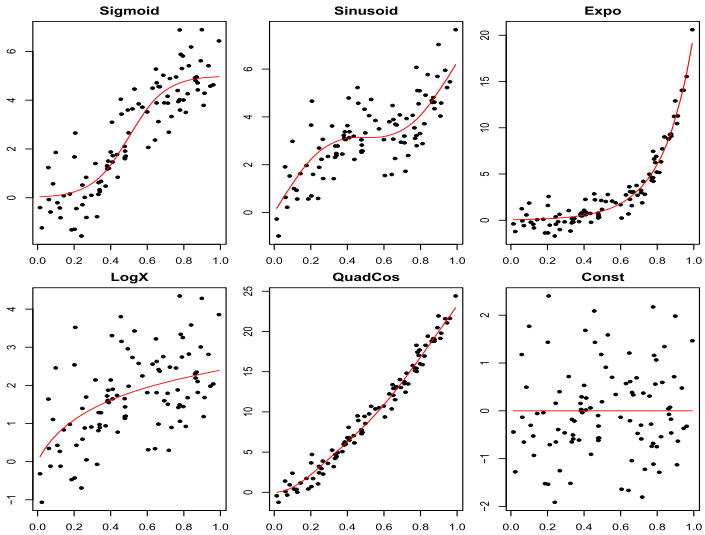

FIG 2. *Simulated data and the true mean curve for monotone and/or convex regression models*

Table 2 summarizes the RMISE's, the standard errors and computing time in seconds. Overall, the simulation results indicate that VBM is slightly worse than the two other MCMC methods for the shape-restricted regression models in terms of RMISE. The worse performance of VBM compared to the MCMC methods is, we believe, due to two factors. The first is the posterior independence assumptions that are inherent to the VB approximation, and which can cause both underestimation of variability as well as bias in point estimates of variance parameters and smoothing parameters. These drawbacks are not confined to this application only but apply to VB approaches more generally (see, e.g., Wang and Titterington (2004) and Turner and Sahani (2011)). The second, less important factor that may explain the poorer performance of the VBM approach is the adaptive truncation of the number of basis functions used to avoid numerical singularities as described in the remark following Algorithm 2. As expected, in all cases, VBM has an advantage over BSARM and BRSM in terms of computational speed.

TABLE 2
Average RMISE and computing time (seconds) for monotone functions over 50 repetitions

Function	n	VBM	BSARM	BRSM
`Sigmoid`	100	0.3	0.21	0.21
	(s.e./time)	(0.120/1.569)	(0.070/21.75)	(0.086/49.54)
	200	0.23	0.15	0.13
	(s.e./time)	(0.058/3.853)	(0.043/34.95)	(0.046/53.43)
	500	0.20	0.11	0.098
	(s.e./time)	(0.025/8.605)	(0.030/56.85)	(0.026/64.50)
`Sinusoid`	100	0.21	0.23	0.25
	(s.e./time)	(0.097/1.692)	(0.068/27.61)	(0.067/49.69)
	200	0.20	0.16	0.19
	(s.e./time)	(0.067/4.214)	(0.041/35.65)	(0.048/53.53)
	500	0.18	0.10	0.11
	(s.e./time)	(0.025/9.117)	(0.029/55.91)	(0.037/64.53)
`Expo`	100	0.45	0.26	0.25
	(s.e./time)	(0.057/2.002)	(0.074/27.34)	(0.074/50.01)
	200	0.3	0.19	0.17
	(s.e./time)	(0.116/4.789)	(0.042/35.39)	(0.051/53.71)
	500	0.37	0.13	0.11
	(s.e./time)	(0.021/8.907)	(0.028/57.85)	(0.027/64.53)
`LogX`	100	0.17	0.16	0.22
	(s.e./time)	(0.049/1.754)	(0.055/28.06)	(0.048/49.84)
	200	0.14	0.13	0.15
	(s.e./time)	(0.0350/4.201)	(0.040/25.94)	(0.043/53.49)
	500	0.11	0.087	0.11
	(s.e./time)	(0.025/9.576)	(0.029/55.86)	(0.030/64.42)
Const	100	0.14	0.14	0.086
	(s.e./time)	(0.039/1.634)	(0.051/28.20)	(0.057/49.78)
	200	0.12	0.094	0.060
	(s.e./time)	(0.065/3.667)	(0.036/26.52)	(0.038/53.42)
	500	0.15	0.066	0.036
	(s.e./time)	(0.131/7.806)	(0.028/58.49)	(0.029/64.29)

Next, consider the following three regression models with monotonicity and convexity,

$$\texttt{Expo} \ : \ Y = \exp(6x - 3) + \epsilon \tag{4.1}$$

$$\texttt{QuadCos} \ : \ Y = 16x^2 - \frac{4}{\pi^2}\cos(2\pi x) - \frac{1}{\pi^2}\cos(4\pi x) - \frac{32}{9\pi^2}\cos(3\pi x) - \frac{32}{\pi^2}\cos(\pi x) + \frac{365}{9\pi^2} + \epsilon \tag{4.2}$$

$$\texttt{LogX} \ : \ Y = \log(1 + 10x) + \epsilon \tag{4.3}$$

where $\epsilon \sim N(0, 1)$. Both the `Expo` and `QuadCos` models are increasing and convex on $[0, 1]$ while the `LogX` model is increasing and concave on $[0, 1]$, as also shown in Figure 2.

Similarly to the previous simulation study, results from VBMC (VB for the monotone and convex model) and its MCMC counterpart BSARMC (BSAR for the monotone and convex model) from the `R` package, `bsamGP` are compared. For the VBMC procedure, we assign starting values of $\mu^q_\psi = 0.5$ and $\mu^q_{\alpha,\theta} = (0.5, 1, 0, ..., 0)^T$ for $n = 50$ and $\mu^q_\psi = 1$ and $\mu^q_{\alpha,\theta} = (1, 1, 0, ..., 0)^T$ for $n = 100$ and 200. Table 3 presents the average RMISE of VBMC and BSARMC with the standard error of each average RMISE as before.

As summarized in Table 3, the average RMISE of VBMC is slightly larger than for the MCMC method BSARMC for the monotonic and convex/concave functions. We believe the worse performance of VBMC compared to BSARMC is for reasons similar to those discussed earlier for the case of monotone constraints. We do note, however, that the performance gap compared to MCMC seems to be reduced for the case of concave/convex constraints compared to monotone constraints. We believe this occurs because with more stringent shape constraints the fit becomes less sensitive to smoothing parameters and to any bias in estimation of them. A comparison of computation times for the algorithms is given in Table 4. For both monotone and convex/concave shape constraints, the VB algorithms are an order of magnitude faster, which justifies some loss of statistical performance in cases where computation time is an important consideration.

TABLE 3
Average RMISE (s.e) for monotonic and convex/concave functions over 50 repetitions

Function	n	VBMC	BSARMC
Expo	50	0.339 (0.006)	0.31 (0.011)
	100	0.255 (0.006)	0.219 (0.007)
	200	0.210 (0.003)	0.163 (0.003)
QuadCos	50	0.27 (0.013)	0.25 (0.011)
	100	0.213 (0.006)	0.198 (0.005)
	200	0.167 (0.003)	0.160 (0.002)
LogX	50	0.22 (0.014)	0.20 (0.011)
	100	0.151 (0.006)	0.142 (0.005)
	200	0.112 (0.003)	0.114 (0.003)

Figure 3 shows the boxplots of the ratios of the individual RMISE values between the BSARMC and VBMC for different sample sizes $n = 50, 100$, and 200. Ratios less than one indicate that BSARMC has better performance than VBMC does for point estimation in terms of RMISE. It seems that the RMISE's of the VBMC and BSARMC approaches are similar for each of the three models.

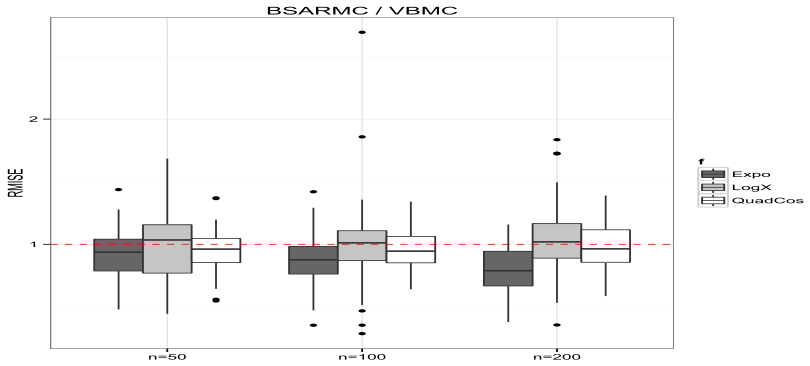

FIG 3. *Boxplots of ratios of two estimated RMISE's:* $\text{RMISE}_j(\text{BSARMC})/\text{RMISE}_j(\text{VBMC})$, $j = 1, \ldots, N = 50$. *The ratio less than 1 implies that BSARMC outperforms VBMC in terms of* RMISE.

TABLE 4
Average computation time (in seconds) over 5 repetitions.

	$n = 100$, $J = 40$		$n = 200$, $J = 50$		$n = 500$, $J = 100$	
Function	VBMC	BSARMC	VBMC	BSARMC	VBMC	BSARMC
Expo	1.07	18.64	3.91	58.80	9.13	603.7
QuadCos	1.43	18.92	3.76	59.89	8.96	606.0
LogX	1.19	18.63	3.65	60.93	10.3	606.4

Table 4 presents the average computation time of VBMC and BSARMC for the above examples. As expected, the variational Bayes approach has a much lower average computation time compared to the MCMC approach. The differences become increasingly significant as the sample size increases. For example, when $n = 500$ and $J = 100$, the amount of time required for the variational approach to converge is a small fraction of the time (less than 2%) required to run an MCMC analysis.

4.2. Credible interval estimation and model selection with application to electricity demand data

In this example, we use the electricity demand data in Yatchew (2003) to compare between the VB and MCMC algorithm based on BSAR. The data contains 288 quarterly observations of Ontario's electricity demand from 1971 to 1994. Following Yatchew (2003), Lenk and Choi (2017) use the log of the electricity demand to GDP as the dependent variable and log price ratio of electricity to natural gas as a covariate in $\boldsymbol{W}$. The choice of dependent variable is intentional as Yatchew (2003) found that the demand for electricity is co-integrated with the gross domestic product. Similar to Lenk and Choi (2017), we use "Temperature", which is the number of heating and cooling degree days relative to 68°F, as the independent variable x. We consider all three models, namely unrestricted, monotone, and monotonic convex in our application.

The hyperparameters for the priors in the variational Bayes approach are set up as follows. We use $\mu_\beta^0 = (0,0)^T$ and $\Sigma_\beta^0 = 100\mathbf{I}_2$ as hyperparameters for the prior of $\boldsymbol{\beta}$. For all other hyperparameters for the priors, we set them to be exactly the same as discussed in the previous sections. Similar to the simulation studies, our choice of starting points is determined by trial and error. For the unrestricted fit, we use $\mu_\psi^q = 1$. For the non-increasing shape-restricted fit, we use $\mu_\psi^q = 5$ and $\mu_\theta^q = (5,5,...,5)^T$ as the starting point. Finally, for the non-increasing convex fit, we use $\mu_\psi^q = 1$ and $\mu_\theta^q = (0.5,0.5,...,0.5)^T$. Further, our choice of the initial truncation point J is set to 60.

In addition to curve fitting for point estimation, we consider credible interval estimation. One advantage of the variational procedure is that we are able to simulate independent samples directly from the variational posterior distribution, which facilitates computations. For example, if we want to estimate a credible interval for $\delta\boldsymbol{\theta}_J^T\boldsymbol{\varphi}_J^a(x)\boldsymbol{\theta}_J$ in the monotone case, we first simulate a sufficiently large number of $\boldsymbol{\theta}_J$ from $N_q(\mu_\theta^q, \Sigma_\theta^q)$ and then for each of these points we plug $\boldsymbol{\theta}_J$ into the function $\delta\boldsymbol{\theta}_J^T\boldsymbol{\varphi}_J^a(x)\boldsymbol{\theta}_J$. A credible interval of $\delta\boldsymbol{\theta}_J^T\boldsymbol{\varphi}_J^a(x)\boldsymbol{\theta}_J$ can then be obtained from the corresponding sample quantiles of these plug-in values. The procedure is similar to the one followed for constructing credible intervals from the MCMC output, except that in the case of MCMC, the approximate posterior samples are dependent.

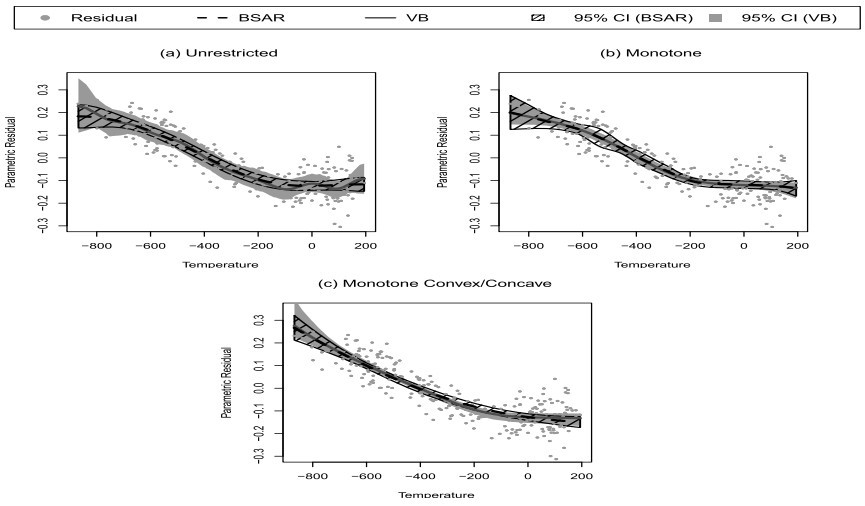

FIG 4. *Estimated fit and credible interval for electricity demand. The dots are the residual, while the solid and dashed lines are the posterior means for the variational procedure and BSAR, respectively.*

Figure 4 shows the estimated posterior mean of f for all three models against temperature and the 95% credible intervals using both VB and MCMC based on BSAR. In particular, Figure 4 (a), (b) and (c) shows the fit of the unrestricted, non-increasing and non-increasing convex models respectively. We observe that

in all three models, the variational Bayes fit follows quite closely the MCMC fit. We also observe that other than the unrestricted case, it seems that the width of the 95% credible interval is similar in both VB and MCMC for both the monotone and monotonic convex case. This result is not always what is expected when using a variational Bayes approximation, since such approximations are known to underestimate variability in some situations. Our findings show that the estimated posterior mean of τ by the VB procedures is much larger than that of the BSAR but this does not seem to result in any corresponding inaccuracy in estimation of mean functions or credible intervals. Computation times for the unrestricted, decreasing, and decreasing convex case were (in seconds) 1.65, 26.21 and 25.94 respectively for MCMC, and 0.02, 12.39, and 4.08 seconds respectively, for VB.

Furthermore, we test the adequacy of the parametric against the semiparametric model for fitting the electricity demand data, by computing the marginal likelihoods of competing models. In particular, we compare a parametric model without "Temperature" (H_0) to a semiparametric model with "Temperature" (H_1) in our application,

$$H_0 \ : \ y = \boldsymbol{w}^T\boldsymbol{\beta} + \epsilon \ \ versus \ \ H_1 \ : \ y = \boldsymbol{w}^T\boldsymbol{\beta} + f(x) + \epsilon,$$

where x denotes "Temperature" as mentioned before. As summarized in Table 5, the semiparametric models H_1 with "Temperature" have larger marginal likelihoods, $\log p(\boldsymbol{y})$, than for the parametric model H_0 and they also have better in sample fits with smaller root mean squared error (RMSE) between the observed Y and estimated regression function than for the parametric model, based on VB as well as MCMC procedures. Here, the marginal likelihood is computed using the Gelfand and Dey approximation (Gelfand and Dey, 1994) for MCMC methods in exactly the same way as described in Lenk (1999) and Lenk and Choi (2017) for BSAR and BSARM. Specifically, let $\boldsymbol{\vartheta}_j$ be a set of unknown parameters involved in BSAR and BSARM for model H_j; let $p_j(\boldsymbol{\vartheta}_j)$ be a prior density of $\boldsymbol{\vartheta}_j$, $p_j(\boldsymbol{y}|\boldsymbol{\vartheta}_j)$ be the likelihood function of $\boldsymbol{y}$ given $\boldsymbol{\vartheta}_j$ under H_j, and $h_j(\boldsymbol{\vartheta}_j)$ be an auxiliary distribution on the support of $\boldsymbol{\vartheta}_j$. Then the Gelfand and Dey approximation $p_j(\boldsymbol{y})$ used for the marginal likelihood under H_j is given by

$$p_j(\boldsymbol{y})^{-1} = \frac{1}{B}\sum_{u=1}^{B} \frac{h_j(\boldsymbol{\vartheta}_j^{(u)})}{p_j(\boldsymbol{y}|\boldsymbol{\vartheta}_j^{(u)})p_j(\boldsymbol{\vartheta}_j^{(u)})}$$

where $\boldsymbol{\vartheta}_j^{(u)}$ is the uth value of $\boldsymbol{\vartheta}_j$ generated from the MCMC algorithm, and B denotes the total number of poster samples after burn-in period. As the auxiliary distribution h_j, we take the same distributions as priors for $\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \tau^2$ and θ_0, while we use the truncated normal distribution for γ. In comparison with BSAR and BSARM, we evaluate the lower bound $\mathcal{L}(q)$ in the VB approximation for marginal likelihood computation.

Further, as shown in Table 5, the marginal likelihoods based on VBM and BSARM are larger than those from VBU and BSAR, which indicates that in the semiparametric models H_1, the shape-restricted model with monotonicity

is favored over the unrestricted model in terms of the marginal likelihoods for both VB and MCMC procedures and that the VB lower bounds could be used for model selection purposes in addition to point estimation and credible interval construction. However, the variational lower bound can have errors of very different magnitudes for the shape restricted and unrestricted cases, which suggests that it should be used with caution in model choice in this setting.

Alternatively, we consider two information criteria in the context of the variational Bayes approach, namely VAIC (Variational AIC) and VBIC (Variational BIC),

$$\begin{aligned} \text{VAIC} &= 2\log p(\boldsymbol{y} \mid \mathrm{E}_q(\boldsymbol{\delta})) - 4\mathrm{E}_q \log p(\boldsymbol{y} \mid \boldsymbol{\delta}) \\ \text{VBIC} &= -2\mathcal{L}(q) + 2\mathrm{E}_q \log p(\boldsymbol{\delta}), \end{aligned}$$

proposed by You, Ormerod and Müller (2014), in particular for the Bayesian linear model and certain diffuse priors. We also speculate that these VAIC and VBIC approaches would be applicable to our problems and that they would ameliorate such a limitation with normalizing constants and diffuse priors we employed, in addition to aforementioned concerns in the variational lower bound for model section. In computing VAIC and VBIC, we need to additionally evaluate $\mathrm{E}_q(\log \sigma)$, and details about this are given in the Appendix. The results summarized in Table 5 also indicate that VBM is still favored over VBU in terms of VAIC, the same as the VB lower bounds, whereas VBU is favored with VBIC as in RMSE. Note that VBIC still relies on the lower bound directly and hence has the same problem as the lower bound for model choice purposes. However, it seems VAIC does not depend directly on the lower bound and hence may be more reliable. Although the two VB information criteria do not agree, it is evident that semiparametric models in H_1 provide adequate descriptions of the electricity demand data, compared to the parametric model H_0 as also shown in RMSE and $\mathcal{L}(q)$ values.

TABLE 5
Summary results of model selection for electricity demand data

Model	H_0	H_1		H_0	H_1	
Approach	Linear VB	VBU	VBM	Linear MCMC	BSAR	BSARM
RMSE	0.120	0.052	0.054	0.120	0.053	0.053
$\mathcal{L}(q)$ ($\log p(\boldsymbol{y})$)	141.6	143.9	182.7	142.0	155.0	234.5
VAIC	-195.6	-781.8	-2419	-	-	-
VBIC	-361.8	-554.9	-467.2	-	-	-

4.3. A large data set with stock price

The last empirical analysis is for an illustration of the merit of the VB approach for dealing with a large data set, specifically, a stock price data set from the London Stock Exchange in the United Kingdom. A similar data set was also analyzed in Luts, Broderick and Wand (2014).

TABLE 6
Summary of Stock Price data

Source	Google Finance (www.google.com/finance)
Model	$\text{HSBC}_i = f(\text{BARC}_i) + \epsilon_i$
# of obs	2906
Date	Jan/01/2005 ∼ Jan/21/2016

The data set is based on London Stock Exchange data during its opening hours, collected through the `R` package `quantmod` setting the time interval from January 1st of 2005 to July 21st of 2016. The source of the data is Google Finance, https://www.google.com/finance. The predictor (x) and response variable (y) consist of the stock prices of two financial institutions: *The Barclays PLC* and *The Hongkong and Shanghai Banking Corporation(HSBC)*, as summarized in Table 6. Although this data set is moderately large, it is also chosen to be small enough that MCMC implementations of shape restricted regression are still feasible for comparison.

We consider six different approaches, VB and MCMC for three models, unrestricted, monotone, and monotonic concave, to analyze the data set. Figure 5 presents the estimated fits for VBU and VBM with 95 % credible intervals for stock price data, and Table 7 summarizes additional information about the fits, including RMSE and computing time in seconds. As summarized in Table 7, the unrestricted model (VBU/BSAR) has the largest marginal likelihood and the smallest RMSE among the three models, and in terms of RMSE, VB approaches provide competitive fits compared to MCMC. The VB approach has computational demands less than for the MCMC approaches by several orders of magnitude.

TABLE 7
Summary results of analysis of Stock Price data

Approach	VBU	VBM	VBMC	BSAR	BSARM	BSARMC
RMSE	57.69	58.14	59.19	57.11	58.43	59.31
Time	0.45	42.22	495.9	42.42	2265	2297
$\mathcal{L}(q)$ $(\log p(\boldsymbol{y}))$	-16003	-17473	-17469	-16028	-16069	-16216

5. Conclusion

In this paper, we presented a variational Bayes approach to a semiparametric regression model based on a spectral analysis of Gaussian process priors. In particular, we developed fast variational Bayes methods for semiparametric regression models with monotone and convex/concave restrictions for the regression function by modeling its derivatives with squared Gaussian processes. The variational approximation schemes we developed were shown to fit the semiparametric regression models based on the framework of Lenk and Choi (2017) comparable to MCMC methods and to reduce computation time relative to MCMC methods. In addition, the variational Bayes methods could provide reasonable

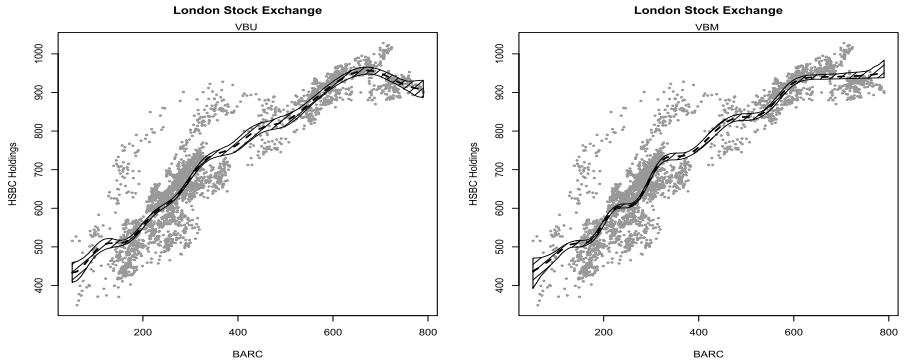

FIG 5. *Estimated fits with VBU and VBM for stock price data. The dots are observations while different lines are the posterior means for the VBU (left panel) and VBM (right panel) with 95 % credible intervals, respectively.*

credible intervals and marginal likelihoods useful for uncertainty quantification and model selection based on real data applications.

There are several issues that could be considered in future work. In our experiments with the variational algorithm, we found that the convergence rate of the variational approach can be quite sensitive to the starting point. In particular the variational algorithm may become stuck in local modes in certain models or exhibit slow convergence. In the current work, suitable starting points were determined by trial and error, and more systematic methods for this are needed. Further, the use of variational methods to obtain better MCMC proposals could be explored. There are also other shape restrictions considered in Lenk and Choi (2017), such as U-Shaped and S-Shaped restrictions, and it would be interesting to attempt to implement a variational Bayes approach in these models. Variational approaches in these and other semiparametric models, for example, functional regression, quantile regression and spatial data analysis, and non-Gaussian data (see, e.g., Goldsmith, Wand and Crainiceanu (2011), Luts and Wand (2015) and Waldmann and Kneib (2015)), may be particularly challenging and important in dealing with high-dimensional problems in the context of Gaussian process priors and shape restrictions. Alternatively, stochastic gradient approaches to variational inference (Ji, Shen and West, 2010; Nott et al., 2012; Paisley, Blei and Jordan, 2012) could be considered in these settings. Further, we plan to adapt the proposed VB methods for shape restrictions into the mean field VB of Neville, Ormerod and Wand (2014) for sparse signal shrinkage and the linear response VB of Giordano, Broderick and Jordan (2015) for overcoming the limitations of mean field variational Bayes in underestimating the variability, incurring bias and posterior dependence (see, e.g., Wang and Titterington (2004), Turner and Sahani (2011) and Neville, Ormerod and Wand (2014)).

Moreover, the proposed VB approach to Fourier series with shape restrictions could be extended to multivariate predictors. Most simply, a multivariate

nonparametric component with an assumed additive structure could be used, with shape constraints on the additive terms. However, the additive assumption is limiting and the more general problem of handling multivariate shape constraints is complex, with a much smaller existing literature than for univariate shape constraints. Expanding the BSAR methods to handle multivariate shape constraints is not easy; the number of basis terms needed grows exponentially with respect to the dimensions. Existing methods for handling multivariate shape constraints include methods using Gaussian processes (Riihimäki and Vehtari, 2010; Lin and Dunson, 2014), as well as methods using multivariate basis functions with shape restrictions such as multivariate splines (e.g., Cai and Dunson (2007)), tensor product bases (e.g. Hofner, Kneib and Hothorn (2016)) and radial basis functions (e.g. Chakraborty, Ghosh and Mallick (2012) and Zhang et al. (2014)). It is fair to say, however, that most of these methods either do not scale well with the dimension or with the sample size. An exception is the recent work of Riihimäki and Vehtari (2010) for monotone Gaussian Process regression and classification using virtual derivative observations. That approach is able to handle genuinely multivariate shape constraints, and they implement their methods using a scalable approximate inference algorithm, expectation propagation.

Appendix

Conjugate variational updates and lower bound for model without shape restriction

We provide details of the updates for q_1–q_4 as follows:

- For $\boldsymbol{\beta}$, the mean field update $q_1(\boldsymbol{\beta})$ takes the form

$$\log q_1(\boldsymbol{\beta}) \doteq \mathrm{E}_{-1}(\log p(\boldsymbol{\beta}|\sigma^2)) + \mathrm{E}_{-1}(\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}_J, \sigma^2)),$$

where

$$\begin{aligned}
\mathrm{E}_{-1}(\log p(\boldsymbol{\beta}|\sigma^2)) &\doteq -\frac{1}{2}\frac{r_{q,\sigma}}{s_{q,\sigma}}(\boldsymbol{\beta}-\mu_\beta^0)^\top \Sigma_\beta^{0\,-1}(\boldsymbol{\beta}-\mu_\beta^0),\\
\mathrm{E}_{-1}(\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}_J, \sigma^2)) &\doteq -\frac{1}{2}\mathrm{E}_{-1}\left(\frac{1}{\sigma^2}\right)\mathrm{E}((\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\beta}-\boldsymbol{\varphi}_J\boldsymbol{\theta}_J)^\top\\
&\quad \times(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\beta}-\boldsymbol{\varphi}_J\boldsymbol{\theta}_J))\\
&\doteq -\frac{1}{2}\frac{r_{q,\sigma}}{s_{q,\sigma}}\left\{\mathrm{tr}(\boldsymbol{\varphi}_J^\top\boldsymbol{\varphi}_J\Sigma_\theta^q)\right.\\
&\quad \left.+(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\beta}-\boldsymbol{\varphi}_J\mu\theta^q)^\top(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\beta}-\boldsymbol{\varphi}_J\mu_\theta^q))\right\}\\
&\doteq -\frac{1}{2}\frac{r_{q,\sigma}}{s_{q,\sigma}}\left\{\boldsymbol{\beta}^\top\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{\beta}\right.\\
&\quad \left.-2\boldsymbol{\beta}^\top\boldsymbol{W}^\top(\boldsymbol{y}-\boldsymbol{\varphi}_J\mu_\theta^q)\right\}.
\end{aligned}$$

Thus, we have

$$\log q_1(\boldsymbol{\beta}) \doteq -\frac{1}{2}\frac{r_{q,\sigma}}{s_{q,\sigma}}\left\{\boldsymbol{\beta}^\top\left(\Sigma_\beta^{0\,-1}+\frac{r_{q,\sigma}}{s_{q,\sigma}}\boldsymbol{W}^\top\boldsymbol{W}\right)\boldsymbol{\beta} -2\boldsymbol{\beta}^\top\left(\Sigma_\beta^{0\,-1}\mu_\beta^0+\frac{r_{q,\sigma}}{s_{q,\sigma}}\boldsymbol{W}^\top(\boldsymbol{y}-\boldsymbol{\varphi}_J\mu_\theta^q)\right)\right\},$$

from which we deduce that $q_1(\boldsymbol{\beta})$ is multivariate normal, $N(\mu_\beta^q, \Sigma_\beta^q)$ with the expressions for Σ_β^q and μ_β^q given in `Algorithm 1`.

- For $\boldsymbol{\theta}_J$, the mean field update $q_2(\boldsymbol{\theta}_J)$ has the form

$$\log q_2(\boldsymbol{\theta}_J) \doteq \mathrm{E}_{-2}(\log p(\boldsymbol{\theta}_J|\sigma^2,\tau^2,\psi)) + \mathrm{E}_{-2}(\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta}_J,\sigma^2)),$$

where

$$\begin{aligned}
\mathrm{E}_{-2}(\log p(\boldsymbol{\theta}_J|\sigma^2,\tau^2,\psi)) &\doteq -\frac{1}{2}\mathrm{E}_{-2}\left(\frac{1}{\sigma^2}\right)\mathrm{E}_{-2}\left(\frac{1}{\tau^2}\right)\boldsymbol{\theta}_J^\top\mathrm{diag}(\mathrm{E}_{-2}(\Gamma^{-1}))\boldsymbol{\theta}_J\\
&\doteq -\frac{1}{2}\frac{r_{q,\sigma}}{s_{q,\sigma}}\frac{r_{q,\tau}}{s_{q,\tau}}\boldsymbol{\theta}_J^\top\mathrm{diag}(\mathrm{E}_{-2}(\Gamma^{-1}))\boldsymbol{\theta}_J,\\
\mathrm{E}_{-2}(\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta}_J,\sigma^2)) &\doteq -\frac{1}{2}\mathrm{E}_{-2}\left(\frac{1}{\sigma^2}\right)\mathrm{E}_{-2}((\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\beta}-\varphi\theta)^\top\\
&\quad\times(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{\beta}-\boldsymbol{\varphi}_J\boldsymbol{\theta}_J))\\
&\doteq -\frac{1}{2}\frac{r_{q,\sigma}}{s_{q,\sigma}}\left\{\mathrm{tr}\boldsymbol{W}^\top\boldsymbol{W}\Sigma_\beta^q)\right.\\
&\quad\left.+(\boldsymbol{y}-\boldsymbol{W}\mu_\beta^q-\boldsymbol{\varphi}_J\boldsymbol{\theta}_J)^\top(\boldsymbol{y}-\boldsymbol{W}\mu_\beta^q-\boldsymbol{\varphi}_J\boldsymbol{\theta}_J)\right\}\\
&\doteq -\frac{1}{2}\frac{r_{q,\sigma}}{s_{q,\sigma}}\left\{\boldsymbol{\theta}_J^\top\boldsymbol{\varphi}_J^\top\boldsymbol{\varphi}_J\boldsymbol{\theta}_J-2\boldsymbol{\theta}_J^\top\boldsymbol{\varphi}_J^\top(\boldsymbol{y}-\boldsymbol{W}\mu_\beta^q)\right\}.
\end{aligned}$$

Here, $\mathrm{E}_{-2}\left(\Gamma^{-1}\right)$ indicates a J-dimensional vector with elements $Q_j(\mu_\psi^q, \sigma_\psi^{q\,2})$, $j=1,\ldots,J$, where

$$\begin{aligned}
Q_j\left(\mu_\psi^q,\sigma_\psi^{q\,2}\right) &= \mathrm{E}_{-2}\left(\exp(j|\psi|)\right)\\
&= \exp\left(\frac{\sigma_\psi^{q\,2}j^2}{2}+\mu_\psi^q j\right)\left\{1-\Phi\left(-\frac{\mu_\psi^q}{\sigma_\psi^q}-\sigma_\psi^q j\right)\right\}\\
&\quad+\exp\left(\frac{\sigma_\psi^{q\,2}j^2}{2}-\mu_\psi^q j\right)\left\{1-\Phi\left(\frac{\mu_\psi^q}{\sigma_\psi^q}-\sigma_\psi^q j\right)\right\},\\
&\quad j=1,\ldots,J,\\
\Gamma^{-1} &= (\exp(|\psi|),\exp(2|\psi|),\ldots,\exp(J|\psi|)^\top. \qquad (5.1)
\end{aligned}$$

Thus, we have

$$\log q_2(\boldsymbol{\theta}_J) \doteq -\frac{1}{2}\left\{\boldsymbol{\theta}_J^\top\left(\frac{r_{q,\sigma}}{s_{q,\sigma}}\boldsymbol{\varphi}_J^\top\boldsymbol{\varphi}_J+\frac{r_{q,\sigma}}{s_{q,\sigma}}\frac{r_{q,\tau}}{s_{q,\tau}}\mathrm{diag}(\mathrm{E}_{-2}(\Gamma^{-1}))\right)\boldsymbol{\theta}_J -2\frac{r_{q,\sigma}}{s_{q,\sigma}}\boldsymbol{\theta}_J^\top\boldsymbol{\varphi}_J^\top(\boldsymbol{y}-\boldsymbol{W}\mu_\beta^q)\right\},$$

from which we deduce that $q_2(\boldsymbol{\theta}_J)$ is a multivariate normal distribution, $N(\mu_\theta^q, \Sigma_\theta^q)$ with the expressions for Σ_θ^q and μ_θ^q given in `Algorithm 1`.

- For σ^2, the mean field update $q_3(\sigma^2)$ has the form

$$\begin{aligned}\log q_2(\sigma^2) \doteq \mathrm{E}_{-3}\big[&\log p(\sigma^2) + \log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}_J, \sigma^2) \\ &+ \log p(\boldsymbol{\theta}_J|\sigma^2, \tau^2, \psi) + \log p(\boldsymbol{\beta}|\sigma^2)\big],\end{aligned}$$

where

$$\begin{aligned}\mathrm{E}_{-3}(\log p(\sigma^2)) &\doteq -\left(\frac{r_{0,\sigma}}{2} + 1\right)\log\sigma^2 - \frac{s_{0,\sigma}}{2\sigma^2}, \\ \mathrm{E}_{-3}(\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}_J, \sigma^2)) &\doteq -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\mathrm{E}((\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\beta} - \boldsymbol{\varphi}_J\boldsymbol{\theta}_J)^\top \\ &\quad \times (\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\beta} - \boldsymbol{\varphi}_J\boldsymbol{\theta}_J)) \\ &\doteq -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\Big\{\mathrm{tr}(\boldsymbol{W}^\top\boldsymbol{W}\Sigma_\beta^q) + \mathrm{tr}(\boldsymbol{\varphi}_J^\top\boldsymbol{\varphi}_J\Sigma_\theta^q) \\ &\quad + (\boldsymbol{y} - \boldsymbol{W}\mu_\beta^q - \boldsymbol{\varphi}_J\mu_\theta^q)^\top(\boldsymbol{y} - \boldsymbol{W}\mu_\beta^q - \boldsymbol{\varphi}_J\mu_\theta^q)\Big\},\end{aligned}$$

$$\begin{aligned}\mathrm{E}_{-3}(\log p(\boldsymbol{\theta}_J|\sigma^2, \tau^2, \psi)) &\doteq -\frac{J}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\mathrm{E}_{-3}\left(\frac{1}{\tau^2}\right)\mathrm{E}(\boldsymbol{\theta}_J^\top\mathrm{diag}(\Gamma^{-1})\boldsymbol{\theta}_J) \\ &\doteq -\frac{J}{2}\log\sigma^2 \\ &\quad - \frac{1}{2\sigma^2}\frac{r_{q,\tau}}{s_{q,\tau}}\mathrm{tr}((\Sigma_\theta^q + \mu_\theta^q\mu_\theta^{q\top})\mathrm{diag}(\mathrm{E}_{-3}(\Gamma^{-1}))), \\ \mathrm{E}_{-3}(\log p(\boldsymbol{\beta}|\sigma^2)) &\doteq -\frac{p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\Big\{(\mu_\beta^q - \mu_\beta^0)^\top\Sigma_\beta^{0\,-1}(\mu_\beta^q - \mu_\beta^0) \\ &\quad + \mathrm{tr}(\Sigma_\beta^{0\,-1}\Sigma_\beta^q)\Big\}.\end{aligned}$$

Thus, we have

$$\begin{aligned}\log q_3(\sigma^2) \doteq -&\left(\frac{r_{0,\sigma} + n + p + J}{2} + 1\right)\log\sigma^2 \\ &- \frac{1}{2\sigma^2}\Big\{s_{0,\sigma} + \mathrm{tr}(\boldsymbol{W}^\top\boldsymbol{W}\Sigma_\beta^q) + \mathrm{tr}(\boldsymbol{\varphi}_J^\top\boldsymbol{\varphi}_J\Sigma_\theta^q) \\ &+ \mathrm{tr}(\Sigma_\beta^{0\,-1}\Sigma_\beta^q) + \frac{r_{q,\tau}}{s_{q,\tau}}\mathrm{tr}((\Sigma_\theta^q + \mu_\theta^q\mu_\theta^{q\top})\mathrm{diag}(\mathrm{E}_{-3}(\Gamma^{-1}))) \\ &+ (\boldsymbol{y} - \boldsymbol{W}\mu_\beta^q - \boldsymbol{\varphi}_J\mu_\theta^q)^\top(\boldsymbol{y} - \boldsymbol{W}\mu_\beta^q - \boldsymbol{\varphi}_J\mu_\theta^q) \\ &+ (\mu_\beta^q - \mu_\beta^0)^\top\Sigma_\beta^{0\,-1}(\mu_\beta^q - \mu_\beta^0)\Big\},\end{aligned}$$

from which we deduce that $q_3(\sigma^2)$ is an inverse gamma, $IG(r_{q,\sigma}/2, s_{q,\sigma}/2)$ with the expressions for $r_{q,\sigma}$ and $s_{q,\sigma}$ given in `Algorithm 1`.

- For τ^2, the mean field udpate $q_4(\tau^2)$ has the form

$$\log q_4(\tau^2) \doteq \mathrm{E}_{-4}\left[\log p(\tau^2) + \log p(\boldsymbol{\theta}_J|\sigma^2, \tau^2, \psi)\right],$$

where

$$\begin{aligned}
\mathrm{E}_{-4}(\log p(\tau^2)) &\doteq -\left(\frac{r_{0,\tau}}{2} + 1\right)\log\tau^2 - \frac{s_{0,\tau}}{2\tau^2}, \\
\mathrm{E}_{-4}(\log p(\boldsymbol{\theta}_J|\sigma^2, \tau^2, \psi)) &\doteq -\frac{J}{2}\log\tau^2 \\
&\quad -\frac{1}{2\tau^2}\mathrm{E}\left(\frac{1}{\sigma^2}\right)\mathrm{E}_{-4}(\boldsymbol{\theta}_J^\top \mathrm{diag}(\mathrm{E}(\Gamma^{-1}))\boldsymbol{\theta}_J) \\
&\doteq -\frac{J}{2}\log\tau^2 \\
&\quad -\frac{1}{2\tau^2}\frac{r_{q,\tau}}{s_{q,\tau}}\mathrm{tr}((\Sigma_\theta^q + \mu_\theta^q \mu_\theta^{q\top})\mathrm{diag}(\mathrm{E}_{-4}(\Gamma^{-1}))).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\log q_4(\tau^2) \doteq &-\left(\frac{r_{0,\tau} + J}{2} + 1\right)\log\tau^2 \\
&-\frac{1}{2\tau^2}\left\{s_{0,\tau} + \frac{r_{q,\tau}}{s_{q,\tau}}\mathrm{tr}((\Sigma_\theta^q + \mu_\theta^q \mu_\theta^{q\top})\mathrm{diag}(\mathrm{E}(\Gamma^{-1})))\right\},
\end{aligned}$$

from which we deduce that $q_4(\tau^2)$ is an inverse gamma, $IG(r_{q,\tau}/2, s_{q,\tau}/2)$ with the expressions for $r_{q,\tau}$ and $s_{q,\tau}$ given in `Algorithm 1`.

Each term in (2.8) is evaluated as

$$\begin{aligned}
\mathrm{E}(\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}_J, \sigma^2)) = &-\frac{n}{2}\log 2\pi - \frac{n}{2}\left\{\log(s_{q,\sigma}/2) - \psi(r_{q,\sigma}/2)\right\} \\
&-\frac{r_{q,\sigma}}{2s_{q,\sigma}}\left\{\mathrm{tr}(\boldsymbol{W}^\top \boldsymbol{W}\Sigma_\beta^q) + \mathrm{tr}(\varphi^\top \varphi \Sigma_\theta^q)\right. \\
&\left.+(y - \boldsymbol{W}\mu_\beta^q - \varphi\mu_\theta^q)^\top (y - \boldsymbol{W}\mu_\beta^q - \varphi\mu_\theta^q)\right\},
\end{aligned}$$

where $\psi(\cdot)$ denotes the digamma function,

$$\begin{aligned}
\mathrm{E}(\log p(\boldsymbol{\beta}|\sigma^2)) &= -\frac{p}{2}\log 2\pi - \frac{p}{2}\log\{\log(s_{q,\sigma}/2) - \psi(r_{q,\sigma}/2)\} \\
&\quad -\frac{1}{2}\log|\Sigma_\beta^0| - \frac{1}{2}\frac{r_{q,\sigma}}{s_{q,\sigma}}\left\{\mathrm{tr}(\Sigma_\beta^{0\,-1}\Sigma_\beta^q)\right. \\
&\quad \left.+(\mu_\beta^q - \mu_\beta^0)^\top \Sigma_\beta^{0\,-1}(\mu_\beta^q - \mu_\beta^0)\right\}, \\
\mathrm{E}(\log p(\boldsymbol{\theta}_J|\sigma^2, \tau^2, \psi)) &= -\frac{J}{2}\left\{\log 2\pi + \log(s_{q,\sigma}/2) - \psi(r_{q,\sigma}/2)\right. \\
&\quad \left.+\log(s_{q,\tau}/2) - \psi(r_{q,\tau}/2)\right\}
\end{aligned}$$

$$
\begin{aligned}
&+\frac{J(J+1)}{4}\left\{\sigma_\psi^q\sqrt{\frac{2}{\pi}}\exp\left(-\frac{{\mu_\psi^q}^2}{2{\sigma_\psi^q}^2}\right)\right.\\
&\left.+\mu_\psi^q\left(1-2\Phi\left(-\frac{\mu_\psi^q}{\sigma_\psi^q}\right)\right)\right\}\\
&-\frac{1}{2}\frac{r_{q,\tau}}{s_{q,\tau}}\frac{r_{q,\sigma}}{s_{q,\sigma}}\mathrm{tr}((\Sigma_\theta^q+\mu_\theta^q{\mu_\theta^q}^\top)\mathrm{diag}(\mathrm{E}(\Gamma^{-1}))),\\
\mathrm{E}(\log p(\psi)) &= \log w_0/2 - w_0 S_1(\mu_\psi^q, {\sigma_\psi^q}^2),\\
\mathrm{E}(\log p(\sigma^2)) &= (r_{0,\sigma}/2)\log(s_{0,\sigma}/2) - \log\Gamma(r_{0,\sigma}/2)\\
&\quad -(r_{0,\sigma}/2+1)\left\{\log(s_{q,\sigma}/2)-\psi(r_{q,\sigma}/2)\right\} - \frac{s_{0,\sigma}}{2}\frac{r_{q,\sigma}}{s_{q,\sigma}},\\
\mathrm{E}(\log p(\tau^2)) &= r_{0,\tau}/2\log(s_{0,\tau}/2) - \log\Gamma(r_{0,\tau}/2)\\
&\quad -(r_{0,\tau}/2+1)\left\{\log s_{q,\tau}/2-\psi(r_{q,\tau}/2)\right\} - \frac{s_{0,\tau}}{2}\frac{r_{q,\tau}}{s_{q,\tau}}.
\end{aligned}
$$

Further, each term in (2.9) is given by

$$
\begin{aligned}
\mathrm{E}(\log q_1(\boldsymbol{\beta})) &= -\frac{p}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_\beta^q| - \frac{p}{2},\\
\mathrm{E}(\log q_2(\boldsymbol{\theta}_J)) &= -\frac{J}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_\theta^q| - \frac{J}{2},\\
\mathrm{E}(\log q_3(\sigma^2)) &= r_{q,\sigma}/2\log(s_{q,\sigma}/2) - \log\Gamma(r_{q,\sigma}/2)\\
&\quad -(r_{q,\sigma}/2+1)\left\{\log(s_{q,\sigma}/2)-\psi(r_{q,\sigma}/2)\right\} - \frac{r_{q,\sigma}}{2},\\
\mathrm{E}(\log q_4(\tau^2)) &= r_{q,\tau}/2\log(s_{q,\tau}/2) - \log\Gamma(r_{q,\tau}/2)\\
&\quad -(r_{q,\tau}/2+1)\left\{\log(s_{q,\tau}/2)-\psi(r_{q,\tau}/2)\right\} - \frac{r_{q,\tau}}{2},\\
\mathrm{E}(\log q_5(\psi)) &= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log {\sigma_\psi^q}^2 - \frac{1}{2}.
\end{aligned}
$$

Conjugate variational updates and lower bound for model with monotone shape restriction

- For $\boldsymbol{\beta}$, the mean field update $q_1(\boldsymbol{\beta})$ takes the form

$$\log q_1(\boldsymbol{\beta}) \doteq \mathrm{E}_{-1}(\log p(\boldsymbol{\beta})) + \mathrm{E}_{-1}(\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta}_J,\sigma^2)),$$

where

$$
\begin{aligned}
\mathrm{E}_{-1}(\log p(\boldsymbol{\beta})) &\doteq -\frac{1}{2}\mathrm{E}_{-1}\left(\frac{1}{\sigma^2}\right)(\boldsymbol{\beta}-\mu_\beta^0)^\top {\Sigma_\beta^0}^{-1}(\boldsymbol{\beta}-\mu_\beta^0),\\
\mathrm{E}_{-1}(\log p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta}_J,\sigma^2)) &\doteq -\frac{1}{2}\mathrm{E}_{-1}\left(\frac{1}{\sigma^2}\right)\mathrm{E}_{-1}
\end{aligned}
$$

$$\times \left(\sum_{i=1}^{n} (y_i - \boldsymbol{w}_i^\top \boldsymbol{\beta} - \delta \boldsymbol{\theta}_J^\top \boldsymbol{\varphi}_J^a(x_i) \boldsymbol{\theta}_J)^2 \right)$$

$$\doteq -\frac{1}{2} \mathrm{E}_{-1}\left(\frac{1}{\sigma^2}\right) \sum_{i=1}^{n} \Big\{ (y_i - \boldsymbol{w}_i^\top \boldsymbol{\beta} - \delta \mathrm{tr}(\boldsymbol{\varphi}_J^a(x_i) \Sigma_\theta^q) - \delta {\mu_\theta^q}^\top \boldsymbol{\varphi}_J^a(x_i) \mu_\theta^q)^2 + 2\mathrm{tr}(\boldsymbol{\varphi}_J^a(x_i) \Sigma_\theta^q \psi(x_i) \Sigma_\theta^q) + 4 {\mu_\theta^q}^\top \boldsymbol{\varphi}_J^a(x_i) \Sigma_\theta^q \boldsymbol{\varphi}_J^a(x_i) \mu_\theta^q \Big\},$$

which is from well-known results about a quadratic form of a multivariate normal random vector. Thus, we have

$$\log q_1(\boldsymbol{\beta}) \doteq -\frac{1}{2} \left\{ \boldsymbol{\beta}^\top \left({\Sigma_\beta^0}^{-1} + \mathrm{E}_{-1}\left(\frac{1}{\sigma^2}\right) \sum_{i=1}^{n} \boldsymbol{w}_i^\top \boldsymbol{w}_i \right) \boldsymbol{\beta} - 2 \left({\Sigma_\beta^0}^{-1} \mu_\beta^0 + \mathrm{E}_{-1}\left(\frac{1}{\sigma^2}\right) \sum_{i=1}^{n} \boldsymbol{w}_i (y_i - \delta \mathrm{tr}(\boldsymbol{\varphi}_J^a(x_i) \Sigma_\theta^q) - \delta {\mu_\theta^q}^\top \boldsymbol{\varphi}_J^a(x_i) \mu_\theta^q) \right) \boldsymbol{\beta} \right\},$$

which implies that $q_1(\boldsymbol{\beta})$ is normal, $N(\mu_\beta^q, \Sigma_\beta^q)$, with μ_β^q and Σ_β^q as given in `Algorithm 2`.

- For τ^2, the derivation of the update for $q_4(\tau^2)$ is the same as in Section 2.3, except that $r_{q,\sigma}/s_{q,\sigma}$ is replaced by $\mathrm{E}_{-4}(1/\sigma)$.

To derive the variational lower bound, $\mathcal{L}(q)$, we need to compute the following terms:

$$\mathrm{E}(\log p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta}_J, \sigma^2)) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\mathrm{E}(\log \sigma^2) - \mathrm{E}\left(\frac{1}{\sigma^2}\right) \left\{ \sum_i (y_i - \boldsymbol{w}_i^\top \mu_\beta^q - \delta {\mu_\theta^q}^\top \boldsymbol{\varphi}_J^a(x_i) \mu_\theta^q - \delta \mathrm{tr}(\boldsymbol{\varphi}_J^a(x_i) \Sigma_\theta^q))^2 + \boldsymbol{w}_i^\top \Sigma_\beta^q \boldsymbol{w}_i + 2\mathrm{tr}(\boldsymbol{\varphi}_J^a(x_i) \Sigma_\theta^q \boldsymbol{\varphi}_J^a(x_i) \Sigma_\theta^q) + 4 {\mu_\theta^q}^\top \boldsymbol{\varphi}_J^a(x_i) \Sigma_\theta^q \boldsymbol{\varphi}_J^a(x_i) \mu_\theta^q \right\},$$

$$\mathrm{E}(\log p(\boldsymbol{\beta}|\sigma^2)) = -\frac{p}{2}\log 2\pi - \frac{p}{2}\mathrm{E}(\log \sigma^2) - \frac{1}{2}\log |\Sigma_\beta^0| - \frac{1}{2}\mathrm{E}\left(\frac{1}{\sigma^2}\right) \left\{ \mathrm{tr}({\Sigma_\beta^0}^{-1} \Sigma_\beta^q) + (\mu_\beta^q - \mu_\beta^0)^\top {\Sigma_\beta^0}^{-1} (\mu_\beta^q - \mu_\beta^0) \right\},$$

$$\mathrm{E}(\log p(\boldsymbol{\theta}_J|\sigma^2, \tau^2, \psi)) = -\frac{(J+1)}{2}\mathrm{E}(\log 2\pi\sigma) - \frac{1}{2}\log \sigma_0^2 - \frac{J}{2}\{\log(s_{q,\tau}/2) - \psi(r_{q,\tau}/2)\}$$

$$
\begin{aligned}
&+\frac{J(J+1)}{4}\left\{\sigma_\psi^q\sqrt{\frac{2}{\pi}}\exp\left(-\frac{{\mu_\psi^q}^2}{2{\sigma_\psi^q}^2}\right)\right.\\
&\left.+\mu_\psi^q\left(1-2\Phi\left(-\frac{\mu_\psi^q}{\sigma_\psi^q}\right)\right)\right\}\\
&-\frac{1}{2}\mathrm{E}\left(\frac{1}{\sigma}\right)\mathrm{tr}\left\{\left(\Sigma_\theta^q+\mu_\theta^q{\mu_\theta^q}^\top\right)\mathrm{diag}(\mathrm{E}(\Upsilon^{-1}))\right\},\\
\mathrm{E}(\log p(\psi)) &= \log\frac{w_0}{2}-w_0S_1(\mu_\psi^q,{\sigma_\psi^q}^2),\\
\mathrm{E}(\log p(\sigma^2)) &= \frac{r_{0,\sigma}}{2}\log\frac{s_{0,\sigma}}{2}-\log\Gamma\left(\frac{r_{0,\sigma}}{2}\right)-(r_{0,\sigma}/2+1)E(\log\sigma^2)\\
&\quad-\frac{s_{0,\sigma}}{2}E\left(\frac{1}{\sigma^2}\right),\\
\mathrm{E}(\log q(\sigma^2)) &= -\log I_1+a\mathrm{E}(\log\sigma^2)+b\mathrm{E}(1/\sigma)-c\mathrm{E}(1/\sigma^2),
\end{aligned}
$$

where I_1 is the normalizing constant of $q(\sigma^2)$, and $\mathrm{E}\left(\frac{1}{\sigma}\right)$ and $\mathrm{E}\left(\frac{1}{\sigma^2}\right)$ are the marginal expectations with respect to $q(\sigma^2)$. In particular, I_1 is given as

$$
\begin{aligned}
I_1 &= \int_0^\infty\left(\frac{1}{\sigma}\right)^{2a}\exp\left(\frac{b}{\sigma}-\frac{c}{\sigma^2}\right)d\sigma^2\\
&= 2\left\{(2c)^{-(a-1)}\Gamma(2a-2)\exp\left(\frac{b^2}{8c}\right)D_{-2a+2}\left(\frac{-b}{\sqrt{2c}}\right)\right\} \qquad (5.2)
\end{aligned}
$$

where $D_\nu(\cdot)$ denotes the parabolic cylinder function of order ν (Neville, Ormerod and Wand, 2014). Then, it follows from Neville, Ormerod and Wand (2014) that additional algebra reduces $\mathrm{E}\left(\frac{1}{\sigma}\right)$ and $\mathrm{E}\left(\frac{1}{\sigma^2}\right)$ to

$$
\begin{aligned}
\mathrm{E}\left(\frac{1}{\sigma}\right) &= I_1^{-1}\int_0^\infty\left(\frac{1}{\sigma}\right)^{2a+1}\exp\left(\frac{b}{\sigma}-\frac{c}{\sigma^2}\right)d\sigma^2\\
&= (2c)^{-1/2}\frac{\Gamma(2a-1)}{\Gamma(2a-2)}R_{2a-3}\left(\frac{-b}{\sqrt{2c}}\right), \qquad (5.3)\\
\mathrm{E}\left(\frac{1}{\sigma^2}\right) &= (2c)^{-1}\frac{\Gamma(2a)}{\Gamma(2a-2)}R_{2a-2}\left(\frac{-b}{\sqrt{2c}}\right)R_{2a-3}\left(\frac{-b}{\sqrt{2c}}\right), \qquad (5.4)
\end{aligned}
$$

respectively, where $R_\nu(x)=\frac{D_{-\nu-2}(x)}{D_{-\nu-1}(x)}$. Note that due to the numerical underflow when directly evaluating $R_\nu(x)$, Neville, Ormerod and Wand (2014) implement continued fraction approach to obtain both exact and numerically stable result. In addition to these "exact" results in (5.2)–(5.4), the Laplace approximation could be considered for an alternative method, which may account for the lower limit of zero in the integral by fitting an unnormalized truncated normal distribution to the integrand. Further, in computing VAIC and VBIC for VBM, we evaluate $\mathrm{E}_q(\log\sigma)$ by approximating $\mathrm{E}_q(\log\sigma)\approx\log\mathrm{E}_q(\sigma)$ with the first order Taylor series approximation of $\log\sigma$ about $E_q(\sigma)$, namely,

$\log \sigma \approx \log \mathrm{E}_q(\sigma) + \mathrm{E}_q^{-1}(\sigma - \mathrm{E}_q(\sigma))$, where

$$\begin{aligned}\mathrm{E}(\sigma) &= I_1^{-1} \int_0^\infty \left(\frac{1}{\sigma}\right)^{2a-1} \exp\left(\frac{b}{\sigma} - \frac{c}{\sigma^2}\right) d\sigma^2 \\ &= (2c)^{1/2} \left[(2a-3) R_{2a-4}\left(\frac{-b}{\sqrt{2c}}\right)\right]^{-1}.\end{aligned}$$

Acknowledgments

David Kwamena Mensah was supported by a Singapore International Graduate Award (SINGA). David Nott and Victor Meng Hwee Ong were supported by a Singapore Ministry of Education Academic Research Fund Tier 2 grant (R-155-000-143-112). Research of Seongil Jo was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2017R1D1A3B03035235). Research of Taeryon Choi was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2016R1D1A1B03932178). We greatly appreciate the editor, AE and reviewers for their helpful comments that improved the style and substance of the paper.

References

Adler, R. J. and Taylor, J. E. (2007). *Random fields and geometry. Springer Monographs in Mathematics*. Springer, New York. MR2319516

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

Cai, B. and Dunson, D. B. (2007). Bayesian multivariate isotonic regression splines: applications to carcinogenicity studies. *J. Amer. Statist. Assoc.* **102** 1158–1171. MR2412540

Chakraborty, S., Ghosh, M. and Mallick, B. K. (2012). Bayesian nonlinear regression for large p small n problems. *J. Multivariate Anal.* **108** 28–40. MR2903131

Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data. Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc., Hoboken, NJ. MR2848400

Curtis, S. M. and Ghosh, S. K. (2011). A variable selection approach to monotonic regression with Bernstein polynomials. *J. Appl. Stat.* **38** 961–976. MR2782409

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *J. Roy. Statist. Soc. Ser. B* **56** 501–514. MR1278223

Ghahramani, Z. and Beal, M. J. (2001). Propagation algorithms for variational Bayesian learning. *Advances in neural information processing systems 13* 507–513.

Giordano, R. J., Broderick, T. and Jordan, M. I. (2015). Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational

Bayes. In *Advances in Neural Information Processing Systems 28* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, eds.) 1441–1449.

Goldsmith, J., Wand, M. P. and Crainiceanu, C. (2011). Functional regression via variation Bayes. *Electron. J. Stat.* **5** 572–602. MR2813555

Grenander, U. (1981). *Abstract inference*. John Wiley & Sons, Inc., New York Wiley Series in Probability and Mathematical Statistics. MR0599175

Hofner, B., Kneib, T. and Hothorn, T. (2016). A unified framework of constrained regression. *Stat. Comput.* **26** 1–14. MR3439355

Hu, Y., Zhao, K. and Lian, H. (2015). Bayesian quantile regression for partially linear additive models. *Stat. Comput.* **25** 651–668. MR3334423

Ji, C., Shen, H. and West, M. (2010). Bounded approximations for marginal likelihoods Technical Report No. 10-05, Institute of Decision Sciences, Duke University.

Jo, S., Choi, T., Park, B. and Lenk, P. J. (2017). bsamGP: Bayesian Spectral Analysis Models using Gaussian Process Priors R package version 1.0.2.

Jordan, M. I. (2004). Graphical models. *Statist. Sci.* **19** 140–155. MR2082153

Jordan, M., Ghahramani, Z., Jaakkola, T. and Saul, L. (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning* **37** 183-233.

Knowles, D. A. and Minka, T. (2011). Non-conjugate Variational Message Passing for Multinomial and Binary Regression. In *Advances in Neural Information Processing Systems 24* (J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger, eds.) 1701–1709. Curran Associates, Inc. MR0402994

Ko, K., Qu, L. and Vannucci, M. (2009). Wavelet-based Bayesian estimation of partially linear regression models with long memory errors. *Statist. Sinica* **19** 1463–1478. MR2589192

Lázaro-Gredilla, M., Quiñonero-Candela, J., Rasmussen, C. E. and Figueiras-Vidal, A. R. (2010). Sparse spectrum Gaussian process regression. *J. Mach. Learn. Res.* **11** 1865–1881. MR2660655

Lenk, P. J. (1999). Bayesian inference for semiparametric regression using a Fourier representation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 863–879. MR1722244

Lenk, P. J. and Choi, T. (2017). Bayesian analysis of shape-restricted functions using Gaussian process priors. *Statist. Sinica* **27** 43–69. MR3642448

Lin, L. and Dunson, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika* **101** 303–317. MR3215349

Luts, J., Broderick, T. and Wand, M. P. (2014). Real-time semiparametric regression. *J. Comput. Graph. Statist.* **23** 589–615. MR3224647

Luts, J. and Wand, M. P. (2015). Variational inference for count response semiparametric regression. *Bayesian Anal.* **10** 991–1023. MR3432247

Meyer, M. C., Hackstadt, A. J. and Hoeting, J. A. (2011). Bayesian estimation and inference for generalised partial linear models using shape-restricted splines. *J. Nonparametr. Stat.* **23** 867–884. MR2854243

Neville, S. E., Ormerod, J. T. and Wand, M. P. (2014). Mean field vari-

ational Bayes for continuous sparse signal shrinkage: pitfalls and remedies. *Electron. J. Stat.* **8** 1113–1151. MR3263115

Nott, D. J., Tan, S. L., Villani, M. and Kohn, R. (2012). Regression density estimation with variational methods and stochastic approximation. *J. Comput. Graph. Statist.* **21** 797–820. MR2970920

O'Hagan, A. (1978). Curve fitting and optimal design for prediction. *J. Roy. Statist. Soc. Ser. B* **40** 1–42. MR0512140

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

Paciorek, C. J. (2007). Bayesian smoothing with Gaussian processes using Fourier basis functions in the spectralGP package. *Journal of Statistical Software* **19** 2.

Paisley, J. W., Blei, D. M. and Jordan, M. I. (2012). Variational Bayesian Inference with Stochastic Search. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

Riihimäki, J. and Vehtari, A. (2010). Gaussian processes with monotonicity information. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Y. W. Teh and M. Titterington, eds.). *Proceedings of Machine Learning Research* **9** 645–652.

Shively, T. S., Sager, T. W. and Walker, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 159–175. MR2655528

Tan, L. L., Ong, V. H., Nott, D. and Jasra, A. (2016). Variational inference for sparse spectrum Gaussian process regression. *Stat. Comput.* to appear. MR3538635

Titterington, D. M. (2004). Bayesian methods for neural networks and related models. *Statist. Sci.* **19** 128–139. MR2082152

Turner, R. E. and Sahani, M. (2011). Two problems with variational expectation maximisation for time-series models. In *Bayesian Time series models* (D. Barber, T. Cemgil and S. Chiappa, eds.) 5, 109–130. Cambridge University Press. MR2894235

Waldmann, E. and Kneib, T. (2015). Variational approximations in geoadditive latent Gaussian regression: mean and quantile regression. *Stat. Comput.* **25** 1247–1263. MR3401884

Wand, M. P. (2014). Fully simplified multivariate normal updates in non-conjugate variational message passing. *J. Mach. Learn. Res.* **15** 1351–1369. MR3214787

Wand, M. P. and Ormerod, J. T. (2011). Penalized wavelets: embedding wavelets into semiparametric regression. *Electron. J. Stat.* **5** 1654–1717. MR2870147

Wang, X. and Berger, J. O. (2016). Estimating shape constrained functions using Gaussian processes. *SIAM/ASA J. Uncertain. Quantif.* **4** 1–25. MR3452261

Wang, B. and Titterington, D. M. (2004). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *In Workshop on Artificial Intelligence and Statistics* 373–380.

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

Winn, J. and Bishop, C. M. (2005). Variational message passing. *J. Mach. Learn. Res.* **6** 661–694. MR2249835

Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge University Press.

You, C., Ormerod, J. T. and Müller, S. (2014). On variational Bayes estimation and variational information criteria for linear regression models. *Aust. N. Z. J. Stat.* **56** 73–87. MR3200293

Zhang, Z., Wang, D., Dai, G. and Jordan, M. I. (2014). Matrix-variate Dirichlet process priors with applications. *Bayesian Anal.* **9** 259–285. MR3216996

Zhao, K. and Lian, H. (2014). Variational inferences for partially linear additive models with variable selection. *Comput. Statist. Data Anal.* **80** 223–239. MR3240489