Fast and Scalable Variational Bayes Estimation of Spatial Econometric Models for Gaussian Data

Guohui Wu[1]

Abstract

Spatial econometric models have been widely used for analyzing cross-sectional data in which spatial dependence is of primary interest. Although proven successful, Bayesian estimation via Markov chain Monte Carlo (MCMC) for spatial econometric models can be computationally demanding as the size of data and complexity of models grow. This paper proposes two variational Bayes methods that are more scalable and computationally faster in estimating general spatial autoregressive and matrix exponential spatial specification models: the hybrid mean-field variational Bayes (MFVB) method and the integrated nonfactorized variational Bayes (INFVB) method. The hybrid MFVB method assumes posterior independence and, when applicable, can yield accurate results but tends to underestimate posterior variances. In comparison, the INFVB method provides more robust results by accounting for posterior dependence and is computationally appealing due to parallelization. We demonstrate that variational Bayesian inference can be a faster and more scalable alternative to the MCMC approach for Bayesian spatial econometric modeling. The effectiveness of our proposed methods for spatial econometric models is demonstrated through simulated examples and a real-world data application.

KEY WORDS: Big Data, Markov Chain Monte Carlo, Spatial Econometrics, Variational Bayesian Inference

[1]SAS Institute Inc., 100 SAS Campus Drive, Cary, NC 27513, raywu2014@gmail.com

1 Introduction

As a subfield of econometrics, spatial econometrics concentrates on accounting for spatial dependence in cross-sectional spatial data in a regression setting (Anselin, 2001; Elhorst, 2013). According to the type of the response variable in the data, cross-sectional spatial data can be divided into two types: Gaussian and non-Gaussian data. Depending on whether data is Gaussian, either linear or nonlinear spatial econometric models can be used for econometric analysis (LeSage and Pace, 2009). Despite the wide availability of spatial econometric models, estimating these models can be computationally challenging for practitioners, especially in the presence of big data. For Gaussian spatial data, there are many dedicated software packages for spatial econometric modeling; examples are the SPATIALREG procedure in SAS/ETS® (SAS Institute Inc., 2016) and the CSPATIALREG procedure in SAS® Econometrics (SAS Institute Inc., 2017).

Among a wide range of spatial econometric models, the spatial autoregressive (SAR) model (Whittle, 1954) and the matrix exponential spatial specification (MESS) model (LeSage and Pace, 2007) are often used for modeling Gaussian spatial data (LeSage and Pace, 2009; Rodrigues et al., 2014; Debarsy et al., 2015; Figueiredo and Da Silva, 2015; Strauß et al., 2017). From a modeling perspective, the primary difference between SAR and MESS models is that SAR models imply a geometric decay of spatial dependence whereas MESS models imply an exponential decay (LeSage and Pace, 2007). Computationally, MESS models simplify the log-likelihood calculation and hence are more advantageous than SAR models (LeSage and Pace, 2007). Apart from these differences, there is a close correspondence between SAR and MESS models if the same row-standardized spatial weights matrix is used in both models (LeSage and Pace, 2009; Debarsy et al., 2015).

Many estimation methods have been proposed for spatial econometric models, including maximum likelihood, quasi-maximum likelihood, generalized method of moments, and

Bayesian estimation (see LeSage and Pace, 2009; Elhorst, 2013, and the references therein). Recent years have seen diverse applications of Bayesian estimation via Markov chain Monte Carlo (MCMC) for spatial econometric models; for examples, see LeSage and Pace (2009) and the references therein. Although MCMC can provide asymptotically exact samples (Robert, 2004), it can be computationally intensive and time-consuming for complex spatial econometric models with big data. This is partially attributed to challenges that are associated with drawing representative samples from nonstandard posterior distributions and to the need to run Markov chain long enough to achieve an adequately precise summary of posterior quantities for the posterior distribution.

To overcome the computational bottleneck of MCMC, Bivand et al. (2014) propose integrated nested Laplace approximation (INLA) as an alternative to estimating some spatial econometric models for both Gaussian and non-Gaussian data. At the core of the INLA framework, model parameters are divided into two subsets—hyperparameters and latent effects parameters—and the key assumption is that the conditional distribution of latent effects given hyperparameters follows a multivariate Gaussian distribution (Rue et al., 2009). However, the performance of the INLA method is adversely affected by an increase in the number of hyperparameters and grid points chosen for them. Specifically for spatial data analysis, Bivand et al. (2015) discuss how to fit some spatial statistical models and spatial econometric models using R-INLA software. Recently, Gómez-Rubio et al. (2017) present a detailed introduction about how some spatial econometric models are implemented in the R-INLA software, and they demonstrate its computational speed relative to MCMC. Although the general SAR model is supported in R-INLA software (Gómez-Rubio and Palmí-Perales, 2017), the MESS models we study in this paper are currently not supported in R-INLA software.

With a growing number of applications in many disciplines, variational Bayesian inference has risen to be a faster alternative to MCMC for many complex models (see Ormerod

and Wand, 2010; Blei et al., 2017, and the references therein). Loosely speaking, variational Bayes methods aim to find a variational distribution to approximate the joint posterior distribution by minimizing the Kullback-Leibler divergence between the target and approximate distributions. Unlike the MCMC approach, which uses simulation to infer the joint posterior distribution, variational Bayes methods use optimization to fulfill the same inference purpose (Blei et al., 2017).

Two of the most commonly used variational Bayes methods are the mean-field variational Bayes (MFVB) method (Jordan et al., 1999) and the fixed-form variational Bayes (FFVB) method (Wainwright and Jordan, 2008; Honkela et al., 2010). The MFVB method relies on the assumption that the variational distribution can be factorized into a product form; this assumption can severely impact the estimates for posterior variances (Wang and Titterington, 2005). The FFVB method avoids the product-form assumption but restricts each variational factor to a parametric form and hence the method can be more computationally involved (Salimans et al., 2013). The distinction between the MFVB and FFVB methods is that the MFVB method applies to conjugate models. To relax the product-form assumption and preserve posterior dependences, Han et al. (2013) propose the integrated nonfactorized variational Bayes (INFVB) method and demonstrate its improved performance over the MFVB method for the Bayesian LASSO regression model. The key advantage of the INFVB method over its MFVB counterpart is that the INFVB method can provide more accurate and robust estimates by taking posterior dependence into account. Moreover, the INFVB algorithm can be parallelized and is therefore promising when scaling to big data.

Motivated by the success of variational Bayes methods, we propose a hybrid MFVB algorithm and the INFVB algorithm to achieve fast and reliable Bayesian estimation for spatial econometric models. In particular, we focus on two of these models for Gaussian data—the spatial autoregressive confused (SAC) model and MESS(1,1) model. These models nest many well-known spatial econometric models, including but not limited to the SAR and

MESS models. Our contributions are manifold. First, we are the first to consider variational Bayesian inference for spatial econometric models. Second, we develop our hybrid MFVB and INFVB algorithms to achieve fast and reliable Bayesian estimation for many widely used spatial econometric models. Relative to MCMC, the two variational Bayes algorithms we propose are preferable because of their increased computational speed and scalability to big data, which are demonstrated through simulated examples and a real application.

This paper is organized as follows. Section 2 introduces MCMC and the SAC and MESS(1,1) models. Section 3 provides a brief introduction to our hybrid MFVB and INFVB algorithms and outlines the corresponding algorithms we propose for the SAC and MESS(1,1) models. Section 4 presents two simulated examples to illustrate the effectiveness of our hybrid MFVB and INFVB algorithms and examines the scalability of these two algorithms to big data sets. Section 5 applies the two variational Bayes methods we propose to a real application. Discussion is provided in Section 6.

2 Spatial Econometric Models for Gaussian Data

Among the many spatial econometric models, we restrict our attention to two models for Gaussian spatial data—the spatial autoregressive confused (SAC) model and the matrix exponential spatial specification (MESS) model. Before we describe the SAC and MESS models, we need to introduce some notation. Consider some Gaussian areal data that have been collected across n areal units in space. To describe neighbor relationships among these areal units, we consider two $n \times n$ valid spatial weights matrices $\mathbf{W}_1$ and $\mathbf{W}_2$, which can be the same or different. Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, where y_i denotes the value of the continuous response variable for an areal unit $i = 1, 2, \ldots, n$. Furthermore, let $\mathbf{Z}$ be an $n \times p$ design matrix. For a vector $\boldsymbol{a}$, we denote $\|\boldsymbol{a}\|^2 = \boldsymbol{a}'\boldsymbol{a}$.

2.1 Spatial Autoregressive Confused (SAC) Model

The SAC model, also known as SARAR(1,1), is defined as (Manski, 1993)

$$\begin{aligned} \boldsymbol{y} &= \rho \mathbf{W}_1 \boldsymbol{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{u} \\ \boldsymbol{u} &= \lambda \mathbf{W}_2 \boldsymbol{u} + \boldsymbol{\epsilon} \end{aligned}, \tag{1}$$

where ρ is a spatial autoregressive coefficient that measures spatial dependence in the dependent variable and λ is a spatial autocorrelation coefficient that measures spatial dependence among the disturbance terms. The design matrix $\mathbf{X}$ may be the same as $\mathbf{Z}$ or may take the form $[\mathbf{Z}\ \mathbf{W}_1\mathbf{Z}]$. When $\mathbf{X}$ includes the spatial lag of covariates $\mathbf{W}_1\mathbf{Z}$, the SAC model in (1) accounts for local spillover. Moreover, $\boldsymbol{\beta}$ denotes the vector of regression coefficients whose length is clear in the context.

By convention (for example, see LeSage and Pace, 2009), $\mathbf{W}_1$ and $\mathbf{W}_2$ are spatial weights matrices with real eigenvalues. In addition, restrictions on ρ and λ are often imposed such that $\frac{1}{\omega_{min}} < \rho < \frac{1}{\omega_{max}}$ and $\frac{1}{r_{min}} < \lambda < \frac{1}{r_{max}}$. Here ω_{min} and ω_{max} denote the minimum and maximum eigenvalues of $\mathbf{W}_1$, and r_{min} and r_{max} denote the minimum and maximum eigenvalues of $\mathbf{W}_2$. When $\mathbf{W}_1$ and $\mathbf{W}_2$ are row-standardized, both ω_{max} and r_{max} are equal to 1. In this case, it is often of practical interest to consider the restrictions on ρ and λ to be $-1 < \rho < 1$ and $-1 < \lambda < 1$, although ω_{min} and r_{min} are usually greater than -1.

From the modeling perspective, the SAC model in (1) is versatile in accounting for three different sources of spatial dependence in the data: the endogenous interaction effect, the exogenous interaction effect, and the interaction effects in the error term (Elhorst, 2013). The SAC model in (1) is general and nests four well-known spatial econometric models: the spatial error model ($\rho = 0$ and $\mathbf{X} = \mathbf{Z}$), the spatial Durbin error model ($\rho = 0$ and $\mathbf{X} = [\mathbf{Z}\ \mathbf{W}_1\mathbf{Z}]$), the spatial autoregressive (SAR) model ($\lambda = 0$ and $\mathbf{X} = \mathbf{Z}$), and the spatial Durbin model ($\lambda = 0$ and $\mathbf{X} = [\mathbf{Z}\ \mathbf{W}_1\mathbf{Z}]$).

The likelihood function for the SAC model in (1) takes the form

$$p(\boldsymbol{y}|\boldsymbol{\beta},\rho,\sigma^2,\lambda) = (2\pi\sigma^2)^{-\frac{n}{2}}|\mathbf{A}||\mathbf{B}|\exp\left\{-\frac{[\mathbf{B}(\mathbf{A}\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})]'[\mathbf{B}(\mathbf{A}\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})]}{2\sigma^2}\right\}, \tag{2}$$

where $\mathbf{A} = \mathbf{I}_n - \rho\mathbf{W}_1$, $\mathbf{B} = \mathbf{I}_n - \lambda\mathbf{W}_2$, and $\mathbf{I}_n$ is the identity matrix of size n. Close examination of (2) raises a computational concern when dealing with big data—that is, computing matrix determinants $|\mathbf{A}|$ and $|\mathbf{B}|$ becomes prohibitive when n is large. To resolve this computational issue, two approximation techniques have been proposed: Taylor approximation (Barry and Pace, 1999) and Chebyshev approximation (Pace and LeSage, 2004). For the sake of brevity, we provide only the approximation that is based on the Taylor series expansion:

$$\ln|\mathbf{I}_n - \rho\mathbf{W}_1| \approx -\sum_{o=1}^{O}\frac{\rho^o \text{trace}(\mathbf{W}_1^o)}{o}, \tag{3}$$

and we refer to LeSage and Pace (2009) for a detailed discussion about the implementation of (3).

We now specify the prior distributions that will be needed later. Specifically, we consider the following prior specification: $\boldsymbol{\beta} \sim \text{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$; $\sigma^2 \sim \text{IG}(q_{\sigma^2}, r_{\sigma^2})$, where $\text{IG}(C, D)$ denotes an inverse-gamma distribution with shape parameter C and scale parameter D; $\rho \sim \text{Uniform}(1/\omega_{min}, 1/\omega_{max})$; and $\lambda \sim \text{Uniform}(1/r_{min}, 1/r_{max})$.

2.2 Matrix Exponential Spatial Specification (MESS) Model

Another type of spatial econometric model for Gaussian data is the MESS(1,1) model, which can be defined as

$$\begin{aligned} \mathbf{S}_1\boldsymbol{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{u} \\ \mathbf{S}_2\boldsymbol{u} &= \boldsymbol{\epsilon} \end{aligned}, \tag{4}$$

where $\mathbf{S}_1 = e^{\alpha\mathbf{W}_1}$, $\mathbf{S}_2 = e^{\tau\mathbf{W}_2}$, and $\boldsymbol{\epsilon} \sim \text{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$, as discussed in Debarsy et al. (2015). Here α and τ are two scalar spatial coefficients. Similar to the preceding SAC model, the

design matrix $\mathbf{X}$ may be the same as $\mathbf{Z}$ or may take the form $[\mathbf{Z}\ \mathbf{W}_1\mathbf{Z}]$, the latter of which incorporates local spatial dependence. Unlike the constrained spatial coefficients ρ and λ in the SAC model, α and τ are unconstrained because $\mathbf{S}_1$ and $\mathbf{S}_2$ are guaranteed to be invertible (Debarsy et al., 2015). We note that the MESS(1,1) model in (4) nests the MESS model ($\tau = 0$ and $\mathbf{X} = \mathbf{Z}$), MESS Durbin model ($\tau = 0$ and $\mathbf{X} = [\mathbf{Z}\ \mathbf{W}_1\mathbf{Z}]$), MESS error model ($\alpha = 0$ and $\mathbf{X} = \mathbf{Z}$), and MESS Durbin error model ($\alpha = 0$ and $\mathbf{X} = [\mathbf{Z}\ \mathbf{W}_1\mathbf{Z}]$) as special cases.

For spatial weights matrices $\mathbf{W}_1$ and $\mathbf{W}_2$, it holds that $|\mathbf{S}_1| = |\mathbf{S}_2| = 1$. As a result, the likelihood function for the MESS(1,1) model takes the form of

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \alpha, \tau) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{[\mathbf{S}_2(\mathbf{S}_1\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})]'\,[\mathbf{S}_2(\mathbf{S}_1\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})]}{2\sigma^2}\right\}. \tag{5}$$

To evaluate (5), we need to compute the product of the matrix exponential and a vector of the form $e^{\xi\mathbf{W}}\boldsymbol{v}$, where $\mathbf{W}$ is a square matrix and $\boldsymbol{v}$ is a column vector, both of size n. For finite ξ, one way to compute $e^{\xi\mathbf{W}}\boldsymbol{v}$ is to first compute the matrix exponential as

$$e^{\xi\mathbf{W}} = \sum_{j=0}^{\infty} \frac{\xi^j}{j!}\mathbf{W}^j$$

(Horn and Johnson, 1990) and then perform the matrix vector multiplication. As pointed out by Sidje (1998), a more efficient alternative is to avoid explicitly computing the matrix exponential. We explore these two alternatives by taking advantage of the R `expm` package (Goulet et al., 2017).

Comparing (2) and (5) reveals a key advantage of the MESS(1,1) model versus the SAC model. That is, the MESS(1,1) model is more computationally appealing than the SAC model because the former does not involve the calculation of matrix determinants whereas the latter does.

To complete the specification of the MESS(1,1) model, we need to choose prior distributions for the model parameters. In particular, we consider $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$, $\sigma^2 \sim \mathrm{IG}(q_{\sigma^2}, r_{\sigma^2})$, $\alpha \sim \mathrm{N}(\mu_\alpha, \sigma_\alpha^2)$, and $\tau \sim \mathrm{N}(\mu_\tau, \sigma_\tau^2)$.

2.3 Markov Chain Monte Carlo Algorithm

To facilitate later discussion, we present a brief introduction to MCMC. Let $p(\boldsymbol{y}|\boldsymbol{\Theta})$ denote the likelihood function of the data given a set of parameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M\}$ under a given model, where $\boldsymbol{\theta}_l$ denotes the lth block of model parameters for $l = 1, 2, \ldots, M$. In addition, let $p(\boldsymbol{\Theta})$ be the prior distribution function for $\boldsymbol{\Theta}$. By Bayes' theorem, we have

$$p(\boldsymbol{\Theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}, \boldsymbol{\Theta})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}),$$

where $p(\boldsymbol{y}, \boldsymbol{\Theta}) = p(\boldsymbol{y}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})$ is the joint distribution function of $\boldsymbol{y}$ and $\boldsymbol{\Theta}$. In most cases, this joint posterior distribution is intractable, and thus the primary objective of the MCMC approach lies in learning about $p(\boldsymbol{\Theta}|\boldsymbol{y})$ via a collection of samples $\{\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)}, \ldots, \boldsymbol{\Theta}^{(m)}\}$ from the intractable posterior distribution.

At the core of the MCMC method is the drawing of samples from the joint posterior distribution $p(\boldsymbol{\Theta}|\boldsymbol{y})$. To this end, the MCMC approach breaks down the problem of sampling from $p(\boldsymbol{\Theta}|\boldsymbol{y})$ into sampling from the conditional distributions $p(\boldsymbol{\theta}_l|\boldsymbol{y}, \boldsymbol{\Theta}_{-l})$, where $\boldsymbol{\Theta}_{-l} = \boldsymbol{\Theta} \setminus \boldsymbol{\theta}_l$ for $l = 1, 2, \ldots, M$. Among the many MCMC algorithms that produce samples from a target distribution, the two most widely used algorithms are the Gibbs sampler and the Metropolis-Hastings algorithm (Gelman et al., 2003). If each conditional distribution $p(\boldsymbol{\theta}_l|\boldsymbol{y}, \boldsymbol{\Theta}_{-l})$ is a standard distribution, iterating the sampling from this conditional distribution over the M blocks of parameters leads to the Gibbs sampler. Otherwise, the Metropolis-Hastings algorithm is often used to sample from nonstandard full conditional distributions. In general, it is common to combine both the Gibbs sampler and the Metropolis-Hastings algorithm: the former for standard full conditional distributions and the latter for nonstandard full conditional distributions. Such a hybrid approach is called the Metropolis-within-Gibbs algorithm.

We now develop MCMC algorithms for the SAC and MESS(1,1) models that are described in Section 2.1 and Section 2.2. Figure 1 presents the derived full conditional distributions for these two models. According to Figure 1, the full conditional distributions for spatial

dependence parameters ρ and λ in the SAC model are nonstandard. Similarly, the full conditional distributions for α and τ in the MESS(1,1) model are nonstandard. Although the inverse CDF method proposed by LeSage and Pace (2009) is applicable, its efficiency can be impaired when it is difficult to compute the marginal distribution $p(\rho, \lambda|\boldsymbol{y})$ in the SAC model and $p(\alpha, \tau|\boldsymbol{y})$ in the MESS(1,1) model. Instead, we take advantage of elliptical slice sampling (Murray et al., 2010) to draw samples for ρ and λ in the SAC model and for α and τ in the MESS(1,1) model. In the current context, the elliptical slice sampling algorithm does not require data-specific user-defined tuning, and it is very general.

1. $\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2, \rho, \lambda \sim \mathrm{N}(\widehat{\boldsymbol{\mu}}_\beta, \widehat{\boldsymbol{\Sigma}}_\beta)$, where

$$\widehat{\boldsymbol{\Sigma}}_\beta = \left\{ \frac{(\mathbf{BX})'(\mathbf{BX})}{\sigma^2} + \boldsymbol{\Sigma}_\beta^{-1} \right\}^{-1} \text{ and } \quad \widehat{\boldsymbol{\mu}}_\beta = \widehat{\boldsymbol{\Sigma}}_\beta \left\{ \frac{(\mathbf{BX})'\mathbf{BA}\boldsymbol{y}}{\sigma^2} + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta \right\}.$$

2. $\sigma^2|\boldsymbol{y}, \boldsymbol{\beta}, \rho, \lambda \sim \mathrm{IG}(\widehat{q}_{\sigma^2}, \widehat{r}_{\sigma^2})$, where

$$\widehat{q}_{\sigma^2} = \frac{n}{2} + q_{\sigma^2} \text{ and } \widehat{r}_{\sigma^2} = r_{\sigma^2} + \frac{[\mathbf{B}(\mathbf{A}\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})]'[\mathbf{B}(\mathbf{A}\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})]}{2}.$$

3. $p(\rho|\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \lambda) \propto |\mathbf{I}_n - \rho\mathbf{W}_1| \exp\left\{ -\frac{[\mathbf{B}(\mathbf{A}\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})]'[\mathbf{B}(\mathbf{A}\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})]}{2\sigma^2} \right\} 1\left(\frac{1}{\omega_{min}} < \rho < \frac{1}{\omega_{max}} \right)$.

4. $p(\lambda|\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \rho) \propto |\mathbf{I}_n - \lambda\mathbf{W}_2| \exp\left\{ -\frac{[\mathbf{B}(\mathbf{A}\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})]'[\mathbf{B}(\mathbf{A}\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})]}{2\sigma^2} \right\} 1\left(\frac{1}{r_{min}} < \lambda < \frac{1}{r_{max}} \right)$.

1. $\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2, \alpha, \tau \sim \mathrm{N}(\widehat{\boldsymbol{\mu}}_\beta, \widehat{\boldsymbol{\Sigma}}_\beta)$, where

$$\widehat{\boldsymbol{\Sigma}}_\beta = \left\{ \frac{(\mathbf{S}_2\mathbf{X})'(\mathbf{S}_2\boldsymbol{X})}{\sigma^2} + \boldsymbol{\Sigma}_\beta^{-1} \right\}^{-1} \text{ and } \widehat{\boldsymbol{\mu}}_\beta = \widehat{\boldsymbol{\Sigma}}_\beta \left\{ \frac{(\mathbf{S}_2\mathbf{X})'\mathbf{S}_2\mathbf{S}_1\boldsymbol{y}}{\sigma^2} + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta \right\}.$$

2. $\sigma^2|\boldsymbol{y}, \boldsymbol{\beta}, \alpha, \tau \sim \mathrm{IG}(\widehat{q}_{\sigma^2}, \widehat{r}_{\sigma^2})$, where

$$\widehat{q}_{\sigma^2} = \frac{n}{2} + q_{\sigma^2} \text{ and } \widehat{r}_{\sigma^2} = r_{\sigma^2} + \frac{[\mathbf{S}_2(\mathbf{S}_1\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})]'[\mathbf{S}_2(\mathbf{S}_1\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})]}{2}.$$

3. $p(\alpha|\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \tau) \propto \exp\left\{ -\frac{[\mathbf{S}_2(\mathbf{S}_1\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})]'[\mathbf{S}_2(\mathbf{S}_1\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})]}{2\sigma^2} \right\} \exp\left\{ -\frac{(\alpha-\mu_\alpha)^2}{2\sigma_\alpha^2} \right\}$.

4. $p(\tau|\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \alpha) \propto \exp\left\{ -\frac{[\mathbf{S}_2(\mathbf{S}_1\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})]'[\mathbf{S}_2(\mathbf{S}_1\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})]}{2\sigma^2} \right\} \exp\left\{ -\frac{(\tau-\mu_\tau)^2}{2\sigma_\tau^2} \right\}$.

Figure 1: Full conditional distributions for the SAC model (top) and the MESS(1,1) model (bottom).

3 Variational Bayes

In this section, we propose variational Bayesian inference for the SAC and MESS(1,1) models that are described in Section 2. We start by introducing the mean-field variational Bayes (MFVB) algorithm, which imposes a product-form restriction. To relax the product-form restriction, we then introduce the integrated nonfactorized variational Bayes (INFVB) algorithm. Lastly, we develop both the MFVB and INFVB algorithms for the SAC and MESS(1,1) models.

In essence, variational Bayes approaches aim to find variational distributions $q(\boldsymbol{\Theta})$ to approximate the target posterior density $p(\boldsymbol{\Theta}|\boldsymbol{y})$. Among many candidate $q(\boldsymbol{\Theta})$s, the optimal approximate variational distribution, $q^*(\boldsymbol{\Theta})$, is the one that has a closest match to the joint $p(\boldsymbol{\Theta}|\boldsymbol{y})$ in the following sense

$$q^*(\boldsymbol{\Theta}) = \arg\min_{q \in Q} \int q(\boldsymbol{\Theta}) \ln \frac{q(\boldsymbol{\Theta})}{p(\boldsymbol{\Theta}|\boldsymbol{y})} d\boldsymbol{\Theta} \tag{6}$$

$$= \arg\max_{q \in Q} \underline{p}(\boldsymbol{y}; q) \tag{7}$$

(Ormerod and Wand, 2010), where $\ln \underline{p}(\boldsymbol{y}; q) = \int q(\boldsymbol{\Theta}) \ln \frac{p(\boldsymbol{y}, \boldsymbol{\Theta})}{q(\boldsymbol{\Theta})} d\boldsymbol{\Theta}$ is the lower bound on the logarithm of marginal likelihood, $\ln p(\boldsymbol{y})$. This lower bound is often called the evidence lower bound (ELBO). Put simply, the optimal approximate density function $q^*(\boldsymbol{\Theta})$ minimizes $\mathrm{KL}\left(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}|\boldsymbol{y})\right)$, which is the Kullback-Leibler divergence between $q(\boldsymbol{\Theta})$ and $p(\boldsymbol{\Theta}|\boldsymbol{y})$. Or equivalently, the optimal approximate density $q^*(\boldsymbol{\Theta})$ maximizes the ELBO.

3.1 Mean-Field Variational Bayes

To make the optimization in (6) or (7) more convenient to solve, the MFVB algorithm imposes the product-form restriction as follows (Wainwright and Jordan, 2008):

$$q(\boldsymbol{\Theta}) = \prod_{l=1}^{M} q_l(\boldsymbol{\theta}_l). \tag{8}$$

According to (8), the approximate density function $q(\boldsymbol{\Theta})$ is assumed to have a product form, which implies posterior independence between different blocks of model parameters. Thanks to this product-form restriction, each variational factor $q_l(\boldsymbol{\theta}_l)$ can be shown to have the following form:

$$q_l(\boldsymbol{\theta}_l) \propto \exp\left\{\mathrm{E}_{\boldsymbol{\Theta}_{-l}}\left[\ln p(\boldsymbol{\theta}_l|\boldsymbol{y}, \boldsymbol{\Theta}_{-l})\right]\right\}, \tag{9}$$

(Ormerod and Wand, 2010), where $\mathrm{E}_{\boldsymbol{\Theta}_{-l}}\left[\ln p(\boldsymbol{\theta}_l|\boldsymbol{y}, \boldsymbol{\Theta}_{-l})\right] = \int \ln p(\boldsymbol{\theta}_l|\boldsymbol{y}, \boldsymbol{\Theta}_{-l}) \prod_{s=1, s\neq l}^{M} q_s(\boldsymbol{\theta}_s) d\boldsymbol{\Theta}_{-l}$ for $l = 1, 2, \ldots, M$. The implication of (9) is that solving $q_l(\boldsymbol{\theta}_l)$ can be trivial if conjugacy holds for all blocks of parameters. However, if conjugacy does not hold for $\boldsymbol{\theta}_l$, then $q_l(\boldsymbol{\theta}_l)$ does not have a recognizable form. Consequently, the MFVB approach is often applicable to conjugate models, and it bears some resemblance to the Gibbs sampler in the MCMC paradigm.

Akin to the Metropolis-within-Gibbs algorithm under the MCMC umbrella, the MFVB approach can be extended to accommodate nonconjugate models by further assuming a parametric form for any nonstandard variational distribution function $q_l(\boldsymbol{\theta}_l)$, as is done in the FFVB approach; see Salimans et al. (2013) and the references therein. Hereafter, we refer to such an extended MFVB method as a *hybrid MFVB* approach to emphasize the product-form restriction. Note that a parametric distribution form can result from either an explicit assumption (such as FFVB) or an implicit assumption (such as Laplace approximation) according to various methods for nonconjugate models. For a comprehensive review of these methods, see Wang and Blei (2013) and Tran et al. (2016) and the references therein. For the sake of brevity and clarity of presentation, following Wang and Blei (2013), we focus on using Laplace approximation to approximate any continuous and nonstandard variational distribution in our hybrid MFVB algorithm.

For a continuous and nonstandard variational distribution function $q_l(\boldsymbol{\theta}_l)$, Laplace ap-

proximation proceeds by using a second-order Taylor expansion for $\ln q_l(\boldsymbol{\theta}_l)$,

$$\ln q_l(\boldsymbol{\theta}_l) \approx \texttt{const} + \ln q_l(\widehat{\boldsymbol{\theta}}_l) + \frac{1}{2}(\boldsymbol{\theta}_l - \widehat{\boldsymbol{\theta}}_l)'\mathbf{H}(\widehat{\boldsymbol{\theta}}_l)(\boldsymbol{\theta}_l - \widehat{\boldsymbol{\theta}}_l), \tag{10}$$

where `const` denotes a constant, $\widehat{\boldsymbol{\theta}}_l = \arg\max_{\boldsymbol{\theta}_l} \ln q_l(\boldsymbol{\theta}_l)$ and $\mathbf{H}(\widehat{\boldsymbol{\theta}}_l) = \frac{\partial^2 \ln q_l(\boldsymbol{\theta}_l)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_l}$. It follows from (10) that $q_l(\boldsymbol{\theta}_l) \approx \mathrm{N}(\widehat{\boldsymbol{\theta}}_l, -\mathbf{H}^{-1}(\widehat{\boldsymbol{\theta}}_l))$. Hence, the use of Laplace approximation implies a Gaussian distribution for a continuous and nonstandard variational distribution.

Taking advantage of Laplace approximation, our hybrid MFVB algorithms for the SAC and MESS(1,1) models are provided in Algorithm 1 and Algorithm 2, respectively. Note that the parameters in each model are divided into four blocks: $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \sigma^2, \rho, \lambda\}$ for the SAC model and $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \sigma^2, \alpha, \tau\}$ for the MESS(1,1) model. Here, Laplace approximation is used for deriving variational distributions for ρ and λ in the SAC model and for α and τ in the MESS(1,1) model.

Algorithm 1: Hybrid MFVB Algorithm for the SAC Model (Closed-form updates are used for $q(\boldsymbol{\beta})$ and $q(\sigma^2)$ whereas Laplace approximation is used for $q(\rho)$ and $q(\lambda)$.)

1. Initialize $\boldsymbol{\Omega} = \{\widetilde{\boldsymbol{\mu}}_\beta,\ \widetilde{\boldsymbol{\Sigma}}_\beta,\ \widetilde{q}_{\sigma^2},\ \widetilde{r}_{\sigma^2},\ \widetilde{\mu}_\rho,\ \widetilde{\sigma}^2_\rho,\ \widetilde{\mu}_\lambda,\ \widetilde{\sigma}^2_\lambda\}$.

2. Update $q(\boldsymbol{\beta}) = \mathrm{N}(\widetilde{\boldsymbol{\mu}}_\beta, \widetilde{\boldsymbol{\Sigma}}_\beta)$ with

$$\widetilde{\boldsymbol{\Sigma}}_\beta = \left\{ \frac{\widetilde{q}_{\sigma^2}}{\widetilde{r}_{\sigma^2}} \left[(\widetilde{\mathbf{B}}\mathbf{X})'(\widetilde{\mathbf{B}}\mathbf{X}) + \mathrm{Var}(\lambda)(\mathbf{W}_2\mathbf{X})'(\mathbf{W}_2\mathbf{X}) \right] + \boldsymbol{\Sigma}_\beta^{-1} \right\}^{-1}$$
$$\widetilde{\boldsymbol{\mu}}_\beta = \widetilde{\boldsymbol{\Sigma}}_\beta \left\{ \frac{\widetilde{q}_{\sigma^2}}{\widetilde{r}_{\sigma^2}} \left[(\widetilde{\mathbf{B}}\mathbf{X})'\widetilde{\mathbf{B}}\widetilde{\mathbf{A}}\boldsymbol{y} + \mathrm{Var}(\lambda)(\mathbf{W}_2\mathbf{X})'\mathbf{W}_2\widetilde{\mathbf{A}}\boldsymbol{y} \right] + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta \right\},$$

where $\widetilde{\mathbf{A}} = \mathbf{I}_n - \mathrm{E}_{q(\rho)}(\rho)\mathbf{W}_1$, $\widetilde{\mathbf{B}} = \mathbf{I}_n - \mathrm{E}_{q(\lambda)}(\lambda)\mathbf{W}_2$, and $\mathrm{Var}(\lambda) = \mathrm{E}_{q(\lambda)}(\lambda^2) - \mathrm{E}^2_{q(\lambda)}(\lambda)$.

3. Update $q(\sigma^2) = \mathrm{IG}(\widetilde{q}_{\sigma^2}, \widetilde{r}_{\sigma^2})$, where $\widetilde{q}_{\sigma^2} = \frac{n}{2} + q_{\sigma^2}$ and

$$\widetilde{r}_{\sigma^2} = r_{\sigma^2} + 0.5 \left\{ \left\| \widetilde{\mathbf{B}}(\widetilde{\mathbf{A}}\boldsymbol{y} - \mathbf{X}\widetilde{\boldsymbol{\mu}}_\beta) \right\|^2 + \mathrm{Var}(\rho) \left\| \widetilde{\mathbf{B}}\mathbf{W}_1\boldsymbol{y} \right\|^2 + \mathrm{trace}\left((\widetilde{\mathbf{B}}\mathbf{X})'\widetilde{\boldsymbol{\Sigma}}_\beta(\widetilde{\mathbf{B}}\mathbf{X}) \right) \right.$$
$$\left. + \mathrm{Var}(\lambda) \left[\left\| \mathbf{W}_2(\widetilde{\mathbf{A}}\boldsymbol{y} - \mathbf{X}\widetilde{\boldsymbol{\mu}}_\beta) \right\|^2 + \mathrm{trace}\left((\mathbf{W}_2\mathbf{X})'\widetilde{\boldsymbol{\Sigma}}_\beta(\mathbf{W}_2\mathbf{X}) \right) \right] \right\},$$

where $\mathrm{Var}(\rho) = \mathrm{E}_{q(\rho)}(\rho^2) - \mathrm{E}^2_{q(\rho)}(\rho)$.

4. Update $q(\rho) = \mathrm{N}\left(\widetilde{\mu}_\rho, \widetilde{\sigma}^2_\rho\right) 1\left(\frac{1}{\omega_{min}} < \rho < \frac{1}{\omega_{max}}\right)$, where $\widetilde{\mu}_\rho = \arg\max_\rho g(\rho)$, $\widetilde{\sigma}^2_\rho = -\frac{1}{g''(\widetilde{\mu}_\rho)}$, and

$$g(\rho) = \ln|\mathbf{I} - \rho\mathbf{W}_1| - \frac{\widetilde{q}_{\sigma^2}}{2\widetilde{r}_{\sigma^2}}(\mathbf{A}\boldsymbol{y} - \mathbf{X}\widetilde{\boldsymbol{\mu}}_\beta)' \left(\widetilde{\mathbf{B}}'\widetilde{\mathbf{B}} + \mathrm{Var}(\lambda)\mathbf{W}_2'\mathbf{W}_2 \right) (\mathbf{A}\boldsymbol{y} - \mathbf{X}\widetilde{\boldsymbol{\mu}}_\beta).$$

5. Update $q(\lambda) = \mathrm{N}\left(\widetilde{\mu}_\lambda, \widetilde{\sigma}^2_\lambda\right) 1\left(\frac{1}{r_{min}} < \lambda < \frac{1}{r_{max}}\right)$, where $\widetilde{\mu}_\lambda = \arg\max_\lambda f(\lambda)$, $\widetilde{\sigma}^2_\lambda = -\frac{1}{f''(\widetilde{\mu}_\lambda)}$, and

$$f(\lambda) = \ln|\mathbf{I}_n - \lambda\mathbf{W}_2| - \frac{\widetilde{q}_{\sigma^2}}{2\widetilde{r}_{\sigma^2}} \left\{ \left\| \mathbf{B}(\widetilde{\mathbf{A}}\boldsymbol{y} - \mathbf{X}\widetilde{\boldsymbol{\mu}}_\beta) \right\|^2 + \mathrm{Var}(\rho) \left\| \mathbf{B}\mathbf{W}_1\boldsymbol{y} \right\|^2 \right.$$
$$\left. + \mathrm{trace}\left((\mathbf{B}\mathbf{X})'\widetilde{\boldsymbol{\Sigma}}_\beta(\mathbf{B}\mathbf{X}) \right) \right\}.$$

6. Repeat Steps 2–5 until convergence.

Algorithm 2: Hybrid MFVB Algorithm for the MESS(1,1) Model (Closed-form updates are used for $q(\boldsymbol{\beta})$ and $q(\sigma^2)$ whereas Laplace approximation is used for $q(\alpha)$ and $q(\tau)$.)

1. Initialize $\boldsymbol{\Omega} = \{\widetilde{\boldsymbol{\mu}}_\beta,\ \widetilde{\boldsymbol{\Sigma}}_\beta,\ \widetilde{q}_{\sigma^2},\ \widetilde{r}_{\sigma^2},\ \widetilde{\mu}_\alpha,\ \widetilde{\sigma}^2_\alpha,\ \widetilde{\mu}_\tau,\ \widetilde{\sigma}^2_\tau\}$.
2. Update $q(\boldsymbol{\beta}) = \mathrm{N}(\widetilde{\boldsymbol{\mu}}_\beta, \widetilde{\boldsymbol{\Sigma}}_\beta)$ with

$$\widetilde{\boldsymbol{\Sigma}}_\beta = \left\{ \frac{\widetilde{q}_{\sigma^2}}{\widetilde{r}_{\sigma^2}} \mathbf{X}' \widetilde{\mathbf{S}_2'\mathbf{S}_2} \mathbf{X} + \boldsymbol{\Sigma}_\beta^{-1} \right\}^{-1} \text{ and } \widetilde{\boldsymbol{\mu}}_\beta = \widetilde{\boldsymbol{\Sigma}}_\beta \left\{ \frac{\widetilde{q}_{\sigma^2}}{\widetilde{r}_{\sigma^2}} \mathbf{X}' \widetilde{\mathbf{S}_2'\mathbf{S}_2} \widetilde{\mathbf{S}_1} \boldsymbol{y} + \boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\mu}_\beta \right\},$$

where $\widetilde{\mathbf{S}_2'\mathbf{S}_2} = \mathrm{E}_{q(\tau)}(\mathbf{S}_2'\mathbf{S}_2)$ and $\widetilde{\mathbf{S}_1} = \mathrm{E}_{q(\alpha)}(\mathbf{S}_1)$.
3. Update $q(\sigma^2) = \mathrm{IG}(\widetilde{q}_{\sigma^2}, \widetilde{r}_{\sigma^2})$, where $\widetilde{q}_{\sigma^2} = \frac{n}{2} + q_{\sigma^2}$ and

$$\begin{aligned}\widetilde{r}_{\sigma^2} = r_{\sigma^2} + 0.5 \Big\{ &\mathrm{E}_{q(\alpha)} \left[(\mathbf{S}_1\boldsymbol{y})' \widetilde{\mathbf{S}_2'\mathbf{S}_2} (\mathbf{S}_1\boldsymbol{y}) \right] - 2(\mathbf{X}\widetilde{\boldsymbol{\mu}}_\beta)' \widetilde{\mathbf{S}_2'\mathbf{S}_2} \widetilde{\mathbf{S}_1} \boldsymbol{y} \\ &+ (\mathbf{X}\widetilde{\boldsymbol{\mu}}_\beta)' \widetilde{\mathbf{S}_2'\mathbf{S}_2} (\mathbf{X}\widetilde{\boldsymbol{\mu}}_\beta) + \mathrm{trace}(\widetilde{\mathbf{S}_2'\mathbf{S}_2} \mathbf{X} \widetilde{\boldsymbol{\Sigma}}_\beta \mathbf{X}') \Big\}.\end{aligned}$$

4. Update $q(\alpha) = \mathrm{N}(\widetilde{\mu}_\alpha, \widetilde{\sigma}^2_\alpha)$, where $\widetilde{\mu}_\alpha = \arg\max_\alpha g(\alpha)$, $\widetilde{\sigma}^2_\alpha = -\frac{1}{g''(\widetilde{\mu}_\alpha)}$, and

$$g(\alpha) = -\frac{\widetilde{q}_{\sigma^2}}{2\widetilde{r}_{\sigma^2}} (\mathbf{S}_1\boldsymbol{y} - \mathbf{X}\widetilde{\boldsymbol{\mu}}_\beta)' \widetilde{\mathbf{S}_2'\mathbf{S}_2} (\mathbf{S}_1\boldsymbol{y} - \mathbf{X}\widetilde{\boldsymbol{\mu}}_\beta) - \frac{(\alpha - \mu_\alpha)^2}{2\sigma^2_\alpha}.$$

5. Update $q(\tau) = \mathrm{N}(\widetilde{\mu}_\tau, \widetilde{\sigma}^2_\tau)$, where $\widetilde{\mu}_\tau = \arg\max_\tau h(\tau)$, $\widetilde{\sigma}^2_\tau = -\frac{1}{h''(\widetilde{\mu}_\tau)}$, and

$$h(\tau) = -\frac{\widetilde{q}_{\sigma^2}}{2\widetilde{r}_{\sigma^2}} \left\{ \mathrm{E}_{q(\alpha)} \left\| \mathbf{S}_2(\mathbf{S}_1\boldsymbol{y} - \mathbf{X}\widetilde{\boldsymbol{\mu}}_\beta) \right\|^2 + \mathrm{trace}\left((\mathbf{S}_2\mathbf{X})' \widetilde{\boldsymbol{\Sigma}}_\beta (\mathbf{S}_2\mathbf{X}) \right) \right\} - \frac{(\tau - \mu_\tau)^2}{2\sigma^2_\tau}.$$

6. Repeat Steps 2–5 until convergence.

3.2 Integrated Nonfactorized Variational Bayes

The product-form restriction for the MFVB has some practical consequences. When the posterior dependence between different blocks of model parameters is weak or negligible, the MFVB approach can yield accurate estimates for posterior means and variances. However, when the assumption of posterior independence is untenable, MFVB tends to underestimate posterior variances yet captures posterior means (Blei and Jordan, 2006). To relax the

product-form restriction in MFVB, Han et al. (2013) propose an integrated nonfactorized variational Bayes (INFVB) method. The primary advantages of the INFVB method are twofold. Inferentially, the INFVB approach can provide more robust and accurate results than its MFVB counterpart by capturing posterior dependence between different blocks of model parameters. Computationally, the INFVB algorithm can be parallelized and scalable to big data.

To capture posterior dependence between model parameters, the INFVB method begins with the following decomposition:

$$q_{\text{INFVB}}(\boldsymbol{\Theta}) = q(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d)q(\boldsymbol{\Theta}_d), \tag{11}$$

where $\boldsymbol{\Theta} = \boldsymbol{\Theta}_c \bigcup \boldsymbol{\Theta}_d$ and where $\boldsymbol{\Theta}_c$ and $\boldsymbol{\Theta}_d$ denote two disjoint subsets of model parameters, the meaning of which shall become clear later. Substituting (11) into (6) yields the following (Han et al., 2013):

$$q^*_{\text{INFVB}}(\boldsymbol{\Theta}) = \underset{q_{\text{INFVB}}(\boldsymbol{\Theta})}{\arg\min} \int q(\boldsymbol{\Theta}_d) \left[\int q(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d) \ln \frac{q(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d)}{p(\boldsymbol{\Theta}_c, \boldsymbol{\Theta}_d|\boldsymbol{y})} d\boldsymbol{\Theta}_c + \ln q(\boldsymbol{\Theta}_d) \right] d\boldsymbol{\Theta}_d. \tag{12}$$

Compared with (6), (12) can be more challenging to solve. As a result, Han et al. (2013) consider discretizing the variational distribution $q(\boldsymbol{\Theta}_d)$ on a set of K grid points $\left\{\boldsymbol{\Theta}_d^{(1)}, \boldsymbol{\Theta}_d^{(2)}, \ldots, \boldsymbol{\Theta}_d^{(K)}\right\}$; that is, $q(\boldsymbol{\Theta}_d) = \sum_{k=1}^{K} q(\boldsymbol{\Theta}_d^{(k)}) 1(\boldsymbol{\Theta}_d = \boldsymbol{\Theta}_d^{(k)})$. It is now clear that $\boldsymbol{\Theta}_d$ is the subset of model parameters to be discretized and that $\boldsymbol{\Theta}_c = \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_d$. Hence, the dimension of $\boldsymbol{\Theta}_d$ equals to the total number of model parameters that need to be discretized. Importantly, the selected grid points for $\boldsymbol{\Theta}_d$ should cover the domain of the marginal posterior, $p(\boldsymbol{\Theta}_d|\boldsymbol{y})$, as much as possible. Algorithm 3 outlines the steps involved in solving the optimization problem in (12) (Han et al., 2013).

Algorithm 3: Sketch of the INFVB Algorithm

1. For each grid point $\boldsymbol{\Theta}_d^{(k)} \in \left\{\boldsymbol{\Theta}_d^{(1)}, \boldsymbol{\Theta}_d^{(2)}, \ldots, \boldsymbol{\Theta}_d^{(K)}\right\}$, do the following in parallel:

1a) Obtain the optimal variational distribution $q^*(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d^{(k)})$ according to

$$q^*(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d^{(k)}) = \underset{q(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d^{(k)})}{\arg\min} \int q(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d^{(k)}) \ln \frac{q(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d^{(k)})}{p(\boldsymbol{\Theta}_c|\boldsymbol{y}, \boldsymbol{\Theta}_d^{(k)})} d\boldsymbol{\Theta}_c.$$

1b) Evaluate the optimal variational distribution $q^*(\boldsymbol{\Theta}_d)$ at $\boldsymbol{\Theta}_d^{(k)}$ as

$$\ln q^*(\boldsymbol{\Theta}_d^{(k)}) = \texttt{const} - \int q^*(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d^{(k)}) \ln \frac{q^*(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d^{(k)})}{p(\boldsymbol{\Theta}_c, \boldsymbol{\Theta}_d^{(k)}|\boldsymbol{y})} d\boldsymbol{\Theta}_c.$$

2. Compute the approximate marginal posteriors as

$$q^*(\boldsymbol{\Theta}_d) = \sum_{k=1}^{K} q^*(\boldsymbol{\Theta}_d^{(k)}) 1(\boldsymbol{\Theta}_d = \boldsymbol{\Theta}_d^{(k)}) \text{ and } q^*(\boldsymbol{\Theta}_c) = \sum_{k=1}^{K} q^*(\boldsymbol{\Theta}_d^{(k)}) q^*(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d^{(k)}).$$

A few remarks about Algorithm 3: First, Step 1 suggests that the INFVB algorithm is naturally parallelizable. Second, there is often a need to balance between tractability and computation cost in Step 1. On one hand, the dimensionality of $\boldsymbol{\Theta}_d$ should be low to save computation cost. On the other hand, enforcing low dimensionality for $\boldsymbol{\Theta}_d$ tends to make it less tractable to solve for $q^*(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d^{(k)})$. When $q^*(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d^{(k)})$ does not have a closed form, we can appeal to a parametric-form assumption or the product-form restriction, as described in Section 3.1, for $q(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d^{(k)})$.

For the SAC model, we consider $\boldsymbol{\Theta}_d = \{\rho, \lambda\}$ and $\boldsymbol{\Theta}_c = \{\boldsymbol{\beta}, \sigma^2\}$. Similarly, we consider $\boldsymbol{\Theta}_d = \{\alpha, \tau\}$ and $\boldsymbol{\Theta}_c = \{\boldsymbol{\beta}, \sigma^2\}$ for the MESS(1,1) model. With this partition, the INFVB

algorithms for the SAC and MESS(1,1) models are presented in Algorithm 4 and Algorithm 5.

Algorithm 4: The INFVB Algorithm for the SAC Model

1. Denote $\boldsymbol{\Theta}_d^{-}\{\rho, \lambda\}$. For each grid point $\boldsymbol{\Theta}_d^{(k)} \in \left\{\boldsymbol{\Theta}_d^{(1)}, \boldsymbol{\Theta}_d^{(2)}, \ldots, \boldsymbol{\Theta}_d^{(K)}\right\}$, do the following in parallel:

1a) Initialize $\boldsymbol{\Omega} = \{\widetilde{\boldsymbol{\mu}}_\beta^{(k)}, \widetilde{\boldsymbol{\Sigma}}_\beta^{(k)}, \widetilde{q}_{\sigma^2}^{(k)}, \widetilde{r}_{\sigma^2}^{(k)}\}$ and $\text{ELBO}^{(k)}$.

1b) **while** *the increase in ELBO*$^{(k)}$ *is not negligible* **do**

i) Update $q(\boldsymbol{\beta}^{(k)}) = \text{N}(\widetilde{\boldsymbol{\mu}}_\beta^{(k)}, \widetilde{\boldsymbol{\Sigma}}_\beta^{(k)})$ with

$$\widetilde{\boldsymbol{\Sigma}}_\beta^{(k)} = \left\{\frac{\widetilde{q}_{\sigma^2}^{(k)}}{\widetilde{r}_{\sigma^2}^{(k)}}(\mathbf{B}_k\mathbf{X})'(\mathbf{B}_k\mathbf{X}) + \boldsymbol{\Sigma}_\beta^{-1}\right\}^{-1} \text{ and } \widetilde{\boldsymbol{\mu}}_\beta^{(k)} = \widetilde{\boldsymbol{\Sigma}}_\beta^{(k)}\left\{\frac{\widetilde{q}_{\sigma^2}}{\widetilde{r}_{\sigma^2}}(\mathbf{B}_k\mathbf{X})'\mathbf{B}_k\mathbf{A}_k\boldsymbol{y} + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta\right\},$$

where $\mathbf{A}_k = \mathbf{I}_n - \rho_k\mathbf{W}_1$ and $\mathbf{B}_k = \mathbf{I}_n - \lambda_k\mathbf{W}_2$.

ii) Update $q(\sigma^{2(k)}) = \text{IG}(\widetilde{q}_{\sigma^2}^{(k)}, \widetilde{r}_{\sigma^2}^{(k)})$ with $\widetilde{q}_{\sigma^2}^{(k)} = \frac{n}{2} + q_{\sigma^2}$ and
$\widetilde{r}_{\sigma^2}^{(k)} = r_{\sigma^2} + \frac{1}{2}\left\|\mathbf{B}_k(\mathbf{A}_k\boldsymbol{y} - \mathbf{X}\widetilde{\boldsymbol{\mu}}_\beta^{(k)})\right\|^2$.

iii) $\text{ELBO}^{(k)} =$
$-\widetilde{q}_{\sigma^2}^{(k)} \ln \widetilde{r}_{\sigma^2}^{(k)} - 0.5\left\{(\widetilde{\boldsymbol{\mu}}_\beta^{(k)} - \boldsymbol{\mu}_\beta)'\boldsymbol{\Sigma}_\beta^{-1}(\widetilde{\boldsymbol{\mu}}_\beta^{(k)} - \boldsymbol{\mu}_\beta) + \text{trace}\left(\widetilde{\boldsymbol{\Sigma}}_\beta^{(k)}\boldsymbol{\Sigma}_\beta^{-1}\right) - \ln|\widetilde{\boldsymbol{\Sigma}}_\beta^{(k)}|\right\}$.

end

2. The optimal approximate marginal posteriors are

$$q^*(\rho, \lambda) = \sum_{k=1}^{K} \omega_k 1(\boldsymbol{\Theta}_d = \boldsymbol{\Theta}_d^{(k)}), q^*(\boldsymbol{\beta}) = \sum_{k=1}^{K} \omega_k q(\boldsymbol{\beta}^{(k)}), \text{and } q^*(\sigma^2) = \sum_{k=1}^{K} \omega_k q(\sigma^{2(k)}),$$

where $\omega_k = \left(\text{ELBO}^{(k)} + \ln|\mathbf{A}_k| + \ln|\mathbf{B}_k|\right) / \sum_{k=1}^{K}\left(\text{ELBO}^{(k)} + \ln|\mathbf{A}_k| + \ln|\mathbf{B}_k|\right)$.

Algorithm 5: The INFVB Algorithm for the MESS(1,1) Model

1. Denote $\boldsymbol{\Theta}_d^{-}\{\alpha,\tau\}$. For each grid point $\boldsymbol{\Theta}_d^{(k)} \in \left\{\boldsymbol{\Theta}_d^{(1)}, \boldsymbol{\Theta}_d^{(2)}, \ldots, \boldsymbol{\Theta}_d^{(K)}\right\}$, do the following in parallel:

1a) Initialize $\boldsymbol{\Omega} = \{\widetilde{\boldsymbol{\mu}}_\beta^{(k)}, \widetilde{\boldsymbol{\Sigma}}_\beta^{(k)}, \widetilde{q}_{\sigma^2}^{(k)}, \widetilde{r}_{\sigma^2}^{(k)}\}$ and $\text{ELBO}_{(k)}$.

1b) **while** *the increase in ELBO*$^{(k)}$ *is not negligible* **do**

i) Update $q(\boldsymbol{\beta}^{(k)}) = \text{N}(\widetilde{\boldsymbol{\mu}}_\beta^{(k)}, \widetilde{\boldsymbol{\Sigma}}_\beta^{(k)})$ with

$$\widehat{\boldsymbol{\Sigma}}_\beta^{(k)} = \left\{\frac{\widetilde{q}_{\sigma^2}^{(k)}}{\widetilde{r}_{\sigma^2}^{(k)}}(\mathbf{S}_2^{(k)}\mathbf{X})'(\mathbf{S}_2^{(k)}\mathbf{X}) + \boldsymbol{\Sigma}_\beta^{-1}\right\}^{-1} \text{ and } \widehat{\boldsymbol{\mu}}_\beta^{(k)} = \widehat{\boldsymbol{\Sigma}}_\beta^{(k)}\left\{\frac{\widetilde{q}_{\sigma^2}^{(k)}}{\widetilde{r}_{\sigma^2}^{(k)}}(\mathbf{S}_2^{(k)})'\mathbf{S}_2^{(k)}\mathbf{S}_1^{(k)}\boldsymbol{y} + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta\right\},$$

where $\mathbf{S}_1^{(k)} = e^{\alpha_k \mathbf{W}_1}$ and $\mathbf{S}_2^{(k)} = e^{\tau_k \mathbf{W}_2}$.

ii) Update $q(\sigma^{2(k)}) = \text{IG}(\widetilde{q}_{\sigma^2}^{(k)}, \widetilde{r}_{\sigma^2}^{(k)})$ with $\widetilde{q}_{\sigma^2}^{(k)} = \frac{n}{2} + q_{\sigma^2}$ and
$\widetilde{r}_{\sigma^2}^{(k)} = r_{\sigma^2} + 0.5\left\|\mathbf{S}_2^{(k)}(\mathbf{S}_1^{(k)}\boldsymbol{y} - \mathbf{X}\widetilde{\boldsymbol{\mu}}_\beta^{(k)})\right\|^2$.

iii) $\text{ELBO}^{(k)} =$
$-\widetilde{q}_{\sigma^2}^{(k)} \ln \widetilde{r}_{\sigma^2}^{(k)} - 0.5\left\{(\widetilde{\boldsymbol{\mu}}_\beta^{(k)} - \boldsymbol{\mu}_\beta)'\boldsymbol{\Sigma}_\beta^{-1}(\widetilde{\boldsymbol{\mu}}_\beta^{(k)} - \boldsymbol{\mu}_\beta) + \text{trace}\left(\widetilde{\boldsymbol{\Sigma}}_\beta^{(k)}\boldsymbol{\Sigma}_\beta^{-1}\right) - \ln|\widetilde{\boldsymbol{\Sigma}}_\beta^{(k)}|\right\}$.

end

2. The optimal approximate marginal posteriors are

$$q^*(\alpha,\tau) = \sum_{k=1}^{K} \omega_k 1(\boldsymbol{\Theta}_d = \boldsymbol{\Theta}_d^{(k)}), \, q^*(\boldsymbol{\beta}) = \sum_{k=1}^{K} \omega_k q(\boldsymbol{\beta}^{(k)}), \text{ and } q^*(\sigma^2) = \sum_{k=1}^{K} \omega_k q(\sigma^{2(k)}),$$

where $\omega_k = \frac{\text{ELBO}^{(k)} - \frac{(\alpha_k - \mu_\alpha)^2}{2\sigma_\alpha^2} - \frac{(\tau_k - \mu_\tau)^2}{2\sigma_\tau^2}}{\sum_{k=1}^{K}\left(\text{ELBO}^{(k)} - \frac{(\alpha_k - \mu_\alpha)^2}{2\sigma_\alpha^2} - \frac{(\tau_k - \mu_\tau)^2}{2\sigma_\tau^2}\right)}$.

3.3 Convergence and Accuracy Assessment

We now describe the convergence and accuracy assessment for our hybrid MFVB and INFVB algorithms. In general, we can terminate the hybrid MFVB and INFVB algorithms when the increase in the ELBO is less than a specified tolerance ε. Nevertheless, as in the SAC

model, the ELBO does not have a closed form and so it is difficult to compute. Alternatively, we terminate the hybrid MFVB algorithm when $\|\boldsymbol{\Omega}_t - \boldsymbol{\Omega}_{t-1}\|^2 < \varepsilon$, where $\boldsymbol{\Omega}_t$ denotes the values of parameters that are associated with the variational density functions at the tth iteration (Tran et al., 2016). For the INFVB algorithm, we terminate the algorithm when the increase in ELBO is less than ε—that is, when $|\text{ELBO}_t - \text{ELBO}_{t-1}| < \varepsilon$, where ELBO_t is the ELBO at the tth iteration (Ormerod and Wand, 2010).

After obtaining the optimal variational density function $q^*(\boldsymbol{\Theta})$, we need to assess its accuracy relative to the exact posterior density $p(\boldsymbol{\Theta}|\boldsymbol{y})$ from the MCMC fitting. To this end, we follow Wand et al. (2011) and consider the accuracy score, which is defined as follows for all $\theta \in \boldsymbol{\Theta}$:

$$\text{accuracy}(q^*(\theta)) = 100\left(1 - \frac{1}{2}\int_{-\infty}^{\infty} |q^*(\theta) - p(\theta|\boldsymbol{y})| d\theta\right)\%.$$

By this definition, the accuracy score is expressed as a percentage and ranges between 0% and 100%. The greater the accuracy score, the more accurate is the variational density function relative to the exact posterior density for a specific model parameter.

4 Simulated Examples

We consider two simulated examples to illustrate the effectiveness of two variational Bayes fittings that use our hybrid MFVB and INFVB algorithms for the SAC and MESS(1,1) models that are discussed in Section 2. We implement the algorithms for MCMC, hybrid MFVB, and INFVB using R (R Core Team, 2016). For the two simulated examples presented in this section and the analysis of Boston housing data in Section 5, we report computation times for the MCMC and two variational Bayes fittings that are calibrated on a Ubuntu laptop with an Intel® Core™ i7-4710MQ 2.5GHz processor with 4GB of random access memory. Specifically for the INFVB algorithm, we accomplish parallelization using the `foreach` (Revolution Analytics, 2015) and `doParallel` (Microsoft Corporation, 2017) packages in R by

using all eight CPUs that are available. Knowing that for the MCMC fitting, the computation time depends on the run length of Markov chain, we emphasize that our main interest is to demonstrate the effectiveness of the two variational Bayes algorithms we develop for the SAC and MESS(1,1) models rather than to choose an optimal run length. For the hybrid MFVB and INFVB algorithms, we choose the tolerance threshold $\varepsilon = 10^{-6}$, unless specified otherwise.

4.1 Simulated Example for the SAC Model

We first consider a simulated example for the SAC model that is defined in (1). For data simulation, we set the number of observations $n = 200$ and the number of covariates $p = 6$. For regression coefficients $\boldsymbol{\beta}$, we set $\boldsymbol{\beta} = (1.2, -0.9, 0.6, -0.7, 0.8, -1.0)'$. For covariates x_j, we simulate $x_{ij} \overset{iid}{\sim} N(0,1)$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$. In addition, we set spatial coefficients $\rho = 0.5$ and $\lambda = 0.6$. For the variance parameter σ^2, we set $\sigma^2 = 0.4$. To create two spatial weights matrices $\mathbf{W}_1$ and $\mathbf{W}_2$, we first simulate two $n \times n$ matrices $\mathbf{C}$ and $\mathbf{M}$, whose (i,j)th entries C_{ij} and M_{ij} are

$$C_{ij} = \begin{cases} \text{Bernoulli}(8/n) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}, M_{ij} = \begin{cases} \text{Bernoulli}(9/n) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases},$$

for $i = 1, 2, \ldots, n-1$ and $j = i+1, \ldots, n$, and then we set $C_{ij} = C_{ji}$ and $M_{ij} = M_{ji}$ to ensure symmetry. The matrices $\mathbf{C}$ and $\mathbf{M}$ are then row-standardized to obtain $\mathbf{W}_1$ and $\mathbf{W}_2$, respectively.

For prior specification, we consider the following: $\boldsymbol{\beta} \sim N(\mathbf{0}, 10^2 \boldsymbol{I}_p)$, $\sigma^2 \sim \text{IG}(0.01, 0.01)$, $\rho \sim \text{Uniform}(-1, 1)$, and $\lambda \sim \text{Uniform}(-1, 1)$. For the MCMC implementation of the model, we run a total of 20,000 iterations, with the first 10,000 discarded as burn-in and the remaining samples used to draw inference. In the MCMC and two variational Bayes fittings, we perform exact computation for the log-determinant and matrix inverse rather

than appealing to an approximation, because n is small. For the INFVB fitting, we use 100 irregularly spaced grid points for each of ρ and λ, with 20 gridded points evenly spaced between -1 and -0.001 and the rest evenly spaced between 0 and 1.0. We choose more grid points between 0 and 1.0 to reflect that positive spatial dependence is more common.

Figure 2 compares the MCMC result against the approximate density functions from the hybrid MFVB for all parameters in the SAC model. From Figure 2, we see that the hybrid MFVB estimates are in close agreement with the MCMC estimates, with accuracy scores ranging between 96% and 99%. Figure 3 compares the MCMC result against the INFVB-based approximate density functions. As in the hybrid MFVB approach, there is a close correspondence between the INFVB estimates and MCMC estimates, with accuracies between 98% and 99%. According to Figure 2 and Figure 3, both the hybrid MFVB and INFVB methods capture posterior means and variances very well. Although the two variational Bayes fittings are excellent for the approximate posterior density functions, we see that the INFVB method outperforms its hybrid MFVB counterpart with slightly higher accuracy scores.

Because the hybrid MFVB assumes posterior independence by construction, the arguably indistinguishable difference in the accuracy scores between the hybrid MFVB and INFVB estimates leads to the conjecture that the posterior dependence among four parameter blocks $\boldsymbol{\beta}$, ρ, λ, and σ^2 is not strong enough to produce biased estimates for posterior variances in the hybrid MFVB algorithm. To better understand the strength of posterior dependence between model parameters, we compute pairwise Spearman's rank correlation coefficients by using MCMC samples and provide the results in Supplemental Appendix A. According to these coefficients, we do not see strong posterior dependence between any pair of model parameters, which explains why the hybrid MFVB fitting performs satisfactorily.

The computation time for MCMC fitting is about 67.3 seconds, compared to 0.64 seconds for the hybrid MFVB and 4.64 seconds for the INFVB. Given that the MCMC fitting takes

only about one minute, it is tempting to trivialize the computational advantages of both the hybrid MFVB and INFVB methods. To assess how MCMC and the two variational Bayes methods scale to the different sizes of data, we simulate data sets that have higher values of n for the SAC model by using the same simulation setup and MCMC specification that are previously described. Supplemental Appendix B presents the computation times that the MCMC, hybrid MFVB, and INFVB methods require for the fitting. These computation times show that both the hybrid MFVB and INFVB algorithms scale to large data sets better than their MCMC counterpart.

4.2 Simulated Example for the MESS(1,1) Model

We now consider a simulated example for the MESS(1,1) model. For data generation, we use the same setup for n, p, $\boldsymbol{\beta}$, σ^2, $\mathbf{W}_1$, $\mathbf{W}_2$, and covariates x_j as in the simulated example for the SAC model in Section 4.1. For spatial coefficients, we set $\alpha = \ln(0.5)$ and $\tau = \ln(0.4)$. With this particular choice of α and τ, the corresponding MESS(1,1) model is equivalent to the SAC model with $\rho = 0.5$ and $\lambda = 0.6$, provided that $\mathbf{W}_1$ is row-standardized (for a discussion about such an equivalence, see LeSage and Pace, 2007; Debarsy et al., 2015).

For prior specification, we consider the following: $\boldsymbol{\beta} \sim N(\mathbf{0}, 10^2\boldsymbol{I}_p)$, $\sigma^2 \sim \text{IG}(0.01, 0.01)$, $\alpha \sim \text{N}(0, 10^2)$, and $\tau \sim \text{N}(0, 10^2)$. Regarding the MCMC implementation, we run a total of 20,000 iterations with the first 10,000 discarded as burn-in and the remaining samples used to draw inference. For the INFVB fitting, we use 100 equally spaced grid points for each of α and τ, which range between $2\ln(0.5)$ and 0 and between $2\ln(0.4)$ and 0, respectively.

Figure 4 compares the MCMC result against approximate density functions from the hybrid MFVB. From Figure 4, we see that the hybrid MFVB estimates are in close agreement with the MCMC estimates, with accuracy scores ranging between 98% and 99%. Figure 5 compares the MCMC result against the INFVB-based approximate density functions. As

with the hybrid MFVB, there is a close correspondence between the INFVB estimates and MCMC estimates, with accuracies between 98% and 99%. Both the hybrid MFVB and INFVB algorithms are excellent for the approximate posterior density functions with identical accuracy scores for all model parameters.

Supplemental Appendix A presents Spearman's rank correlation coefficients between any pair of model parameters, which are computed using the MCMC samples. These coefficients indicate that there is no strong posterior dependence between parameters in the model. As a result, it is not unexpected that the hybrid MFVB fitting performs as well as the INFVB fitting despite the fact that the former assumes posterior independence. The computation time for the MCMC fitting is around 5504.94 seconds, compared to 100.08 seconds for the hybrid MFVB fitting and 36.48 seconds for the INFVB fitting. Because of parallelization, the INFVB fitting takes about 36.5% of the computation time that the hybrid MFVB fitting takes.

To summarize, we can see that, based on two simulated examples, both the hybrid MFVB and INFVB algorithms perform very well in approximating the exact posterior distributions for the SAC and MESS(1,1) models. However, as the size of data grows, the computation speedup for the two variational Bayes methods relative to MCMC becomes more pronounced. As the model becomes computationally involved, the INFVB algorithm can be more appealing than its hybrid MFVB counterpart because of the flexibility of parallelization in the former.

5 Application

To demonstrate the application of the hybrid MFVB and INFVB methods for the SAC and MESS(1,1) models, we consider the Boston housing data by Harrison and Rubinfeld (1978). This data set was considered by Bivand et al. (2014) to exemplify the use of INLA

for approximate Bayesian computation for linear and nonlinear spatial econometric models. The data set is available in the R package **spdep** (Bivand and Piras, 2015). The data contain 506 observations and 20 variables. The dependent variable `CMEDV` refers to corrected median values of owner-occupied housing (measured in $1,000s) in each tract. We refer interested readers to the **spdep** package for a detailed description about the data set.

5.1 The SAC Model

For the purpose of illustration, we consider the following SAC model:

$$\begin{aligned}\boldsymbol{y} &= \rho \mathbf{W}_1 \boldsymbol{y} + \beta_0 + \beta_1 \texttt{CRIM} + \beta_2 \texttt{ZN} + \beta_3 \texttt{INDUS} + \beta_4 \texttt{CHAS} \\ &\quad + \beta_5 \texttt{NOX}^2 + \beta_6 \texttt{RM}^2 + \beta_7 \texttt{AGE} + \beta_8 \ln(\texttt{DIS}) + \beta_9 \ln(\texttt{RAD}) \\ &\quad + \beta_{10} \texttt{TAX} + \beta_{11} \texttt{PTRATIO} + \beta_{12} \texttt{B} + \beta_{13} \ln(\texttt{LSTAT}) + \boldsymbol{u} \\ \boldsymbol{u} &= \lambda \mathbf{W}_1 \boldsymbol{u} + \boldsymbol{\epsilon}\end{aligned},$$

where the response variable is $y_i = \ln(\texttt{CMEDV}_i)$ for tract i and $\mathbf{W}_1$ is a row-standardized spatial weights matrix for all tracts that is based on the adjacency that is defined in the R object `boston.soi`.

For prior distributions, we consider the same specification as in Section 4.1. For the MCMC implementation, we run a total of 50,000 iterations with the first 10,000 iterations discarded as burn-in and every fourth of the remaining samples kept for inference. In contrast to the simulated examples in Section 4, we need a longer run length and a thinning factor of 4 to produce a Markov chain that has a satisfactory mixing.

The computation time for the MCMC fitting is around 1186.02 seconds, compared to 11.31 seconds for the hybrid MFVB and 10.46 seconds for the INFVB. Figure 6 compares the approximate density functions from the hybrid MFVB against the MCMC results. From Figure 6, we can see that the hybrid MFVB captures the posterior means of all model parameters well but underestimates posterior variances for β_0, ρ, and λ. In addition, the

accuracy scores for regression coefficients $\boldsymbol{\beta}$ range between 70% and 99%. However, the accuracy scores for spatial coefficients ρ and λ are as low as 13% and 58%, respectively. For the variance parameter σ^2, the accuracy score is 96% (plot not shown). Accordingly, the accuracy scores for the hybrid MFVB fitting suggest that better approximate posterior density distributions are desired. To help understand the poor performance of the hybrid MFVB fitting in capturing posterior variances, we use MCMC samples to compute pairwise Spearman's rank correlation coefficients for parameters in the SAC model and provide the results in the Supplementary Appendix C. Spearman's correlation coefficients are -0.82 between ρ and β_0, 0.67 between λ and β_0, and -0.87 between ρ and λ; these coefficients are significant at the 5% level. Given such a strong dependence between these three pairs of parameters, it is not surprising that the underlying assumption of posterior independence for the hybrid MFVB is fragile and the optimal variational distribution functions for β_0, ρ, and λ from the hybrid MFVB are poor approximates to the corresponding exact posterior distributions.

Figure 7 compares the INFVB-based approximate density functions against the MCMC result. From Figure 7, we can see that the accuracy of INFVB is excellent with scores ranging between 92% and 99%. In terms of posterior variances, there is a close agreement between the INFVB result and its MCMC counterpart. This is expected because the INFVB algorithm is more robust to the violation of the posterior independence assumption than its hybrid MFVB counterpart is. Computationally, the INFVB fitting takes less time than the hybrid MFVB fitting, partially because parallelization is accomplished in the former case.

5.2 The MESS(1,1) Model

Similar to Section 5.1, we consider the following MESS(1,1) model:

$$\begin{aligned}\mathbf{S}_1\boldsymbol{y} &= \rho\mathbf{W}_1\boldsymbol{y} + \beta_0 + \beta_1\texttt{CRIM} + \beta_2\texttt{ZN} + \beta_3\texttt{INDUS} + \beta_4\texttt{CHAS} \\ &\quad + \beta_5\texttt{NOX}^2 + \beta_6\texttt{RM}^2 + \beta_7\texttt{AGE} + \beta_8\ln(\texttt{DIS}) + \beta_9\ln(\texttt{RAD}) \\ &\quad + \beta_{10}\texttt{TAX} + \beta_{11}\texttt{PTRATIO} + \beta_{12}\texttt{B} + \beta_{13}\ln(\texttt{LSTAT}) + \boldsymbol{u} \\ \mathbf{S}_2\boldsymbol{u} &= \lambda\mathbf{W}_1\boldsymbol{u} + \boldsymbol{\epsilon}\end{aligned},$$

where $\mathbf{S}_1 = e^{\alpha\mathbf{W}_1}$ and $\mathbf{S}_2 = e^{\tau\mathbf{W}_1}$.

For prior distributions, we consider the same specification as in Section 4.2. Regarding the MCMC fitting of the MESS(1,1), we adopt the same specification as in Section 5.1. For the hybrid MFVB fitting, we set $\varepsilon = 10^{-4}$ because of both computational and convergence concerns. However, with $\varepsilon = 10^{-4}$, the hybrid MFVB algorithm stops short and leads to extremely poor accuracy. Consequently, the hybrid MFVB results are not presented here (they are available upon request). For the INFVB fitting, we place 80 grid points for α and τ, all of which are equally spaced between -0.8 and 0.0 (inclusive).

The computation time for the MCMC fitting is around 21.78 hours, compared to about 2.11 minutes for the INFVB fitting. Figure 8 compares the optimal variational density functions from the INFVB fitting against its MCMC counterparts. The accuracy scores for $\boldsymbol{\beta}$, α, and τ in Figure 8 range between 97% and 99%. For σ^2, the accuracy score is 98% (plot not shown). Hence we can conclude that the accuracy of the INFVB is extremely satisfactory. Supplemental Appendix C shows the pairwise Spearman's rank correlation coefficient between two variables in the model, which are calculated based on the MCMC samples. According to these coefficients, there is a strong monotonic relationship between α and β_0 (Spearman coefficient = 0.74), between α and τ (Spearman coefficient = -0.73), and between τ and β_0 (Spearman coefficient = -0.52), all of which are significant at the 5% level.

6 Discussion

The spatial autoregressive confused model and the matrix exponential spatial specification model are two widely used spatial econometric models for Gaussian areal data. In this paper, we consider variational Bayesian inference for these two models by using the hybrid MFVB and INFVB methods. As a deterministic approach, variational Bayes methods find a variational density function to optimally approximate the joint posterior density function, which is achieved by solving an optimization problem that minimizes the Kullback-Leibler divergence between the joint posterior density function and the candidate variational density functions.

The MFVB and INFVB methods are computationally faster than the MCMC method. The fundamental difference between the MFVB and INFVB methods is that the MFVB method assumes that the joint posterior distribution can be factorized into a product form, whereas the INFVB method is free of such an assumption. As we have shown through simulated examples and a real-data application, both the hybrid MFVB method and the INFVB method can yield very accurate estimates for posterior means of model parameters. However, the hybrid MFVB method is prone to underestimate posterior variances when posterior independence between different blocks of model parameters breaks down. Unlike the hybrid MFVB method, the INFVB method can capture both posterior means and variances very well and is more robust to violation of the posterior dependence assumption. Moreover, the INFVB algorithm can be parallelized and scaled, making it extremely useful for handling big data.

Despite the ability to parallelize and scale the INFVB method, its computation time depends on the dimension of $\boldsymbol{\Theta}_d$ and the choice of grid points that are used for discretization. To save computation time, it is important to keep the dimension of $\boldsymbol{\Theta}_d$ low and to conctrol the number of grid points while the grid spans the domain of the marginal posterior $p(\boldsymbol{\Theta}_d|\boldsymbol{y})$.

When the hybrid MFVB results are easy to obtain, we can appeal to the results to guide our choice of grid points in the INFVB algorithm by accounting for potential underestimation of posterior variances. For example, consider a parameter $\theta \in \boldsymbol{\Theta}_d$ with posterior mean μ_θ and posterior variance σ_θ from the hybrid MFVB fitting, we can specify the grid points to be spaced between $\max(\mu_\theta - 10\sigma_\theta, \min(\theta))$ and $\min(\mu_\theta + 10\sigma_\theta, \max(\theta))$ for the INFVB algorithm. Alternatively, we can use sensitivity analysis to facilitate our choice of the grid points for the INFVB fitting.

We now discuss how to use the hybrid MFVB and INFVB methods to make joint inference. Because of the product-form restriction, the hybrid MFVB method provides insufficient information about the covariance between model parameters, which prevents us from making joint inference. In contrast, the INFVB method enables us to compute the covariance matrix of model parameters (approximately) and to make joint inference. When analytically possible, one can compute the covariance between model parameters based on the approximate joint posterior distribution, $q^*(\boldsymbol{\Theta}_c, \boldsymbol{\Theta}_d) = q^*(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d)q^*(\boldsymbol{\Theta}_d)$, and on the two approximate distributions $q^*(\boldsymbol{\Theta}_c)$ and $q^*(\boldsymbol{\Theta}_d)$. Alternatively, one can realize joint inference using the INFVB method through a simulation process that is naturally parallelizable. To begin, a sample $\boldsymbol{\theta}_d$ is drawn from the optimal variational distribution $q^*(\boldsymbol{\Theta}_d)$. Conditioning on this $\boldsymbol{\theta}_d$, a sample for $\boldsymbol{\Theta}_c$ can be drawn from $q^*(\boldsymbol{\Theta}_c|\boldsymbol{\Theta}_d = \boldsymbol{\theta}_d)$. This simulation process is repeated until sufficient samples are collected. Approximately, these collected samples are also samples from the true joint posterior distribution (since $p(\boldsymbol{\Theta}_c, \boldsymbol{\Theta}_d|, \boldsymbol{y}) \approx q^*(\boldsymbol{\Theta}_c, \boldsymbol{\Theta}_d)$) and hence they can be used for summarizing the covariance matrix and for joint inference, in the same way that the MCMC samples are used.

Although we restrict our attention to spatial econometric models for Gaussian data, the hybrid MFVB method and the INFVB method we propose can be extended to other Gaussian and non-Gaussian spatial econometric models, such as spatial panel models and spatial nonlinear models. As an example, our hybrid MFVB and INFVB methods are applicable

to spatial probit models for both binary and ordered multiple-choice outcomes. Yet there are some computational aspects that are model-specific and need to be resolved. Given that Bayesian estimation for nonlinear spatial econometric models is more involved, we believe that variational Bayesian inference can play a bigger role. However, we do not pursue such a direction here because it is the subject of ongoing research.

Acknowledgements

We thank Xuejun Liao for helpful discussion on the INFVB algorithm and Anne Baxter for editing the paper. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Anselin, L. (2001). "Spatial Econometrics." In *A Companion to Theoretical Econometrics*, ed. B. Baltagi. Oxford, UK: Blackwell.

Barry, R. P. and Pace, R. K. (1999). "Monte Carlo Estimates of the Log Determinant of Large Sparse Matrices." *Linear Algebra and Its Applications*, 289, 1–3, 41–54.

Bivand, R., Gómez-Rubio, V., and Rue, H. (2015). "Spatial Data Analysis with R-INLA with Some Extensions." *Journal of Statistical Software*, 63, 20, 1–31.

Bivand, R. S., Gómez-Rubio, V., and Rue, H. (2014). "Approximate Bayesian Inference for Spatial Econometrics Models." *Spatial Statistics*, 9, 146–165.

Bivand, R. S. and Piras, G. (2015). "Comparing Implementations of Estimation Methods for Spatial Econometrics." *Journal of Statistical Software*, 63, 18, 1–36.

Blei, D. M. and Jordan, M. I. (2006). "Variational Inference for Dirichlet Process Mixtures." *Bayesian Analysis*, 1, 1, 121–143.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). "Variational Inference: A Review for Statisticians." *Journal of the American Statistical Association*, 112, 518, 859–877.

Debarsy, N., Jin, F., and Lee, L.-F. (2015). "Large Sample Properties of the Matrix Exponential Spatial Specification with an Application to FDI." *Journal of Econometrics*, 188, 1, 1–21.

Elhorst, J. P. (2013). *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. New York, NY: Springer.

Figueiredo, C. and Da Silva, A. R. (2015). "A Matrix Exponential Spatial Specification Approach to Panel Data Models." *Empirical Economics*, 49, 1, 115–129.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Boca Raton, FL: CRC Press.

Gómez-Rubio, V., Bivand, R. S., and Rue, H. (2017). "Estimating Spatial Econometrics Models with Integrated Nested Laplace Approximation." *arXiv preprint arXiv:1703.01273*.

Gómez-Rubio, V. and Palmí-Perales, F. (2017). "Spatial Models with the Integrated Nested Laplace Approximation within Markov Chain Monte Carlo."

Goulet, V., Dutang, C., Maechler, M., Firth, D., Shapira, M., and Stadelmann, M. (2017). *expm: Matrix Exponential, Log, 'etc'*. R package version 0.999-2.

Han, S., Liao, X., and Carin, L. (2013). "Integrated Non-Factorized Variational Inference." In *Advances in Neural Information Processing Systems*, 2481–2489.

Harrison, D. and Rubinfeld, D. L. (1978). "Hedonic Housing Prices and the Demand for Clean Air." *Journal of Environmental Economics and Management*, 5, 1, 81–102.

Honkela, A., Raiko, T., Kuusela, M., Tornio, M., and Karhunen, J. (2010). "Approximate Riemannian Conjugate Gradient Learning for Fixed-Form Variational Bayes." *Journal of Machine Learning Research*, 11, 3235–3268.

Horn, R. A. and Johnson, C. R. (1990). *Matrix Analysis*. Cambridge, UK: Cambridge University Press.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). "An introduction to variational methods for graphical models." *Machine learning*, 37, 2, 183–233.

LeSage, J. P. and Pace, R. K. (2007). "A Matrix Exponential Spatial Specification." *Journal of Econometrics*, 140, 1, 190–214.

— (2009). *Introduction to Spatial Econometrics*. Boca Raton, FL: CRC Press.

Manski, C. F. (1993). "Identification of Endogenous Social Effects: The Reflection Problem." *The Review of Economic Studies*, 60, 3, 531–542.

Microsoft Corporation (2017). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R Package Version 1.0.11.

Murray, I., Adams, R., and MacKay, D. (2010). "Elliptical Slice Sampling." In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 9, 541–548.

Ormerod, J. T. and Wand, M. P. (2010). "Explaining Variational Approximations." *American Statistician*, 64, 2, 140–153.

Pace, R. K. and LeSage, J. P. (2004). "Chebyshev Approximation of Log-Determinants of Spatial Weight Matrices." *Computational Statistics & Data Analysis*, 45, 2, 179–196.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Revolution Analytics (2015). *foreach: Provides Foreach Looping Construct for R*. R Package Version 1.4.3.

Robert, C. P. (2004). *Monte Carlo Statistical Methods*. 2nd ed. New York, NY: Springer.

Rodrigues, E., Assunção, R., and Dey, D. K. (2014). "A Closer Look at the Spatial Exponential Matrix Specification." *Spatial Statistics*, 9, 109–121.

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 2, 319–392.

Salimans, T., Knowles, D. A., et al. (2013). "Fixed-Form Variational Posterior Approximation through Stochastic Linear Regression." *Bayesian Analysis*, 8, 4, 837–882.

SAS Institute Inc. (2016). *SAS/ETS 14.2 User's Guide*. SAS Institute Inc., Cary, NC.

— (2017). *SAS® Econometrics 8.2: Econometrics Procedures*. SAS Institute Inc., Cary, NC.

Sidje, R. B. (1998). "Expokit: A Software Package for Computing Matrix Exponentials." *ACM Transactions on Mathematical Software*, 24, 1, 130–156.

Strauß, M. E., Mezzetti, M., and Leorato, S. (2017). "Is a Matrix Exponential Specification Suitable for the Modeling of Spatial Correlation Structures?" *Spatial Statistics*, 20, 221–243.

Tran, M.-N., Nott, D. J., Kuk, A. Y., and Kohn, R. (2016). "Parallel Variational Bayes for Large Datasets with An Application to Generalized Linear Mixed Models." *Journal of Computational and Graphical Statistics*, 25, 2, 626–646.

Wainwright, M. J. and Jordan, M. I. (2008). "Graphical Models, Exponential Families, and Variational Inference." *Foundations and Trends in Machine Learning*, 1, 1–2, 1–305.

Wand, M. P., Ormerod, J. T., Padoan, S. A., and Frühwirth, R. (2011). "Mean Field Variational Bayes for Elaborate Distributions." *Bayesian Analysis*, 6, 4, 847–900.

Wang, B. and Titterington, D. (2005). "Inadequacy of Interval Estimates Corresponding to Variational Bayesian Approximations." In *AISTATS 2005 Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, eds. R. Cowell and Z. Ghahramani, 373–380.

Wang, C. and Blei, D. M. (2013). "Variational Inference in Nonconjugate Models." *Journal of Machine Learning Research*, 14, 1, 1005–1031.

Whittle, P. (1954). "On Stationary Processes in the Plane." *Biometrika*, 434–449.

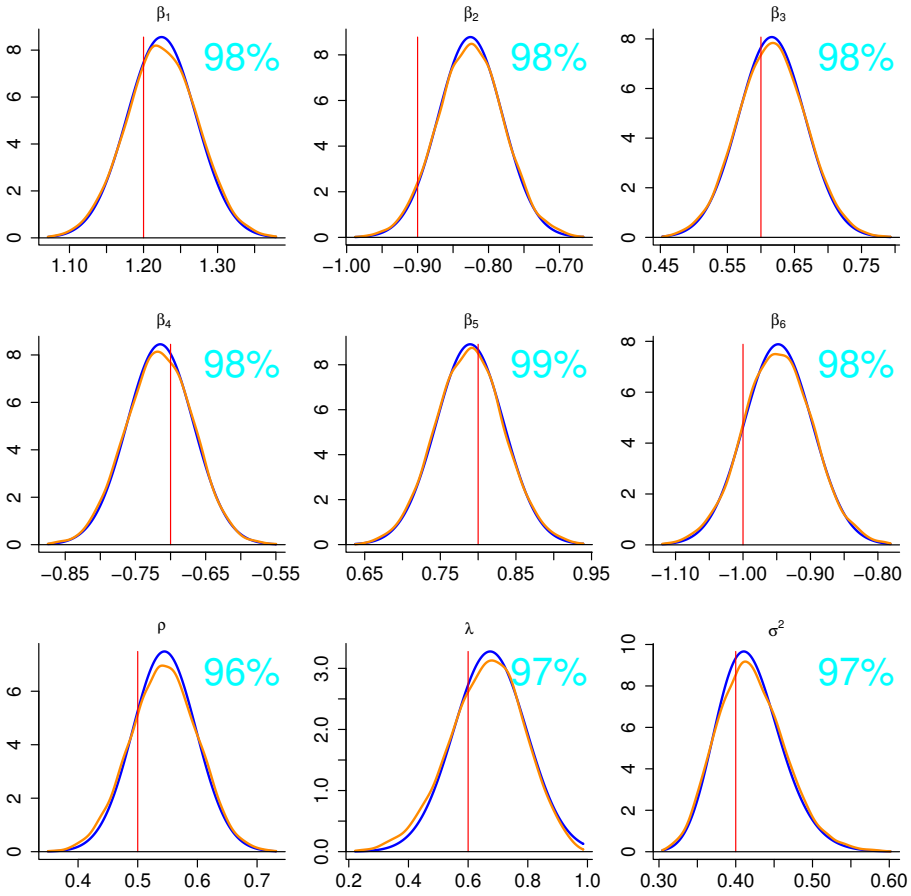

Figure 2: Comparison of the MCMC result (orange) and approximate posterior density functions (blue) using the hybrid MFVB approach for the SAC model in the simulated example (Section 4.1). Vertical lines indicate the true values.

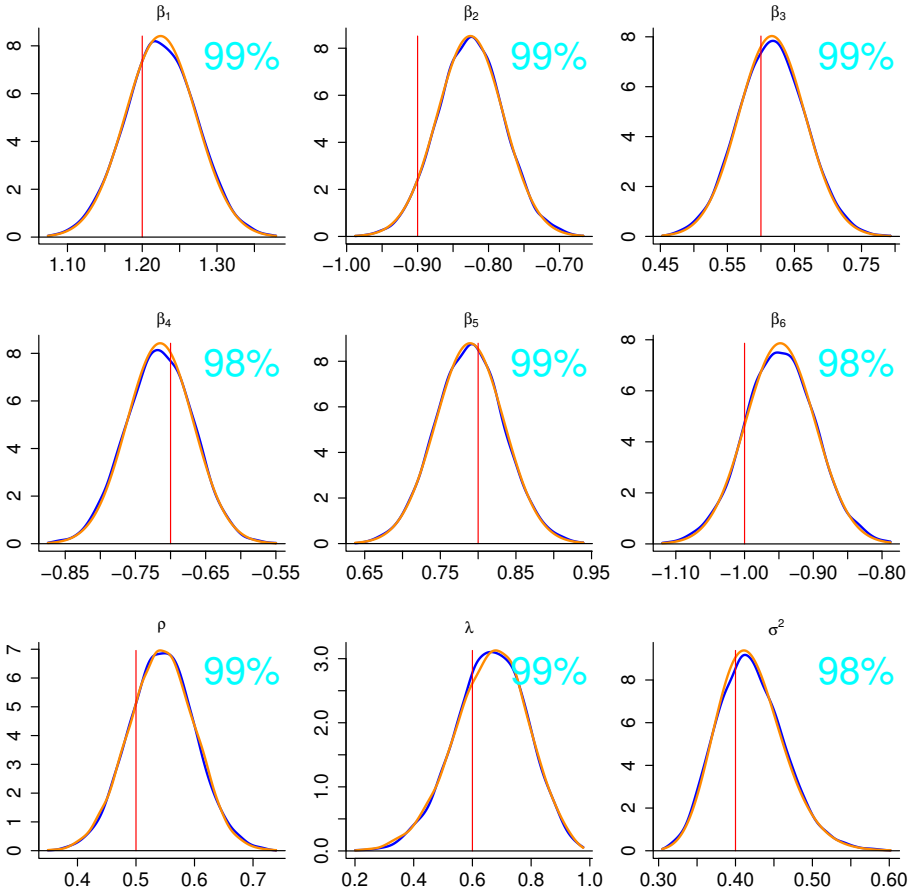

Figure 3: Comparison of the MCMC result (orange) and the INFVB-based approximate posterior density functions (blue) for the SAC model in the simulated example (Section 4.1). Vertical lines indicate the true values.

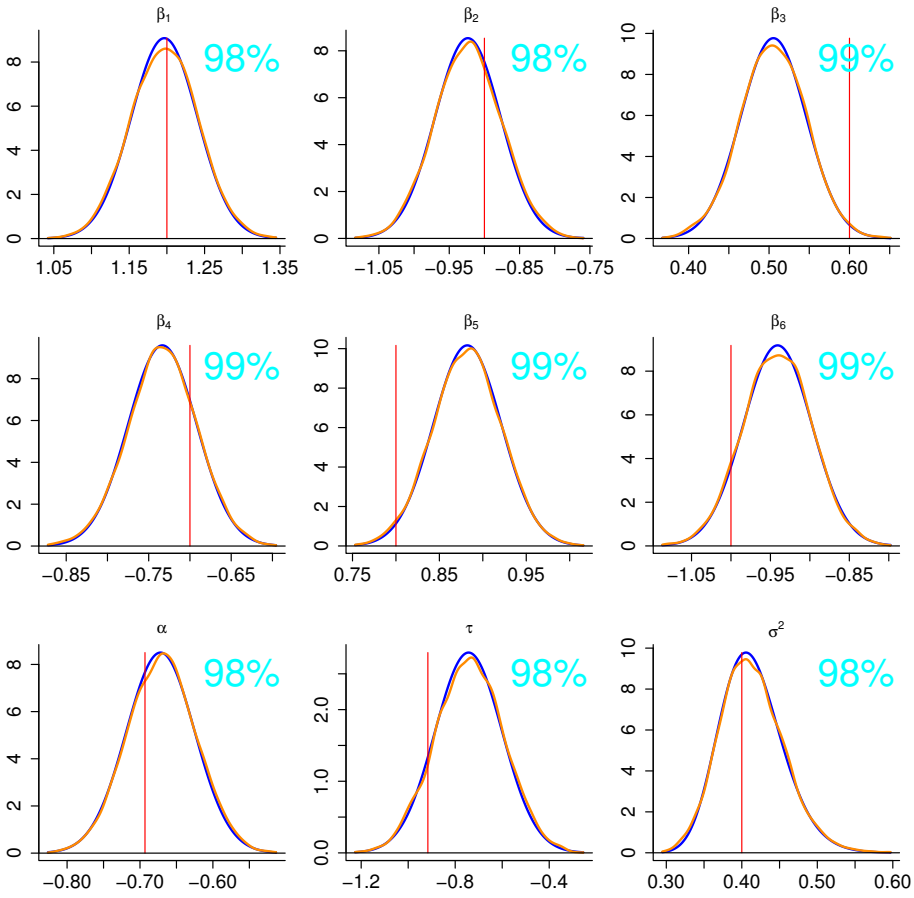

Figure 4: Comparison of the MCMC result (orange) and approximate posterior density functions using the hybrid MFVB approach (blue) for the MESS(1,1) model in the simulated example (Section 4.2). Vertical lines indicate the true values.

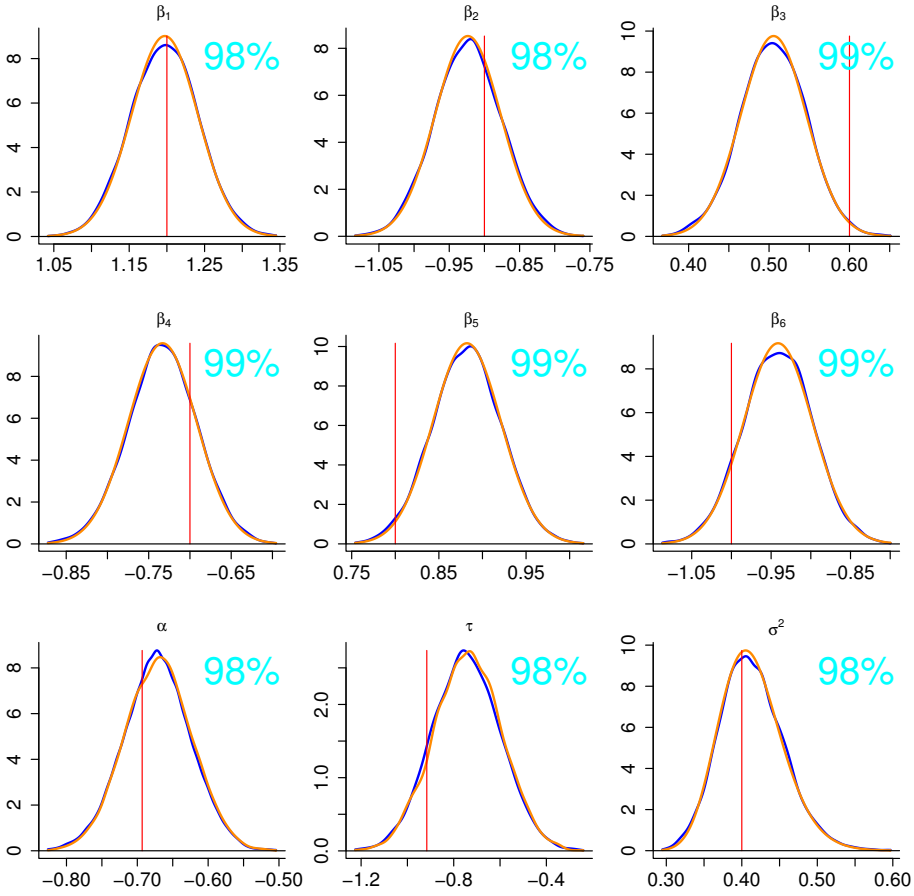

Figure 5: Comparison of the MCMC result (orange) and the INFVB-based approximate posterior density functions (blue) for the SAC model in the simulated example (Section 4.2). Vertical lines indicate the true values.

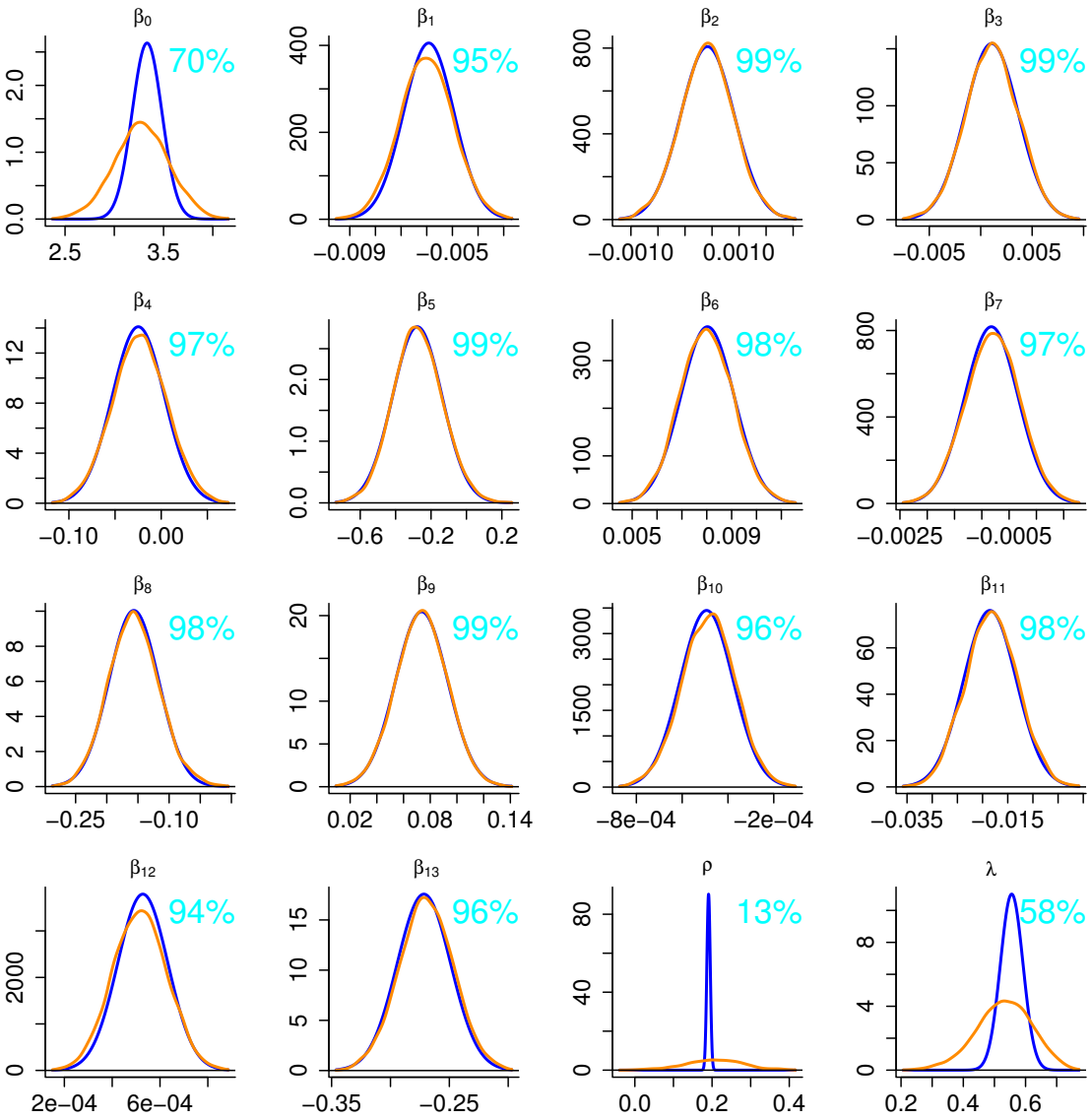

Figure 6: Comparison of the MCMC result (orange) and approximate posterior density functions (blue) from the hybrid MFVB approach for the SAC model using the Boston housing data (Section 5).

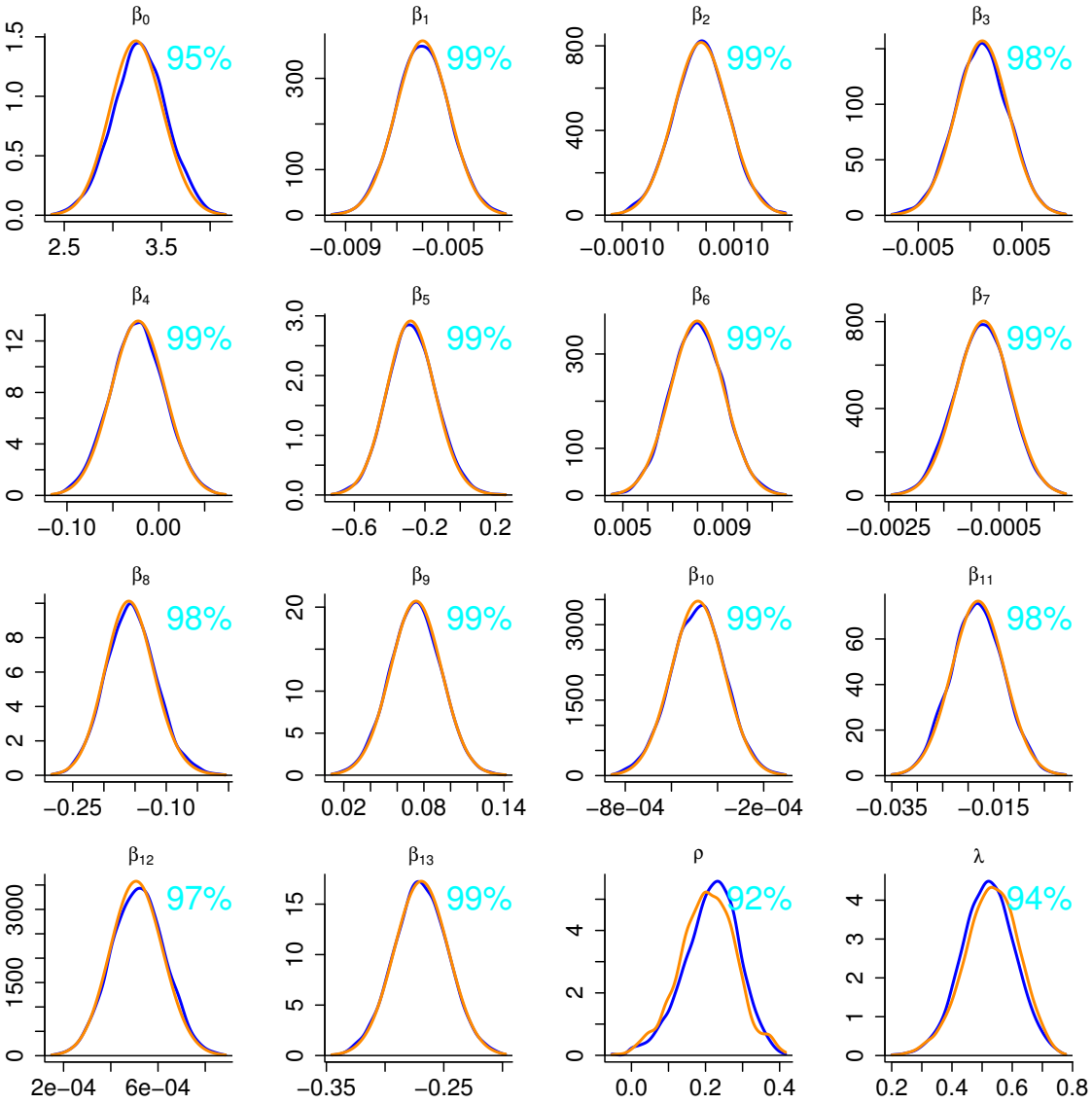

Figure 7: Comparison of the MCMC result (orange) and the INFVB-based approximate posterior density functions (blue) for the SAC model using the Boston housing data (Section 5).

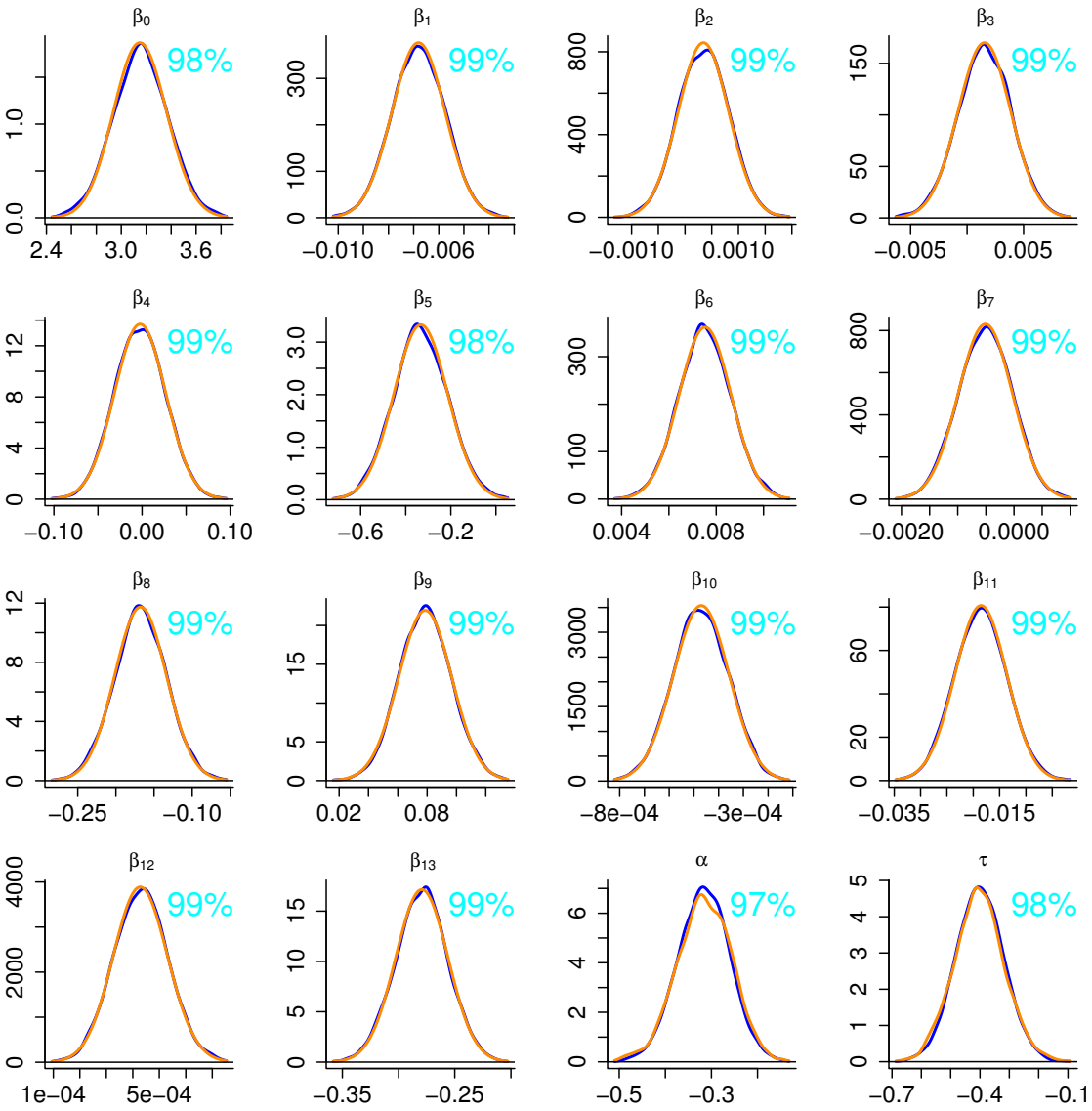

Figure 8: Comparison of the MCMC result (orange) and the INFVB-based approximate posterior density functions (blue) for the MESS(1,1) model using the Boston housing data (Section 5).