Statistics and Computing (2020) 30:1255–1272
https://doi.org/10.1007/s11222-020-09944-8

Conditionally structured variational Gaussian approximation with importance weights

Linda S. L. Tan[1] · **Aishwarya Bhaskaran**[1] · **David J. Nott**[2]

Received: 21 April 2019 / Accepted: 16 April 2020 / Published online: 28 April 2020

Abstract
We develop flexible methods of deriving variational inference for models with complex latent variable structure. By splitting the variables in these models into "global" parameters and "local" latent variables, we define a class of variational approximations that exploit this partitioning and go beyond Gaussian variational approximation. This approximation is motivated by the fact that in many hierarchical models, there are global variance parameters which determine the scale of local latent variables in their posterior conditional on the global parameters. We also consider parsimonious parametrizations by using conditional independence structure and improved estimation of the log marginal likelihood and variational density using importance weights. These methods are shown to improve significantly on Gaussian variational approximation methods for a similar computational cost. Application of the methodology is illustrated using generalized linear mixed models and state space models.

Keywords Gaussian variational approximation · Sparse precision matrix · Stochastic variational inference · Importance weighted lower bound · Rényi's divergence

1 Introduction

In many modern statistical applications, it is necessary to model complex dependent data. In these situations, models which employ observation specific latent variables such as random effects and state space models are widely used because of their flexibility, and Bayesian approaches dealing naturally with the hierarchical structure are attractive in principle. However, incorporating observation specific latent variables leads to a parameter dimension increasing with sample size, and standard Bayesian computational methods can be challenging to implement in very high-dimensional settings. For this reason, approximate inference methods are attractive for these models, both in exploratory settings where many models need to be fitted quickly, as well as in applications involving large datasets where exact methods are infeasible. One of the most common approximate inference paradigms is variational inference (Ormerod and Wand 2010; Blei et al. 2017), which is the approach considered here.

Our main contribution is to consider partitioning the unknowns in a local latent variable model into "global" parameters and "local" latent variables and to suggest ways of structuring the dependence in a variational approximation that match the specification of these models. We go beyond standard Gaussian approximations by defining the variational approximation sequentially, through a marginal density for the global parameters and a conditional density for local parameters given global parameters. Each term in our approximation is Gaussian, but we allow the conditional covariance matrix for the local parameters to depend on the global parameters, which leads to an approximation that is not jointly Gaussian. We are particularly interested in improved inference on global variance and dependence parameters which determine the scale and dependence struc-

Linda Tan and Aishwarya Bhaskaran are supported by the start-up Grant R-155-000-190-133.

✉ Linda S. L. Tan
statsll@nus.edu.sg

Aishwarya Bhaskaran
a0127385@u.nus.edu

David J. Nott
standj@nus.edu.sg

[1] Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore

[2] Department of Statistics and Applied Probability and Institute of Operations Research and Analtyics, National University of Singapore, Singapore, Singapore

ture of local latent variables. With this objective, we suggest a parametrization of our conditional approximation to the local variables that is well motivated and respects the exact conditional independence structure in the true posterior distribution. Our approximations are parsimonious in terms of the number of required variational parameters, which is important since a high-dimensional variational optimization is computationally burdensome. The methods suggested improve on Gaussian variational approximation methods for a similar computational cost. Besides defining a novel and useful variational family appropriate to local latent variable models, we also employ importance weighted variational inference methods (Burda et al. 2016; Domke and Sheldon 2018) to further improve the quality of inference and elaborate further on the connections between this approach and the use of Rényi's divergence within the variational optimization (Li and Turner 2016; Regli and Silva 2018; Yang et al. 2019).

Our method is a contribution to the literature on the development of flexible variational families, and there are many interesting existing methods for this task. One fruitful approach is based on normalizing flows (Rezende and Mohamed 2015), where a variational family is defined using an invertible transformation of a random vector with some known density function. To be useful, the transformation should have an easily computed Jacobian determinant. In the original work of Rezende and Mohamed (2015), compositions of simple flows called radial and planar flows were considered. Later authors have suggested alternatives, such as autoregressive flows (Germain et al. 2015), inverse autoregressive flows (Kingma et al. 2016), and real-valued non-volume preserving transformations (Dinh et al. 2017), among others. Spantini et al. (2018) gives a theoretical framework connecting Markov properties of a target posterior distribution to representations involving transport maps, with normalizing flows being one way to parametrize such mappings. The variational family we consider here can be thought of as a simple autoregressive flow, but carefully constructed to preserve the conditional independence structure in the true posterior and to achieve parsimony in the representation of dependence between local latent variables and global scale parameters. Our work is also related to the hierarchically structured approximations considered in Salimans and Knowles (2013, Sect. 7.1); these authors also consider other flexible approximations based on mixture models and a variety of innovative numerical approaches to the variational optimization. Hoffman and Blei (2015) propose an approach called structured stochastic variational inference which is applicable in conditionally conjugate models. Their approach is similar to ours, in the sense that local variables depend on global variables in the variational posterior. However, conditional conjugacy does not hold in the examples we consider.

The methods we describe can be thought of as extending the Gaussian variational approximation (GVA) of Tan and Nott (2018), where parametrization of the variational covariance matrix was considered in terms of a sparse Cholesky factor of the precision matrix. Similar approximations have been considered for state space models in Archer et al. (2016). The sparse structure reduces the number of free variational parameters and allows matching the exact conditional independence structure in the true posterior. Tan (2018) propose an approach called reparametrized variational Bayes, where the model is reparametrized by applying an invertible affine transformation to the local variables to minimize their posterior dependency on global variables, before applying a mean field approximation. The affine transformation is obtained by considering a second order Taylor series approximation to the posterior of the local variables conditional on the global variables. One way of improving on Gaussian approximations is to consider mixtures of Gaussians (Jaakkola and Jordan 1998; Salimans and Knowles 2013; Miller et al. 2016; Guo et al. 2016). However, even with a parsimonious parametrization of component densities, a large number of additional variational parameters are added with each mixture component. Other flexible variational families can be formed using copulas (Tran et al. 2015; Han et al. 2016; Smith et al. 2019), hierarchical variational models (Ranganath et al. 2016) or implicit approaches (Huszár 2017).

A closely related work is the framework for hierarchical dynamical systems (HDS) proposed by Roeder et al. (2019) which appeared shortly after this manuscript. To allow their methods to scale up to massive data sets, Roeder et al. (2019) consider amortized variational inference for HDS, which similarly uses doubly reparameterized importance-weighted gradient estimators. A Gaussian mean-field variational posterior was assumed for the latent variables, which were partitioned into three blocks, population, group and individual. The mean and variance parameters of group-level variables were expressed as functions of group indicators, while that of individual-level variables were encoded using neural networks depending on both observations and group indicators. While our parameters are also classified as either local or global, we do not apply the mean-field assumption or amortized variational inference. Instead, we consider conditionally structured variational posteriors which exploit the conditional independence structure of the true posterior and express the local variational parameters as linear functions of the global ones to capture the strong posterior dependencies between local and global variables.

We specify the model and notation in Sect. 2 and introduce the conditionally structured Gaussian variational approximation (CSGVA) in Sect. 3. The algorithm for optimizing the variational parameters is described in Sect. 4, and Sect. 5 highlights the association between GVA and CSGVA.

Section 6 describes how CSGVA can be improved using importance weighting. Experimental results and applications to generalized linear mixed models (GLMMs) and state space models are presented in Sects. 7, 8 and 9, respectively. Section 10 gives some concluding discussion.

2 Model specification and notation

Let $y = (y_1, \ldots, y_n)^T$ be observations from a model with global variables θ_G and local variables $\theta_L = (b_1, \ldots, b_n)^T$, where b_i contains latent variables specific to y_i for $i = 1, \ldots, n$. Suppose θ_G is a vector of length G and each b_i is a vector of length L. Let $\theta = (\theta_G^T, \theta_L^T)^T$. We consider models where the joint density is of the form

$$p(y, \theta) = p(\theta_G)p(b_1, \ldots, b_\ell|\theta_G)\left\{\prod_{i=1}^{n} p(y_i|b_i, \theta_G)\right\} \times \left\{\prod_{i>\ell} p(b_i|b_{i-1}, \ldots, b_{i-\ell}, \theta_G)\right\}.$$

The observations $\{y_i\}$ are conditionally independent given $\{b_i\}$ and θ_G. Conditional on θ_G, the local variables $\{b_i\}$ form a ℓth order Markov chain if $\ell \geq 1$, and they are conditionally independent if $\ell = 0$. This class of models include important models such as GLMMs and state space models. Next, we define some mathematical notation before discussing CSGVA for this class of models.

2.1 Notation

For an $r \times r$ matrix A, let $\text{diag}(A)$ denote the diagonal elements of A and $\text{dg}(A)$ be the diagonal matrix obtained by setting non-diagonal elements in A to zero. Let $\text{vec}(A)$ be the vector of length r^2 obtained by stacking the columns of A under each other from left to right and $v(A)$ be the vector of length $r(r+1)/2$ obtained from $\text{vec}(A)$ by eliminating all superdiagonal elements of A. Let E_r be the $r(r+1)/2 \times r^2$ elimination matrix, K_r be the $r^2 \times r^2$ commutation matrix and D_r be the $r^2 \times r(r+1)/2$ duplication matrix (see Magnus and Neudecker 1980). Then $K_r\text{vec}(A) = \text{vec}(A^T)$, $E_r\text{vec}(A) = v(A)$, $E_r^T v(A) = \text{vec}(A)$ if A is lower triangular, and $D_r v(A) = \text{vec}(A)$ if A is symmetric. Let $\mathbf{1}_r$ be a vector of ones of length r. Scalar functions applied to vector arguments are evaluated element by element. Let d denote the differential operator (see e.g. Magnus and Neudecker 1999).

3 Conditionally structured Gaussian variational approximation

We propose to approximate the posterior distribution $p(\theta|y)$ of the model defined in Sect. 2 by a density of the form

$$q(\theta) = q(\theta_G)q(\theta_L|\theta_G),$$

where $q(\theta_G) = N(\mu_1, \Omega_1^{-1})$, $q(\theta_L|\theta_G) = N(\mu_2, \Omega_2^{-1})$, and Ω_1 and Ω_2 are the precision (inverse covariance) matrices of $q(\theta_G)$ and $q(\theta_L|\theta_G)$, respectively. Here μ_2 and Ω_2 depend on θ_G, but we do not denote this explicitly for notational conciseness. Let $C_1C_1^T$ and $C_2C_2^T$ be unique Cholesky factorizations of Ω_1 and Ω_2, respectively, where C_1 and C_2 are lower triangular matrices with positive diagonal entries. We further define C_1^* and C_2^* to be lower triangular matrices of order G and nL, respectively, such that $C^*_{r,ii} = \log(C_{r,ii})$ and $C^*_{r,ij} = C_{r,ij}$ if $i \neq j$ for $r = 1, 2$. The purpose of introducing C_1^* and C_2^* is to allow unconstrained optimization of the variational parameters in the stochastic gradient ascent algorithm, since diagonal entries of C_1 and C_2 are constrained to be positive. Note that C_2 and C_2^* also depend on θ_G but again we do not show this explicitly in our notation.

As Ω_2 depends on θ_G, the joint distribution $q(\theta)$ is generally non-Gaussian even though $q(\theta_G)$ and $q(\theta_L|\theta_G)$ are individually Gaussian. Here we consider a first order approximation and assume that μ_2 and $v(C_2^*)$ are linear functions of θ_G:

$$\mu_2 = d + C_2^{-T}D(\mu_1 - \theta_G), \quad v(C_2^*) = f + F\theta_G. \tag{1}$$

In (1), d is a vector of length nL, D is a $nL \times G$ matrix, f is a vector of length $nL(nL+1)/2$ and F is a $nL(nL+1)/2 \times G$ matrix. For this specification, $q(\theta)$ is not jointly Gaussian due to dependence of the covariance matrix of $q(\theta_L|\theta_G)$ on θ_G. It is Gaussian if and only if $F \equiv 0$. The set of variational parameters to be optimized is denoted as

$$\lambda = (\mu_1^T, v(C_1^*)^T, d^T, \text{vec}(D)^T, f^T, \text{vec}(F)^T)^T.$$

As motivation for the linear approximation in (1), consider the linear mixed model,

$$y_i = X_i\beta + Z_i b_i + \epsilon_i, \quad (i = 1, \ldots, n),$$
$$\beta \sim N(0, \sigma_\beta^2 I_p), \quad b_i \sim N(0, \Lambda), \quad \epsilon_i \sim N(0, \sigma_\epsilon^2 I_{n_i}),$$

where y_i is a vector of responses of length n_i for the ith subject, X_i and Z_i are covariate matrices of dimensions $n_i \times p$ and $n_i \times L$, respectively, β is a vector of coefficients of length p and $\{b_i\}$ are random effects. Assume σ_ϵ^2 is known. Then the global parameters θ_G consists of β and Λ. The posterior

of θ_L conditional on θ_G is

$$p(\theta_L|y,\theta_G) \propto \prod_{i=1}^{n} p(y_i|\beta,b_i)p(b_i|\Lambda)$$
$$\propto \prod_{i=1}^{n} \exp[-\{b_i^T(Z_i^T Z_i/\sigma_\epsilon^2+\Lambda^{-1})b_i - 2b_i^T Z_i^T(y_i - X_i\beta)/\sigma_\epsilon^2\}/2].$$

Thus $p(\theta_L|y,\theta_G) = \prod_{i=1}^n p(b_i|y,\theta_G)$, where $p(b_i|y,\theta_G)$ is a normal density with precision matrix $Z_i^T Z_i/\sigma_\epsilon^2 + \Lambda^{-1}$ and mean $(Z_i^T Z_i/\sigma_\epsilon^2+\Lambda^{-1})^{-1}Z_i^T(y_i-X_i\beta)/\sigma_\epsilon^2$. The precision matrix depends on Λ^{-1} linearly and the mean depends on β linearly after scaling by the covariance matrix. The linear approximation in (1) tries to mimic this dependence relationship.

The proposed variational density is conditionally structured and highly flexible. Such dependence structure is particularly valuable in constructing variational approximations for hierarchical models, where there are global scale parameters in θ_G which help to determine the scale of local latent variables in the conditional posterior of $\theta_L|\theta_G$. While marginal posteriors of the global variables are often well approximated by Gaussian densities, marginal posteriors of the local variables tend to exhibit more skewness and kurtosis. This deviation from normality can be captured by $q(\theta_L) = \int q(\theta_G)q(\theta_L|\theta_G)d\theta_G$, which is a mixture of normal densities. The formulation in (1) also allows for a reduction in the number of variational parameters if conditional independence structure consistent with that in the true posterior is imposed on the variational approximation.

3.1 Using conditional independence structure

Tan and Nott (2018) incorporate the conditional independence structure of the true posterior into Gaussian variational approximations by using the fact that zeros in the precision matrix correspond to conditional independence for Gaussian random vectors. This incorporation achieves sparsity in the precision matrix of the approximation and leads to a large reduction in the number of variational parameters to be optimized. For high-dimensional θ, this sparse structure is especially important because a full Gaussian approximation involves learning a covariance matrix where the number of elements grows quadratically with the dimension of θ.

Recall that $\theta_L = (b_1,\ldots,b_n)^T$. Suppose b_i is conditionally independent of b_j in the posterior for $|i-j| > \ell$, given θ_G and $\{b_j \mid |i-j| \le \ell\}$. For instance, in a GLMM, $\{b_i\}$ may be subject specific random effects, and these are conditionally independent given the global parameters, so this structure holds with $\ell = 0$. In the case of a state space model for a time series, $\{b_i\}$ are the latent states, and this structure holds with $\ell = 1$. Note that ordering of the latent variables is important here.

Now partition the precision matrix Ω_2 of $q(\theta_L|\theta_G)$ into $L \times L$ blocks with n row and n column partitions corresponding to $\theta_L = (b_1,\ldots,b_n)^T$. Let $\Omega_{2,ij}$ be the block corresponding to b_i horizontally and b_j vertically for $i, j = 1,\ldots,n$. If b_i is conditionally independent of b_j for $|i-j| > \ell$, given θ_G and $\{b_j \mid |i-j| \le \ell\}$, then we set $\Omega_{2,ij} = 0$ for all pairs (i,j) with $|i-j| > \ell$. Let $\mathcal{I}$ denote the indices of elements in $v(\Omega_2)$ which are fixed at zero by this conditional independence requirement. If we choose $f_i = 0$ and $F_{ij} = 0$ for all $i \in \mathcal{I}$ and all j in (1), then C_2^* has the same block sparse structure we desire for the lower triangular part of Ω_2. By Proposition 1 of Rothman et al. (2010), this means that Ω_2 will have the desired block sparse structure. Hence we impose the constraints $f_i = 0$ and $F_{ij} = 0$ for $i \in \mathcal{I}$ and all j, which reduces the number of variational parameters to be optimized.

4 Optimization of variational parameters

To make the dependence on λ explicit, write $q(\theta)$ as $q_\lambda(\theta)$. The variational parameters λ are optimized by minimizing the Kullback-Leibler divergence between $q_\lambda(\theta)$ and the true posterior $p(\theta|y)$, where

$$\mathrm{KL}\{q_\lambda||p(\theta|y)\} = \int q_\lambda(\theta)\log\frac{q_\lambda(\theta)}{p(\theta|y)}d\theta$$
$$= \log p(y) - \int q_\lambda(\theta)\log\frac{p(y,\theta)}{q_\lambda(\theta)}d\theta \ge 0.$$

Minimizing $\mathrm{KL}\{q_\lambda||p(\theta|y)\}$ is therefore equivalent to maximizing an evidence lower bound $\mathcal{L}^{\mathrm{VI}}$ on the log marginal likelihood $\log p(y)$, where

$$\mathcal{L}^{\mathrm{VI}} = E_{q_\lambda}\{\log p(y,\theta) - \log q_\lambda(\theta)\}. \quad (2)$$

In (2), E_{q_λ} denotes expectation with respect to $q_\lambda(\theta)$. We seek to maximize $\mathcal{L}^{\mathrm{VI}}$ with respect to λ using stochastic gradient ascent. Starting with some initial estimate of λ, we perform the following update at each iteration t,

$$\lambda_t = \lambda_{t-1} + \rho_t \widehat{\nabla}_\lambda \mathcal{L}^{\mathrm{VI}},$$

where ρ_t represents a small stepsize taken in the direction of the stochastic gradient $\widehat{\nabla}_\lambda \mathcal{L}^{\mathrm{VI}}$. The sequence $\{\rho_t\}$ should satisfy the conditions $\sum_t \rho_t = \infty$ and $\sum_t \rho_t^2 < \infty$ (Spall 2003).

An unbiased estimate of the gradient $\nabla_\lambda \mathcal{L}^{\mathrm{VI}}$ can be constructed using (2) by simulating θ from $q_\lambda(\theta)$. However, this approach usually results in large fluctuations in the stochastic gradients. Hence we implement the "reparametrization trick"

(Kingma and Welling 2014; Rezende et al. 2014; Titsias and Lázaro-Gredilla 2014), which helps to reduce the variance of the stochastic gradients. This approach writes $\theta \sim q_\lambda(\theta)$ as a function of the variational parameters λ and a random vector s having a density not depending on λ. To explain further, let $s = [s_1^T, s_2^T]^T$, where s_1 and s_2 are vectors of length G and nL corresponding to θ_G and θ_L, respectively. Consider a transformation $\theta = r_\lambda(s)$ of the form

$$\begin{bmatrix}\theta_G \\ \theta_L\end{bmatrix} = \begin{bmatrix}\mu_1 + C_1^{-T} s_1 \\ \mu_2 + C_2^{-T} s_2\end{bmatrix}. \tag{3}$$

Since μ_2 and C_2 are functions of θ_G from (1),

$$\begin{aligned}\mu_2 &= d + C_2^{-T} D(\mu_1 - \theta_G) = d - C_2^{-T} D C_1^{-T} s_1, \\ v(C_2^*) &= f + F\theta_G = f + F(\mu_1 + C_1^{-T} s_1).\end{aligned}$$

Hence μ_2 and C_2 are functions of s_1, and θ_L is a function of both s_1 and s_2. This transformation is invertible since given θ and λ, we can first recover $s_1 = C_1^T(\theta_G - \mu_1)$, find μ_2 and C_2, and then recover $s_2 = C_2^T(\theta_L - \mu_2)$. Applying this transformation,

$$\begin{aligned}\mathcal{L}^{\mathrm{VI}} &= \int \phi(s)\{\log p(y,\theta) - \log q_\lambda(\theta)\} ds \\ &= E_\phi\{\log p(y,\theta) - \log q_\lambda(\theta)\},\end{aligned} \tag{4}$$

where E_ϕ denotes expectation with respect to $\phi(s)$ and $\theta = r_\lambda(s)$.

4.1 Stochastic gradients

Next, we differentiate (4) with respect to λ to find unbiased estimates of the gradients. As $\log q_\lambda(\theta)$ depends on λ directly as well as through θ, applying the chain rule, we have

$$\begin{aligned}\nabla_\lambda \mathcal{L}^{\mathrm{VI}} &= E_\phi[\nabla_\lambda r_\lambda(s)\{\nabla_\theta \log p(y,\theta) - \nabla_\theta \log q_\lambda(\theta)\} \\ &\quad - \nabla_\lambda \log q_\lambda(\theta)] \qquad (5) \\ &= E_\phi[\nabla_\lambda r_\lambda(s)\{\nabla_\theta \log p(y,\theta) - \nabla_\theta \log q_\lambda(\theta)\}]. \qquad (6)\end{aligned}$$

Note that $E_\phi\{\nabla_\lambda \log q_\lambda(\theta)\} = 0$ as it is the expectation of the score function. Roeder et al. (2017) refer to the expressions inside the expectations in (5) and (6) as the *total derivative* and *path derivative*, respectively. In (6), the contributions to the gradient from $\log p(y,\theta)$ and $\log q_\lambda(\theta)$ cancel each other if $q_\lambda(\theta)$ approximates the true posterior well (at convergence). However, the score function $\nabla_\lambda \log q_\lambda(\theta)$ is not necessarily small even if $q_\lambda(\theta)$ is a good approximation to $p(\theta|y)$. This term affects adversely the ability of the algorithm to converge and "stick" to the optimal variational parameters, a phenomenon Roeder et al. (2017) refers to as "sticking the landing". Hence, we consider the path derivative,

$$\widehat{\nabla}_\lambda \mathcal{L}^{\mathrm{VI}} = \nabla_\lambda r_\lambda(s)\{\nabla_\theta \log p(y,\theta) - \nabla_\theta \log q_\lambda(\theta)\} \tag{7}$$

as an unbiased estimate of the true gradient $\nabla_\lambda \mathcal{L}^{\mathrm{VI}}$. Tan and Nott (2018) and Tan (2018) also demonstrate that the path derivative has smaller variation about zero when the algorithm is close to convergence.

Let $\nabla_\theta \log p(y,\theta) - \nabla_\theta \log q_\lambda(\theta) = (\mathcal{G}_1^T, \mathcal{G}_2^T)^T$, where $\mathcal{G}_1$ and $\mathcal{G}_2$ are vectors of length G and nL, respectively, corresponding to the partitioning of $\theta = [\theta_G^T, \theta_L^T]^T$. Then $\widehat{\nabla}_\lambda \mathcal{L}^{\mathrm{VI}} = \nabla_\lambda r_\lambda(s)(\mathcal{G}_1^T, \mathcal{G}_2^T)^T$ is given by

$$\begin{bmatrix} \mathcal{G}_1 + \nabla_{\mu_1}\theta_L \mathcal{G}_2 \\ -D_1^* v[C_1^{-T} s_1 \{\mathcal{G}_1 + (\nabla_{\mu_1}\theta_L - D^T C_2^{-1})\mathcal{G}_2\}^T C_1^{-T}] \\ \mathcal{G}_2 \\ -\mathrm{vec}(C_2^{-1}\mathcal{G}_2 s_1^T C_1^{-1}) \\ \nabla_f \theta_L \mathcal{G}_2 \\ \mathrm{vec}(\nabla_f \theta_L \mathcal{G}_2 \theta_G^T) \end{bmatrix},$$

where

$$\begin{aligned}\nabla_{\mu_1}\theta_L &= F^T \nabla_f \theta_L, \\ \nabla_f \theta_L \mathcal{G}_2 &= -D_2^* v\{C_2^{-T}(s_2 - D C_1^{-T} s_1)\mathcal{G}_2^T C_2^{-T}\}.\end{aligned}$$

Here D_1^* and D_2^* are diagonal matrices of order $G(G+1)/2$ and $nL(nL+1)/2$, respectively, such that $\mathrm{d}v(C_r) = D_r^* \mathrm{d}v(C_r^*)$ for $r = 1, 2$. Formally, $D_1^* = \mathrm{diag}\{v(\mathrm{dg}(C_1) + \mathbf{1}_G \mathbf{1}_G^T - I_G)\}$ and $D_2^* = \mathrm{diag}\{v(\mathrm{dg}(C_2) + \mathbf{1}_{nL}\mathbf{1}_{nL}^T - I_{nL})\}$. The full expression and derivation of $\nabla_\lambda r_\lambda(s)$ are given in "Appendix A". In addition, we show (in "Appendix A") that

$$\begin{aligned}\nabla_\theta \log q_\lambda(\theta) &= \begin{bmatrix}\nabla_{\theta_G} \log q_\lambda(\theta) \\ \nabla_{\theta_L} \log q_\lambda(\theta)\end{bmatrix} \\ &= \begin{bmatrix} F^T[v(I_{nL}) - D_2^* v\{(\theta_L - d)s_2^T\}] - C_1 s_1 - D^T s_2 \\ -C_2 s_2 \end{bmatrix}.\end{aligned}$$

$\nabla_\theta \log p(y,\theta)$ is model specific and we discuss the application to GLMMs and state space models in Sects. 8 and 9, respectively.

4.2 Stochastic variational algorithm

The stochastic gradient ascent algorithm for CSGVA is outlined in Algorithm 1. For computing the stepsize, we use Adam (Kingma and Ba 2015), which uses bias-corrected estimates of the first and second moments of the stochastic gradients to compute adaptive learning rates.

At iteration t, the variational parameter λ is updated as $\lambda_t = \lambda_{t-1} + \Delta_t$. Let g_t denote the stochastic gradient estimate at iteration t. In steps 3 and 4, Adam computes estimates of the mean (first moment) and uncentered variance (second

Initialize $\lambda_0 = 0, m_0 = 0, v_0 = 0$,

For $t = 1, \ldots, N$,

1. Generate $s \sim N(0, I_{nL+G})$ and compute $\theta = r_{\lambda_{t-1}}(s)$.
2. Compute gradient $g_t = \widehat{\nabla}_\lambda \mathcal{L}^{\mathrm{VI}}$.
3. Update biased first moment estimate: $m_t = \tau_1 m_{t-1} + (1-\tau_1) g_t$.
4. Update biased second moment estimate: $v_t = \tau_2 v_{t-1} + (1-\tau_2) g_t^2$.
5. Compute bias-corrected first moment estimate: $\hat{m}_t = m_t/(1-\tau_1^t)$.
6. Compute bias-corrected second moment estimate: $\hat{v}_t = v_t/(1-\tau_2^t)$.
7. Update $\lambda_t = \lambda_{t-1} + \alpha \hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$.

Algorithm 1: CSGVA algorithm.

moment) of the gradients using exponential moving averages, where $\tau_1, \tau_2 \in [0, 1)$ control the decay rates. In step 4, g_t^2 is evaluated as $g_t \odot g_t$, where $\odot$ denotes the Hadamard (element-wise) product. As m_t and v_t are initialized as zero, these moment estimates tend to be biased towards zero, especially at the beginning of the algorithm if τ_1, τ_2 are close to one. As $m_t = (1-\tau_1)\sum_{i=1}^t \tau_1^{t-i} g_i$,

$$E(m_t) = E(g_t)(1-\tau_1^t) + \zeta,$$

where $\zeta = 0$ if $E(g_i) = E(g_t)$ for $1 \le i < t$. Otherwise, ζ can be kept small since the weights for past gradients decrease exponentially. An analogous argument holds for v_t. Thus the bias can be corrected by using the estimates $\hat{m}_t$ and $\hat{v}_t$ in steps 5 and 6. The change is then computed as

$$\Delta_t = \frac{\alpha \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon},$$

where α controls the magnitude of the stepsize and ϵ is a small positive constant which ensures that the denominator is positive. In our experiments, we set $\alpha = 0.001$, $\tau_1 = 0.9$, $\tau_2 = 0.99$ and $\epsilon = 10^{-8}$, values close to what is recommended by Kingma and Ba (2015).

At each iteration t, we can also compute an unbiased estimate of the lower bound,

$$\hat{\mathcal{L}}^{\mathrm{VI}} = \log p(y, \theta) - \log q_{\lambda_{t-1}}(\theta),$$

where θ is computed in step 1. Since these estimates are stochastic, we follow the path traced by $\bar{\mathcal{L}}^{\mathrm{VI}}$, which is an average of the lower bounds averaged over every 1000 iterations, as a means to diagnose the convergence of Algorithm 1. $\bar{\mathcal{L}}^{\mathrm{VI}}$ tends to increase monotonically at the start, but as the algorithm comes close to convergence, the values of $\bar{\mathcal{L}}^{\mathrm{VI}}$ fluctuate close to and about the true maximum lower bound. Hence, we fit a least squares regression line to the past κ values of $\bar{\mathcal{L}}^{\mathrm{VI}}$ and terminate Algorithm 1 once the gradient of the regression line becomes negative (see Tan 2018). For our experiments, we set $\kappa = 6$.

5 Links to Gaussian variational approximation

CSGVA is an extension of Gaussian variational approximation (GVA, Tan and Nott 2018). In both approaches, the conditional posterior independence structure of the local latent variables is used to introduce sparsity in the precision matrix of the approximation. Below we demonstrate that GVA is a special case of CSGVA when $F \equiv 0$.

Tan and Nott (2018) consider a GVA of the form

$$\begin{bmatrix}\theta_L \\ \theta_G\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_L \\ \mu_G\end{bmatrix}, T^{-T}T^{-1}\right) \text{ where } T = \begin{bmatrix}T_{LL} & 0 \\ T_{GL} & T_{GG}\end{bmatrix}.$$

Note that T_{LL} and T_{GG} are lower triangular matrices. Using a vector $s = [s_L^T, s_G^T]^T \sim N(0, I_{nL+G})$, we can write

$$\begin{bmatrix}\theta_L \\ \theta_G\end{bmatrix} = \begin{bmatrix}\mu_L \\ \mu_G\end{bmatrix} + T^{-T}\begin{bmatrix}s_L \\ s_G\end{bmatrix},$$

$$\text{where } T^{-T} = \begin{bmatrix}T_{LL}^{-T} & -T_{LL}^{-T}T_{GL}^T T_{GG}^{-T} \\ 0 & T_{GG}^{-T}\end{bmatrix}.$$

Assuming $F \equiv 0$ for CSGVA, we have from (3) that

$$\begin{bmatrix}\theta_L \\ \theta_G\end{bmatrix} = \begin{bmatrix}d \\ \mu_1\end{bmatrix} + \begin{bmatrix}C_2^{-T} & -C_2^{-T}DC_1^{-T} \\ 0 & C_1^{-T}\end{bmatrix}\begin{bmatrix}s_2 \\ s_1\end{bmatrix},$$

where $[s_2^T, s_1^T]^T \sim N(0, I_{nL+G})$. Hence we can identify

$$\mu_1 = \mu_G,\ d = \mu_L,\ C_1 = T_{GG},\ C_2 = T_{LL},\ D = T_{GL}^T.$$

If the standard way of initializing of Algorithm 1 (by setting $\lambda = 0$) does not work well, we can use this association to initialize Algorithm 1 by using the fit from GVA. This informative initialization can reduce computation time significantly although there may be a risk of getting stuck in a local mode.

6 Importance weighted variational inference

Here we discuss how CSGVA can be improved by maximizing an importance weighted lower bound (IWLB, Burda et al. 2016), which leads to a tighter lower bound on the log marginal likelihood, and a variational approximation less prone to underestimation of the true posterior variance. We also relate the IWLB with Rényi's divergence (Rényi 1961; van Erven and Harremos 2014) between $q_\lambda(\theta)$ and $p(\theta|y)$, demonstrating that maximizing the IWLB instead of the usual

evidence lower bound leads to a transition in the behavior of the variational approximation from "mode-seeking" to "mass-covering". We first define Rényi's divergence and the variational Rényi bound (Li and Turner 2016), before introducing the IWLB as the expectation of a Monte Carlo approximation of the variational Rényi bound.

6.1 Rényi's divergence and variational Rényi bound

Rényi's divergence provides a measure of the distance between two densities q and p, and it is defined as

$$D_\alpha(q||p) = \frac{1}{\alpha - 1} \log \int q(\theta)^\alpha p(\theta)^{1-\alpha} d\theta,$$

for $0 < \alpha < \infty$, $\alpha \neq 1$. This definition can be extended by continuity to the orders 0, 1 and ∞, as well as to negative orders $\alpha < 0$. Note that $D_\alpha(q||p)$ is no longer a divergence measure if $\alpha < 0$, but we can write $D_\alpha(q||p)$ as $\frac{\alpha}{1-\alpha} D_{1-\alpha}(p||q)$ for $\alpha \notin \{0, 1\}$ by the skew symmetry property. As α approaches 1, the limit of $D_\alpha(q||p)$ is the Kullback-Leibler divergence, $\mathrm{KL}(q||p)$. In variational inference, minimizing the Kullback-Leibler divergence between the variational density $q_\lambda(\theta)$ and the true posterior $p(\theta|y)$ is equivalent to maximizing a lower bound $\mathcal{L}^{\mathrm{VI}}$ on the log marginal likelihood due to the relationship:

$$\mathcal{L}^{\mathrm{VI}} = \log p(y) - \mathrm{KL}\{q_\lambda||p(\theta|y)\} = E_{q_\lambda}\left\{\log \frac{p(y,\theta)}{q_\lambda(\theta)}\right\}.$$

Generalizing this relation using Rényi's divergence measure, Li and Turner (2016) define the variational Rényi bound $\mathcal{L}_\alpha$ as

$$\begin{aligned}\mathcal{L}_\alpha &= \log p(y) - D_\alpha\{q_\lambda||p(\theta|y)\} \\ &= \frac{1}{1-\alpha} \log E_{q_\lambda}\left\{\left(\frac{p(y,\theta)}{q_\lambda(\theta)}\right)^{1-\alpha}\right\}.\end{aligned}$$

Note that $\mathcal{L}_1$, the limit of $\mathcal{L}_\alpha$ as $\alpha \to 1$, is equal to $\mathcal{L}^{\mathrm{VI}}$. A Monte Carlo approximation of $\mathcal{L}_\alpha$ when the expectation is intractable is

$$\hat{\mathcal{L}}_{\alpha,K} = \frac{1}{1-\alpha} \log \frac{1}{K} \sum_{k=1}^{K} w_k^{1-\alpha}, \tag{8}$$

where $\Theta_K = [\theta_1, \ldots, \theta_K]$ is a set of K samples generated randomly from $q_\lambda(\theta)$, and

$$w_k = w(\theta_k) = \frac{p(y,\theta_k)}{q_\lambda(\theta_k)}, \quad k = 1, \ldots, K,$$

are importance weights. For each k, $E_{q_\lambda}(w_k) = p(y)$. The limit of $\hat{\mathcal{L}}_{\alpha,K}$ as $\alpha \to 1$ is $\frac{1}{K}\sum_{k=1}^{K} \log \frac{p(y,\theta_k)}{q_\lambda(\theta_k)}$. Hence $\hat{\mathcal{L}}_{1,K}$ is an unbiased estimate of $\mathcal{L}_1$ as $E_{\Theta_K}(\hat{\mathcal{L}}_{1,K}) = \mathcal{L}_1 = \mathcal{L}^{\mathrm{VI}}$, where E_{Θ_K} denotes expectation with respect to $q(\Theta_K) = \prod_{k=1}^{K} q_\lambda(\theta_k)$. For $\alpha \neq 1$, $\hat{\mathcal{L}}_{\alpha,K}$ is not an unbiased estimate of $\mathcal{L}_\alpha$.

6.2 Importance weighted lower bound

The importance weighted lower bound (IWLB, Burda et al. 2016) is defined as

$$\mathcal{L}_K^{\mathrm{IW}} = E_{\Theta_K}(\hat{\mathcal{L}}_{0,K}) = E_{\Theta_K}\left(\log \frac{1}{K}\sum_{k=1}^{K} w_k\right),$$

where $\alpha = 0$ in (8). It reduces to $\mathcal{L}^{\mathrm{VI}}$ when $K = 1$. By Jensen's inequality,

$$\mathcal{L}_K^{\mathrm{IW}} \leq \log E_{\Theta_K}\left(\frac{1}{K}\sum_{k=1}^{K} w_k\right) = \log p(y).$$

Thus $\mathcal{L}_K^{\mathrm{IW}}$ provides a lower bound to the log marginal likelihood for any positive integer K. From Theorem 1 (Burda et al. 2016), this bound becomes tighter as K increases.

Theorem 1 *$\mathcal{L}_K^{IW}$ increases with K and approaches $\log p(y)$ as $K \to \infty$ if w_k is bounded.*

Proof Let $I = \{w_{I_1}, \ldots, w_{I_K}\}$ be selected randomly without replacement from $\{w_1, \ldots, w_{K+1}\}$. Then $E_{I|\Theta_{K+1}}(w_{I_j}) = \frac{1}{K+1}\sum_{k=1}^{K+1} w_k$ for any $j = 1, \ldots, K$, where $E_{I|\Theta_{K+1}}$ denotes the expectation associated with the randomness in selecting I given Θ_{K+1}. Thus

$$\begin{aligned}\mathcal{L}_{K+1}^{\mathrm{IW}} &= E_{\Theta_{K+1}}\left(\log \frac{1}{K+1}\sum\nolimits_{k=1}^{K+1} w_k\right) \\ &= E_{\Theta_{K+1}}\left\{\log E_{I|\Theta_{K+1}}\left(\frac{1}{K}\sum\nolimits_{j=1}^{K} w_{I_j}\right)\right\} \\ &\geq E_{\Theta_{K+1}}\left\{E_{I|\Theta_{K+1}} \log\left(\frac{1}{K}\sum\nolimits_{j=1}^{K} w_{I_j}\right)\right\} \\ &= E_{\Theta_K} \log\left(\frac{1}{K}\sum\nolimits_{k=1}^{K} w_k\right) = \mathcal{L}_K^{\mathrm{IW}}.\end{aligned}$$

If w_k is bounded, then $\frac{1}{K}\sum_{k=1}^{K} w_k \overset{a.s.}{\to} p(y)$ as $K \to \infty$ by the law of large numbers. Hence $\mathcal{L}_K^{\mathrm{IW}} = E_{\Theta_K}(\log \frac{1}{K}\sum_{k=1}^{K} w_k) \to \log p(y)$ as $K \to \infty$. □

Next, we present some properties of Rényi's divergence and $E_{\Theta_K}(\hat{\mathcal{L}}_{\alpha,K})$ which are important in understanding the behavior of the variational density arising from maximizing $\mathcal{L}_K^{\mathrm{IW}}$. The proofs of these properties can be found in van Erven and Harremos (2014) and Li and Turner (2016).

Property 1 D_α is increasing in α, and is continuous in α on $[0, 1] \cup \{\alpha \notin [0, 1] \mid |D_\alpha| < \infty\}$.

Property 2 $E_{\Theta_K}(\hat{\mathcal{L}}_{\alpha,K})$ is continuous and decreasing in α for fixed K.

Theorem 2 *There exists* $0 \leq \alpha_{q_\lambda,K} \leq 1$ *for given* q_λ *and* K *such that*

$$\log p(y) = D_{\alpha_{q_\lambda,K}}\{q_\lambda || p(\theta|y)\} + \mathcal{L}_K^{IW}.$$

Proof From Property 2,

$$\begin{aligned}
\mathcal{L}_1 = E_{\Theta_K}(\hat{\mathcal{L}}_{1,K}) \leq E_{\Theta_K}(\hat{\mathcal{L}}_{0,K}) = \mathcal{L}_K^{\mathrm{IW}} \leq \mathcal{L}_0,\\
\mathcal{L}_1 - \log p(y) \leq \mathcal{L}_K^{\mathrm{IW}} - \log p(y) \leq \mathcal{L}_0 - \log p(y),\\
D_0\{q_\lambda||p(\theta|y)\} \leq \log p(y) - \mathcal{L}_K^{\mathrm{IW}} \leq D_1\{q_\lambda||p(\theta|y)\}.
\end{aligned}$$

From Property 1, since D_α is continuous and decreasing for $\alpha \in [0, 1]$, there exists $0 \leq \alpha_{q_\lambda,K} \leq 1$ such that $\log p(y) - \mathcal{L}_K^{\mathrm{IW}} = D_{\alpha_{q_\lambda,K}}\{q_\lambda||p(\theta|y)\}$. □

Minimizing Rényi's divergence for $\alpha \gg 1$ tends to produce approximations which are mode-seeking (zero-forcing), while maximizing Rényi's divergence for $\alpha \ll 0$ encourages mass-covering behavior. Theorem 2 suggests that maximizing the IWLB results in a variational approximation q_λ whose Rényi's divergence from the true posterior can be captured with $0 \leq \alpha \leq 1$, which represents a mix and certain balance between mode-seeking and mass-covering behavior (Minka 2005). In our experiments, we observe that maximizing the IWLB is highly effective in correcting the underestimation of posterior variance in variational inference.

Alternatively, if we approximate $\mathcal{L}_K^{\mathrm{IW}}$ by considering a second-order Taylor expansion of $\log \bar{w}_K$ about $E_\Theta(\bar{w}_K) = p(y)$, where $\bar{w}_K = \frac{1}{K}\sum_{k=1}^K w_K$, and then take expectations, we have

$$\mathcal{L}_K^{\mathrm{IW}} \approx \log p(y) - \frac{\mathrm{var}(w_k)}{2Kp(y)^2}.$$

Maddison et al. (2017) and Domke and Sheldon (2018) provide bounds on the order of the remainder term in the Taylor approximation above, and demonstrate that the "looseness" of the IWLB is given by $\mathrm{var}(w_k)$ as $K \to \infty$. Minimizing $\mathrm{var}(w_K)$ is equivalent to minimizing the χ^2 divergence $D_2(p||q)$. Note that if $q_\lambda(\theta)$ has thin tails compared to $p(\theta|y)$, then the variance of $\mathrm{var}(w_k)$ will be large. Hence minimizing $\mathrm{var}(w_K)$ attempts to match $p(\theta|y)$ with $q_\lambda(\theta)$ so that $q_\lambda(\theta)$ is able to cover the tails.

6.3 Unbiased gradient estimate of importance weighted lower bound

To maximize the IWLB in CSGVA, we need to find an unbiased estimate of $\nabla_\lambda \mathcal{L}_K^{\mathrm{IW}}$ using the transformation in (3). Let $s_k \sim N(0, I_{G+nL})$, $\theta_k = r_\lambda(s_k)$ for $k = 1, \ldots, K$, and $S_K = [s_1, \ldots, s_K]^T$.

$$\begin{aligned}
\nabla_\lambda \mathcal{L}_K^{\mathrm{IW}} &= \nabla_\lambda E_{\Theta_K}(\log \bar{w}_K) = \nabla_\lambda E_{S_K}(\log \bar{w}_K)\\
&= E_{S_K}\left[\sum_{k=1}^K \frac{\nabla_\lambda w_k}{\sum_{k'=1}^K w_{k'}}\right]\\
&= E_{S_K}\left[\sum_{k=1}^K \frac{w_k \nabla_\lambda \log w_k}{\sum_{k'=1}^K w_{k'}}\right]\\
&= E_{S_K}\left[\sum\nolimits_{k=1}^K \widetilde{w}_k \nabla_\lambda \log w_k\right],
\end{aligned} \tag{9}$$

where $w_k = w(\theta_k) = w\{r_\lambda(s_k)\}$ and $\widetilde{w}_k = w_k / \sum_{k'=1}^K w_{k'}$ for $k = 1, \ldots, K$ are normalized importance weights. Applying chain rule,

$$\nabla_\lambda \log w_k = \nabla_\lambda r_\lambda(s_k) \nabla_{\theta_k} \log w_k - \nabla_\lambda \log q_\lambda(\theta_k). \tag{10}$$

In Sect. 4.1, we note that $E_\phi\{\nabla_\lambda \log q_\lambda(\theta)\} = 0$ as it is the expectation of the score function, and hence, we can omit $\nabla_\lambda \log q_\lambda(\theta)$ to obtain an unbiased estimate of $\nabla_\lambda \mathcal{L}^{\mathrm{VI}}$. However, in this case, it is unclear if

$$E_{S_K}\left[\sum_{k=1}^K \widetilde{w}_k \nabla_\lambda \log q_\lambda(\theta_k)\right] = 0. \tag{11}$$

Roeder et al. (2017) conjecture that (11) is true and report improved results when omitting the term $\nabla_\lambda \log q_\lambda(\theta_k)$ from $\nabla_\lambda \log w_k$ in computing gradient estimates. However, Tucker et al. (2018) demonstrated via simulations that (11) does not hold generally and that such omission will result in biased gradient estimates. Our own simulations using CSGVA also suggest that (11) does not hold even though omission of $\nabla_\lambda \log q_\lambda(\theta_k)$ does lead to improved results. As the stochastic gradient algorithm is not guaranteed to converge with biased gradient estimates, we turn to the *double reparametrized gradient estimate* proposed by Tucker et al. (2018) which allows unbiased gradient estimates to be constructed with the omission of $\nabla_\lambda \log q_\lambda(\theta_k)$ albeit with revised weights.

Since $\tilde{w}_k$ depends on λ directly as well as through θ_k, we use chain rule to obtain

$$\begin{aligned}
\nabla_\lambda E_{\theta_k}(\tilde{w}_k) &= \nabla_\lambda E_{s_k}(\tilde{w}_k)\\
&= E_{s_k}(\nabla_\lambda \theta_k \nabla_{\theta_k} \tilde{w}_k) + E_{s_k}(\nabla_\lambda \tilde{w}_k),
\end{aligned} \tag{12}$$

where

$$\begin{aligned}
\nabla_{\theta_k} \tilde{w}_k &= \left\{\frac{1}{\sum_{k'=1}^K w_{k'}} - \frac{w_k}{(\sum_{k'=1}^K w_{k'})^2}\right\} \nabla_{\theta_k} w_k\\
&= (\tilde{w}_k - \tilde{w}_k^2) \nabla_{\theta_k} \log w_k.
\end{aligned}$$

Alternatively,

$$\begin{aligned}\nabla_\lambda E_{\theta_k}(\tilde{w}_k) &= \nabla_\lambda \int q_\lambda(\theta_k)\tilde{w}_k \, d\theta_k \\ &= \int \tilde{w}_k \nabla_\lambda q_\lambda(\theta_k) + q_\lambda(\theta_k)\nabla_\lambda \tilde{w}_k \, d\theta_k \\ &= \int \tilde{w}_k q_\lambda(\theta_k)\nabla_\lambda \log q_\lambda(\theta_k) \, d\theta_k + E_{\theta_k}(\nabla_\lambda \tilde{w}_k) \\ &= E_{\theta_k}[\tilde{w}_k \nabla_\lambda \log q_\lambda(\theta_k)] + E_{s_k}(\nabla_\lambda \tilde{w}_k).\end{aligned} \tag{13}$$

Comparing (12) and (13), we have

$$\begin{aligned}&E_{\Theta_K}\left(\sum\nolimits_{k=1}^{K} \tilde{w}_k \nabla_\lambda \log q_\lambda(\theta_k)\right) \\ &= \sum\nolimits_{k=1}^{K} E_{\Theta_K \setminus \theta_k} E_{\theta_k}[\tilde{w}_k \nabla_\lambda \log q_\lambda(\theta_k)] \\ &= \sum\nolimits_{k=1}^{K} E_{S_K \setminus s_k} E_{s_k}(\nabla_\lambda \theta_k (\tilde{w}_k - \tilde{w}_k^2)\nabla_{\theta_k} \log w_k) \\ &= E_{S_K}\left\{\sum\nolimits_{k=1}^{K} (\tilde{w}_k - \tilde{w}_k^2)\nabla_\lambda r_\lambda(s_k)\nabla_{\theta_k} \log w_k\right\}.\end{aligned}$$

Combining the above expression with (9) and (10), we find that

$$\nabla_\lambda \mathcal{L}_K^{\mathrm{IW}} = E_{S_K}\left(\sum\nolimits_{k=1}^{K} \tilde{w}_k^2 \nabla_\lambda r_\lambda(s_k)\nabla_{\theta_k} \log w_k\right)$$

An unbiased gradient estimate is thus given by

$$\widehat{\nabla}_\lambda \mathcal{L}_K^{\mathrm{IW}} = \sum_{k=1}^{K} \tilde{w}_k^2 \nabla_\lambda r_\lambda(s_k)\nabla_{\theta_k}\{\log p(y, \theta_k) - \log q_\lambda(\theta_k)\}.$$

Thus, to use CSGVA with important weights, we only need to modify Algorithm 1 by drawing K samples $s_1, \ldots, s_K$ in step 1 instead of a single sample and then compute the unbiased gradient estimate, $g_t = \widehat{\nabla}_\lambda \mathcal{L}_K^{\mathrm{IW}}$, in step 2. The rest of the steps in Algorithm 1 remain unchanged. In the importance weighted version of CSGVA, the gradient estimate based on a single sample s is replaced by a weighted sum of the gradients in (7) based on K samples $s_1, \ldots, s_K$. However, these weights do not necessarily sum to 1. An unbiased estimate of $\mathcal{L}_K^{\mathrm{IW}}$ itself is given by $\hat{\mathcal{L}}_K^{\mathrm{IW}} = \log \frac{1}{K} \sum_{k=1}^{K} w_k$.

7 Experimental results

We apply CSGVA to GLMMs and state space models and compare the results with that of GVA in terms of computation time and accuracy of the posterior approximation. Lower bounds reported exclude constants which are independent of the model variables. We also illustrate how CSGVA can be improved using importance weighting (IW-CSGVA), considering $K \in \{5, 20, 100\}$. We find that IW-CSGVA performs poorly if it is initialized in the standard manner using $\lambda = 0$. This is because, when $q_\lambda(\theta)$ is still far from optimal, a few of the importance weights tend to dominate with the rest close to zero, thus producing inaccurate estimates of the gradients. Hence, we initialize IW-CSGVA using the CSGVA fit, and the algorithm is terminated after a short run of 1000 iterations as IW-CSGVA is computationally intensive and improvements in the IWLB and variational approximation seem to be negligible thereafter. Posterior distributions estimated using MCMC via RStan are regarded as the ground truth. Code for all variational algorithms are written in Julia and are available as supplementary materials. All experiments are run on Intel Core i9-9900K CPU @3.60 GHz with 16GB RAM. As the computation time of IW-CSGVA increases linearly with K, we also investigate the speedup that can be achieved through parallel computing on a machine with 8 cores. Julia retains one worker (or core) as the master process, and parallel computing is implemented using the remaining seven workers.

The parametrization of a hierarchical model plays a major role in the rate of convergence of both GVA and CSGVA. In some cases, it can even affect the ability of the algorithm to converge (to a local mode). We have attempted both the centered and noncentered parametrizations (Papaspiliopoulos et al. 2003, 2007), which are known to have a huge impact on the rate of convergence of MCMC algorithms. These two parametrizations are complementary, and neither is superior to the other. If an algorithm converges slowly under one parametrization, it often converges much faster under the other. Which parametrization works better usually depends on the nature of data. For the datasets that we use in the experiments, the centered parametrization was found to have better convergence properties than the noncentered parametrization for GLMMs while the noncentered parametrization is preferred for stochastic volatility models. These observations are further discussed below.

8 Generalized linear mixed models

Let $y_i = (y_{i1}, \ldots, y_{in_i})^T$ denote the vector of responses of length n_i for the ith subject for $i = 1, \ldots, n$, where y_{ij} is generated from some distribution in the exponential family. The mean $\mu_{ij} = E(y_{ij})$ is connected to the linear predictor η_{ij} via

$$g(\mu_{ij}) = \eta_{ij} = X_{ij}^T \beta + Z_{ij}^T b_i,$$

for some smooth invertible link function $g(\cdot)$. The fixed effects β is a $p \times 1$ vector and $b_i \sim N(0, \Lambda)$ is a $L \times 1$ vector of random effects specific to the ith subject. X_{ij} and Z_{ij} are vectors of covariates of length p and L, respectively. Let $\eta_i = [\eta_{i1}, \ldots, \eta_{in_i}]^T$, $X_i = [X_{i1}, \ldots, X_{in_i}]^T$

and $Z_i = [Z_{i1}, \ldots, Z_{in_i}]^T$. We focus on the one-parameter exponential family with canonical links. This includes the Bernoulli model, $y_{ij} \sim \text{Bern}(\mu_{ij})$, with the logit link $g(\mu_{ij}) = \log\{\mu_{ij}/(1-\mu_{ij})\}$ and the Poisson model, $y_{ij} \sim \text{Pois}(\mu_{ij})$, with the log link $g(\mu_{ij}) = \log(\mu_{ij})$. Let WW^T be the unique Cholesky decomposition of Λ^{-1}, where W is a lower triangular matrix with positive diagonal entries. Define W^* such that $W^*_{ii} = \log(W_{ii})$ and $W^*_{ij} = W_{ij}$ if $i \neq j$, and let $\omega = v(W^*)$. We consider normal priors, $\beta \sim N(0, \sigma^2_\beta I)$ and $\omega \sim N(0, \sigma^2_\omega I)$, where σ^2_β and σ^2_ω are set as 100.

The above parametrization of the GLMM is *noncentered* since b_i has mean 0. Alternatively, we can consider the *centered* parametrization proposed by Tan and Nott (2013). Suppose the covariates for the random effects are a subset of the covariates for the fixed effects and the first column of X_i and Z_i are ones corresponding to an intercept and random intercept, respectively. Then we can write

$$\eta_i = X_i\beta + Z_i b_i = Z_i\beta_R + X_i^G\beta_G + Z_i b_i.$$

where $\beta = [\beta_R^T, \beta_G^T]^T$. We further split X_i^G into covariates which are subject specific (varies only with i and assumes the same value for $j = 1, \ldots, n_i$) and those which are not, and $\beta_G = [\beta_{G_1}^T, \beta_{G_2}^T]^T$ accordingly, where β_{G_1}, β_{G_2} are vectors of length g_1 and g_2, respectively. Then

$$\begin{aligned}\eta_i &= Z_i\beta_R + \mathbf{1}_{n_i} x_i^{G_1 T}\beta_{G_1} + X_i^{G_2}\beta_{G_2} + Z_i b_i \\ &= Z_i(C_i\beta_{RG_1} + b_i) + X_i^{G_2}\beta_{G_2},\end{aligned}$$

where

$$C_i = \begin{bmatrix} I_L & x_i^{G_1 T} \\ & 0_{L-1\times g_1} \end{bmatrix} \quad \text{and} \quad \beta_{RG_1} = \begin{bmatrix}\beta_R \\ \beta_{G_1}\end{bmatrix}.$$

Let $\tilde{b}_i = C_i\beta_{RG_1} + b_i$. The centered parametrization is represented as

$$\eta_i = Z_i\tilde{b}_i + X_i^{G_2}\beta_{G_2}, \quad \tilde{b}_i \sim N(C_i\beta_{RG_1}, \Lambda) \tag{14}$$

for $i = 1, \ldots, n$.

Tan and Nott (2013) compare the centered, noncentered and partially noncentered parametrizations for GLMMs in the context of variational Bayes, showing that the choice of parametrization affects not only the rate of convergence, but also the accuracy of the variational approximation. For CSGVA, we observe that the accuracy of the variational approximation and the rate of convergence can also be greatly affected. Tan and Nott (2013) demonstrate that the centered parametrization is preferred when the observations are highly informative about the latent variables. In practice, a general guideline is to use the centered parametrization for Poisson models when observed counts are large and the noncentered parametrization when most counts are close to zero. For Bernoulli models, differences between using centered and noncentered parametrizations are usually minor. Here we use the centered parametrization in (14), which has been observed to yield gains in convergence rates for the datasets used for illustration.

The global parameters are $\theta_G = (\beta^T, \omega^T)^T$ of dimension $G = p + L(L+1)/2$, and the local variables are $\theta_L = (\tilde{b}_1, \ldots, \tilde{b}_n)^T$. The joint density is

$$p(y, \theta) = p(\beta)p(\omega)\prod_{i=1}^n \left\{ p(\tilde{b}_i|\omega, \beta)\prod_{j=1}^{n_i} p(y_{ij}|\beta, \tilde{b}_i)\right\}.$$

The log of the joint density is given by

$$\begin{aligned}\log p(y, \theta) = &\sum_{i=1}^n \{y_i\eta_i - \mathbf{1}^T h(\eta_i) \\ &- (\tilde{b}_i - C_i\beta_{RG_1})^T WW^T(\tilde{b}_i - C_i\beta_{RG_1})/2\} \\ &- \beta^T\beta/(2\sigma^2_\beta) - \omega^T\omega/(2\sigma^2_\omega) + n\log|W| + c,\end{aligned}$$

where $h(\cdot)$ is the log-partition function and c is a constant independent of θ. For the Poisson model with log link, $h(x) = \exp(x)$. For the Bernoulli model with logit link, $h(x) = \log\{1+\exp(x)\}$. The gradient $\nabla_\theta \log p(y, \theta)$ is given in "Appendix B".

For the GLMM, b_i and b_j are conditionally independent given θ_G for $i \neq j$ in $p(\theta|y)$. Hence we impose the following sparsity structure on Ω_2 and C_2,

$$\Omega_2 = \begin{bmatrix} \Omega_{2,11} & 0 & \ldots & 0 \\ 0 & \Omega_{2,22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Omega_{2,nn} \end{bmatrix},$$

$$C_2^* = \begin{bmatrix} C^*_{2,11} & 0 & \ldots & 0 \\ 0 & C^*_{2,22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & C^*_{2,nn} \end{bmatrix},$$

where each $\Omega_{2,ii}$ is a $L \times L$ block matrix and each $C^*_{2,ii}$ is a $L \times L$ lower triangular matrix. We set $f_i = 0$ and $F_{ij} = 0$ for all $i \in \mathcal{I}$ and all j, where $\mathcal{I}$ denotes the set of indices in $v(C_2^*)$ which are fixed as zero. The number of nonzero elements in $v(C_2^*)$ is $nL(L+1)/2$. Hence the number of variational parameters to be optimized are reduced from $nL(nL+1)/2$ to $nL(L+1)/2$ for f and from $nL(nL+1)G/2$ to $nL(L+1)G/2$ for F.

Table 1 Epilepsy data. Number of iterations I (in thousands), runtimes (in s) and estimates of lower bound (SD in brackets) of the variational methods

	K	I	Time	Parallel	$\hat{\mathcal{L}}_K^{\text{IW}}$
GVA	1	31	13.9	–	3138.3 (1.8)
CSGVA	1	39	16.2	–	3139.2 (1.5)
IW-CSGVA	5	1	2.5	6.1	3139.9 (0.7)
	20	1	6.9	8.1	3140.1 (0.4)
	100	1	33.5	16.0	3140.1 (0.3)

8.1 Epilepsy data

In this epilepsy data (Thall and Vail 1990), $n = 59$ patients are involved in a clinical trial to investigate the effects of the anti-epileptic drug Progabide. The patients are randomly assigned to receive either the drug (Trt $= 1$) or a placebo (Trt $= 0$). The response y_i denotes the number of epileptic attacks patient i had during 4 follow-up periods of two weeks each. Covariates include the log of the age of the patients (Age), the log of 1/4 the baseline seizure count recorded over an eight-week period prior to the trial (Base) and Visit (coded as $\text{Visit}_1 = -0.3$, $\text{Visit}_2 = -0.1$, $\text{Visit}_3 = 0.1$ and $\text{Visit}_4 = 0.3$). Note that Age is centered about its mean. Consider $y_{ij} \sim \text{Pois}(\mu_{ij})$, where

$$\begin{aligned}\log \mu_{ij} = {} & \beta_0 + \beta_{\text{Base}}\text{Base}_i + \beta_{\text{Trt}}\text{Trt}_i + \beta_{\text{Age}}\text{Age}_i \\ & + \beta_{\text{Base}\times\text{Trt}}\text{Base}_i \times \text{Trt}_i + \beta_{\text{Visit}}\text{Visit}_{ij} \\ & + b_{i1} + b_{i2}\text{Visit}_{ij},\end{aligned}$$

for $i = 1, \ldots, 59$, $j = 1, \ldots, 4$ (Breslow and Clayton 1993).

Table 1 shows the results obtained from applying the variational algorithms to this data. The lower bounds are estimated using 1000 simulations in each case, and the mean and standard deviation from these simulations are reported. CSGVA produced an improvement in the estimate of the lower bound (3139.2) as compared to GVA (3138.3) and maximizing the IWLB led to further improvements. Using a larger K of 20 or 100 resulted in minimal improvements compared with $K = 5$. As this dataset is small, parallel computing is slower than serial for a small K. This is because, even though the importance weights and gradients for K samples are computed in parallel, the cost of sending and fetching data from the workers at each iteration dwarfs the cost of computation when K is small. For $K = 100$, parallel computing reduces the computation time by about half.

The estimated marginal posterior distributions of the global parameters are shown in Fig. 1. The plots show that CSGVA (red) produces improved estimates of the posterior distribution as compared to GVA (blue), especially in estimating the posterior variance of the precision parameters ω_2 and ω_3. The posteriors estimated using IW-CSGVA for the different values of K are very close. By using just $K = 5$, we are able to obtain estimates that are virtually indistinguishable from that of MCMC.

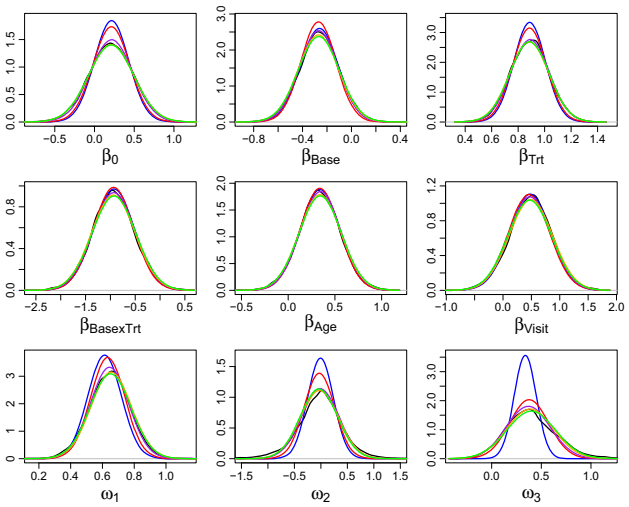

Fig. 1 Epilepsy data. Marginal posterior distributions of global parameters. Black (MCMC), blue (GVA), red (CSGVA), purple ($K = 5$), orange ($K = 20$), green ($K = 100$). (Color figure online)

8.2 Madras data

These data come from the Madras longitudinal schizophrenia study (Diggle et al. 2002) for detecting a psychiatric symptom called "thought disorder." Monthly records showing whether the symptom is present in a patient are kept for $n = 86$ patients over 12 months. The response y_{ij} is a binary indicator for presence of the symptom. Covariates include the time in months since initial hospitalization (t), gender of patient (Gender = 0 if male and 1 if female) and age of patient (Age = 0 if the patient is younger than 20 years and 1 otherwise). Consider $y_{ij} \sim \text{Bern}(\mu_{ij})$ and

$$\begin{aligned}\text{logit}(\mu_{ij}) = {} & \beta_0 + \beta_{\text{Age}}\,\text{Age}_i + \beta_{\text{Gender}}\,\text{Gender}_i + \beta_t t_{ij} \\ & + \beta_{\text{Age}\times t}\,\text{Age}_i \times t_{ij} + \beta_{\text{Gender}\times t}\,\text{Gender}_i \times t_{ij} + b_i,\end{aligned}$$

for $i = 1, \ldots, 86$, $1 \le j \le 12$.

The results in Table 2 are quite similar to that of the epilepsy data. CSGVA yields an improvement in the lower bound estimate as compared to GVA and IW-CSGVA provided further improvements, which are evident starting with a K as small as 5. Parallel computing halved the computation time for $K = 100$ but did not yield any benefits for $K \in \{5, 20\}$. From Fig. 2, CSGVA and IW-CSGVA are again able to capture the posterior variance of the precision parameter ω_1 better than GVA.

Table 2 Madras data. Number of iterations I (in thousands), runtimes (in s) and estimates of lower bound (SD in brackets) of the variational methods

	K	I	Time	Parallel	$\hat{\mathcal{L}}_K^{\text{IW}}$
GVA	1	25	13.1	–	−383.4 (1.4)
CSGVA	1	35	12.6	–	−383.1 (1.4)
IW-CSGVA	5	1	2.4	7.1	−382.5 (0.7)
	20	1	6.8	8.9	−382.4 (0.4)
	100	1	33.9	16.8	−382.3 (0.2)

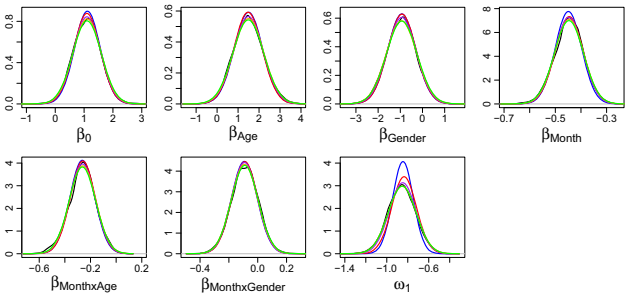

Fig. 2 Madras data. Marginal posterior distributions of global parameters. Black (MCMC), blue (GVA), red (CSGVA), purple ($K = 5$), orange ($K = 20$), green ($K = 100$). (Color figure online)

Table 3 Six cities data. Number of iterations I (in thousands), runtimes (in s) and estimates of lower bound (SD in brackets) of the variational methods

	K	I	Time	parallel	$\hat{\mathcal{L}}_K$
GVA	1	26	60.3	–	−816.4 (4.0)
CSGVA	1	28	36.5	–	−816.0 (3.9)
IW-CSGVA	5	1	6.5	16.3	−812.6 (2.5)
	20	1	23.1	24.5	−811.0 (1.9)
	100	1	115.5	61.4	−809.8 (1.5)

8.3 Six cities data

In the six cities data (Fitzmaurice and Laird 1993), $n = 537$ children from Steubenville, Ohio, are involved in a longitudinal study to investigate the health effects of air pollution. Each child is examined yearly from age 7 to 10, and the response y_{ij} is a binary indicator for wheezing. There are two covariates, Smoke_i (a binary indicator for smoking status of the mother of child i) and Age_{ij} (age of child i at time point j, centered at 9 years). Consider $y_{ij} \sim \text{Bern}(\mu_{ij})$, where

$$\begin{aligned}\text{logit}(\mu_{ij}) = \beta_0 + \beta_{\text{Smoke}}\text{Smoke}_i + \beta_{\text{Age}}\,\text{Age}_{ij} \\ + \beta_{\text{Smoke}\times\text{Age}}\,\text{Smoke}_i \times \text{Age}_{ij} + b_i,\end{aligned}$$

for $i = 1, \ldots, 537$, $j = 1, \ldots, 4$.

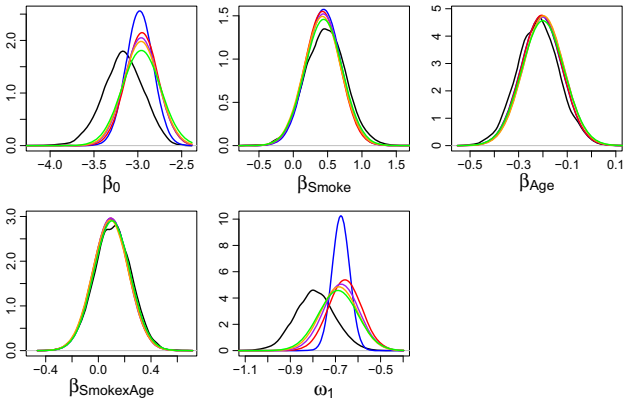

Fig. 3 Six cities data. Marginal posterior distributions of global parameters. Black (MCMC), blue (GVA), red (CSGVA), purple ($K = 5$), orange ($K = 20$), green ($K = 100$). (Color figure online)

From Table 3, CSGVA managed to achieve a higher lower bound than GVA in about half the runtime. As K increases, IW-CSGVA produced tighter lower bounds for the log marginal likelihood. As in the previous two examples, parallel computing is beneficial only when $K = 100$, cutting the runtime by about half. From Fig. 3, there is slight overestimation of the posterior means of β_0 and ω_1 by all the variational methods. However, CSGVA and IW-CSGVA are able to capture the posterior variance of these parameters much better than GVA especially for ω_1.

9 Application to state space models

Here we consider the stochastic volatility model (SVM) widely used for modeling financial time series. Let each observation y_i for $i = 1, \ldots, n$, be generated from a zero-mean Gaussian distribution where the error variance is stochastically evolving over time. The unobserved log volatility b_i is modeled using an autoregressive process of order one with Gaussian disturbances:

$$\begin{aligned}&y_i|b_i, \sigma, \kappa \sim N(0, \mathrm{e}^{\sigma b_i + \kappa}), \quad (i = 1, \ldots, n)\\ &b_i|\phi \sim N(\phi b_{i-1}, 1), \quad (i = 2, \ldots, n)\\ &b_1|\phi \sim N(0, 1/(1-\phi^2)),\end{aligned}$$

where $\kappa \in \mathbb{R}$, $\sigma > 0$ and $0 < \phi < 1$. Here, y_i represents the mean-corrected return at time i and the states $\{b_i\}$ come from a stationary process with b_1 drawn from the stationary distribution. The parametrization of the SVM above is noncentered. The centered parametrization can be obtained by replacing b_i by $(\tilde{b}_i - \kappa)/\sigma$. The performance of GVA and CSGVA are sensitive to the parametrization both in terms of rate of convergence and attained local mode. For the data sets below, the noncentered parametrization was found to have better convergence properties. The

sensitivity of the stochastic volatility model to parametrization when fitted using MCMC algorithms is well known in the literature. To "combine the best of different worlds," Kastner and Frühwirth-Schnatter (2014) introduce a strategy that samples parameters from the centered and noncentered parametrizations alternately. Tan (2017) studies optimal partially noncentered parametrizations for Gaussian state space models fitted using EM, MCMC or variational algorithms.

We use the following transformations to map constrained parameters to $\mathbb{R}$:

$$\alpha = \log(\exp(\sigma) - 1), \quad \psi = \text{logit}(\phi).$$

This transformation for α works better than $\alpha = \log(\sigma)$, which leads to erratic fluctuations in the lower bound and convergence issues more often. The global variables are $\theta_G = [\alpha, \kappa, \psi]^T$ of dimension $G = 3$ and the local variables are $\theta_L = [b_1, \ldots, b_n]^T$ of length n. We assume normal priors for the global parameters, where $\alpha \sim N(0, \sigma_\alpha^2)$, $\kappa \sim N(0, \sigma_\kappa^2)$ and $\psi \sim N(0, \sigma_\psi^2)$, where $\sigma_\alpha^2 = \sigma_\kappa^2 = \sigma_\psi^2 = 10$. The joint density can be written as

$$\begin{aligned} p(y, \theta) =& p(\alpha)p(\kappa)p(\psi)p(b_1|\psi)\left\{\prod_{i=2}^{n} p(b_i|b_{i-1}, \psi)\right\} \\ &\times \left\{\prod_{i=1}^{n} p(y_i|b_i, \alpha, \kappa)\right\}. \end{aligned}$$

The log joint density is

$$\begin{aligned} \log p(y, \theta) =& -\frac{n\kappa}{2} - \frac{\sigma}{2}\sum_{i=1}^{n} b_i - \frac{1}{2}\sum_{i=1}^{n} y_i^2 \mathrm{e}^{-\sigma b_i - \kappa} \\ &- \frac{1}{2}\sum_{i=2}^{n}(b_i - \phi b_{i-1})^2 - \frac{1}{2}b_1^2(1-\phi^2) \\ &+ \frac{1}{2}\log(1-\phi^2) - \frac{\alpha^2}{2\sigma_\alpha^2} - \frac{\kappa^2}{2\sigma_\kappa^2} - \frac{\psi^2}{2\sigma_\psi^2} + c, \end{aligned}$$

where $\phi = \exp(\psi)/\{1 + \exp(\psi)\}$, $\sigma = \log(\exp(\alpha) + 1)$ and c is a constant independent of θ. The gradient $\nabla_\theta \log p(y, \theta)$ is given in "Appendix C". For this model, b_i is conditionally independent of b_j in the posterior if $|i - j| > 1$ given θ_G. Thus, the sparsity structure imposed on Ω_2 and C_2 are

$$\Omega = \begin{bmatrix} \Omega_{2,11} & \Omega_{2,12} & 0 & \ldots & 0 \\ \Omega_{2,21} & \Omega_{2,22} & \Omega_{2,23} & \ldots & 0 \\ 0 & \Omega_{2,32} & \Omega_{2,33} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \Omega_{2,nn} \end{bmatrix},$$

$$C_2 = \begin{bmatrix} C_{2,11} & 0 & 0 & \ldots & 0 \\ C_{2,21} & C_{2,22} & 0 & \ldots & 0 \\ 0 & C_{2,32} & C_{2,33} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & C_{2,nn} \end{bmatrix}.$$

The number of nonzero elements in $v(C_2^*)$ is $2n - 1$. Setting $f_i = 0$ and $F_{ij} = 0$ for all $i \in \mathcal{I}$ and all j, where $\mathcal{I}$ denotes the set of indices in $v(C_2^*)$ which are fixed as zero, the number of variational parameters to be optimized are reduced from $n(n+1)/2$ to $2n-1$ for f, and from $n(n+1)G/2$ to $(2n-1)G$ for F.

9.1 GBP/USD exchange rate data

These data contain 946 observations of the exchange rates of the US Dollar (USD) against the British Pound (GBP), recorded daily from 1 October 1981, to 28 June 1985. These data are available from the "Garch" dataset in the R package `Ecdat`. The mean-corrected responses $\{y_t | t = 1, \ldots, n\}$ are computed from the exchange rates $\{r_t | t = 0, \ldots, n\}$ as

$$y_t = 100\left\{\log\left(\frac{r_t}{r_{t-1}}\right) - \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{r_i}{r_{i-1}}\right)\right\},$$

where $n = 945$.

For these data, CSGVA failed to achieve a higher lower bound when it was initialized using $\lambda = 0$. Hence we initialize CSGVA using the fit from GVA instead, by using the association discussed in Sect. 5. With this informative starting point, CSGVA converged in 16000 iterations and managed to improve upon the GVA fit, attaining a higher lower bound. IW-CSGVA led to further improvements with increasing K. As this dataset contains a large number of observations with correspondingly more variational parameters to be optimized, the computation is more intensive and parallel computing comes in very useful reducing the computation time by factors of 1.8, 2.9 and 4.2 for $K = 5, 20, 100$ respectively (Table 4).

Figure 4 shows the estimated marginal posteriors of the global parameters. CSGVA provides significant improvements in estimating the posterior variance of α and ψ as compared to GVA. With the application of IW-CSGVA, we are able to obtain posterior estimates that are quite close to that of MCMC even with a small K.

Table 4 GBP data. Number of iterations I (in thousands), runtimes (in s) and estimates of lower bound (SD in brackets) of the variational methods

	K	I	Time	parallel	$\hat{\mathcal{L}}_K^{\mathrm{IW}}$
GVA	1	61	239.7	–	−138.2 (1.3)
CSGVA	1	16	58.6	–	−137.8 (1.3)
IW-CSGVA	5	1	18.3	10.2	−137.4 (1.0)
	20	1	71.2	24.4	−137.0 (0.5)
	100	1	355.3	84.3	−136.8 (0.4)

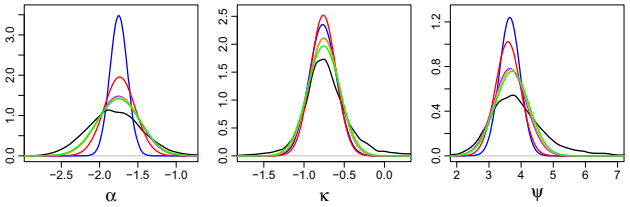

Fig. 4 GBP data. Marginal posterior distributions of global parameters. Black (MCMC), blue (GVA), red (CSGVA), purple ($K = 5$), orange ($K = 20$), green ($K = 100$). (Color figure online)

Figure 5 shows the estimated marginal posteriors of the latent states $\{b_i\}$ summarized using the mean (solid line) and one standard deviation from the mean (dotted line) estimated by MCMC (black) and IW-CSGVA ($K = 5$, purple). For IW-CSGVA, the means and standard deviations are estimated based on 2000 samples, by generating θ_G from $q(\theta_G)$ followed by θ_L from $q(\theta_L|\theta_G)$. For MCMC, estimation was based on 5000 samples. IW-CSGVA estimated the means quite accurately (with a little overestimation), but the standard deviations are underestimated when compared to MCMC. The boxplot shows the difference between the means and standard deviations estimated by IW-CSGVA ($K = 5$) and GVA with MCMC. We can see that IW-CSGVA estimated both the means and standard deviations more accurately as compared to GVA.

9.2 New York stock exchange data

Next we consider 2000 observations of the returns of the New York Stock Exchange (NYSE) from February 2, 1984 to December 31, 1991. These data are available as the dataset "nyse" from the `R` package `astsa`. We consider 100 times the mean-corrected returns as responses $\{y_i\}$.

From Table 5, CSGVA was able to attain a higher lower bound than GVA when initialized in the standard manner using $\lambda = 0$. Applying IW-CSGVA led to further improvements as K increases. For this massive data set, parallel computing yields significant reductions in computation time, by factors of 2.9, 4.5 and 5.5 corresponding to $K = 5$, 20, 100, respectively.

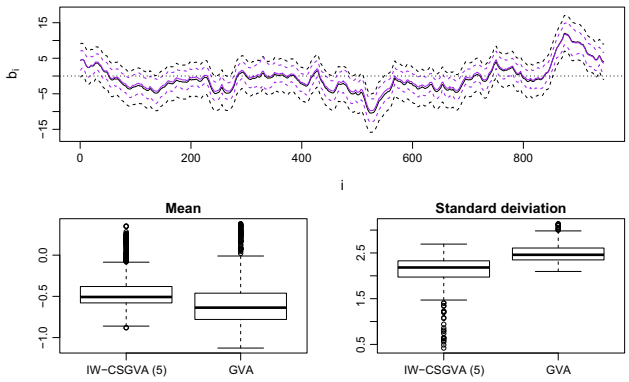

Fig. 5 GBP data. Top: Posterior means (solid line) of the latent states and one standard deviation from the mean (dotted line) estimated using MCMC (black) and IW-CSGVA ($K = 5$, purple). Bottom: Boxplots of $\mathrm{mean}_{\mathrm{MCMC}} - \mathrm{mean}_{\mathrm{VA}}$ and $\mathrm{sd}_{\mathrm{MCMC}} - \mathrm{sd}_{\mathrm{VA}}$. (Color figure online)

Table 5 NYSE data. Number of iterations I (in thousands), runtimes (in s) and estimates of lower bound (SD in brackets) of the variational methods

	K	I	Time	Parallel	$\hat{\mathcal{L}}_K^{\mathrm{IW}}$
GVA	1	43	679.0	–	−570.8 (1.8)
CSGVA	1	49	749.2	–	−570.7 (2.0)
IW-CSGVA	5	1	76.0	26.1	−569.4 (1.1)
	20	1	305.0	67.9	−569.0 (0.7)
	100	1	1503.0	274.0	−568.7 (0.4)

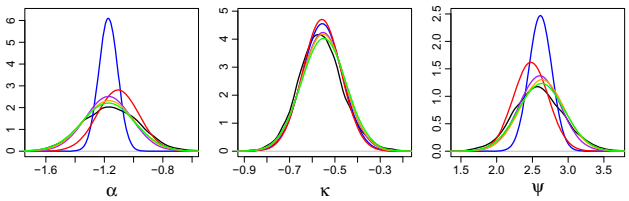

Fig. 6 NYSE data. Marginal posterior distributions of global parameters. Black (MCMC), blue (GVA), red (CSGVA), purple ($K = 5$), orange ($K = 20$), green ($K = 100$). (Color figure online)

Figure 6 shows that the marginal posteriors estimated using CSGVA are quite close to that of MCMC, while GVA underestimated the posterior variance of α and ψ quite severely. Posteriors estimated by IW-CSGVA are virtually indistinguishable from MCMC.

From Fig. 7, we can see that the marginal posteriors of the latent states are also estimated very well using IW-CSGVA ($K = 5$), and there is no systematic underestimation of the posterior variance unlike the previous example. GVA captures the posterior means very well but did not perform as well as IW-CSGVA in estimating the posterior variance.

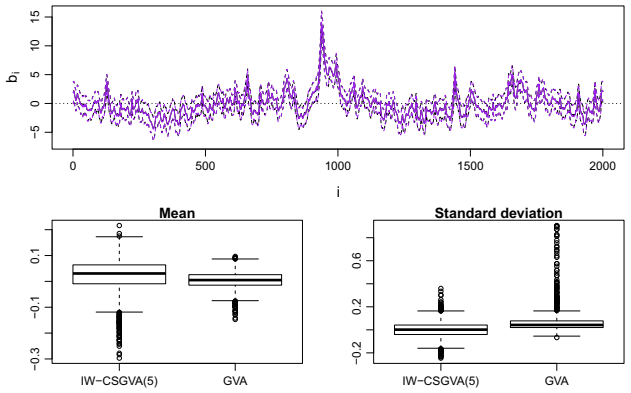

Fig. 7 NYSE data. Top: Posterior means (solid line) of the latent states and one standard deviation from the mean (dotted line) estimated using MCMC (black) and IW-CSGVA ($K = 5$, purple). Bottom: Boxplots of $\text{mean}_{\text{MCMC}} - \text{mean}_{\text{VA}}$ and $\text{sd}_{\text{MCMC}} - \text{sd}_{\text{VA}}$. (Color figure online)

10 Conclusion

In this article, we have proposed a Gaussian variational approximation for hierarchical models that adopts a conditional structure $q(\theta) = q(\theta_G)q(\theta_L|\theta_G)$. The dependence of the local variables θ_L on global variables θ_G are then captured using a linear approximation. This structure is very useful when there are global scale parameters in θ_G which help to determine the scale of local variables in the conditional posterior of $\theta_L|\theta_G$. We further demonstrate how CSGVA can be improved by maximizing the importance weighted lower bound. From our experiments, using a K as small as 5 can lead to significant improvements in the variational approximation, with just a short run. Moreover, for massive datasets, computation time can be further reduced through parallel computing. Our experiments indicate that CSGVA coupled with importance weighting is particularly useful in improving the estimation of the posterior variance of precision parameters ω in GLMMs, and the persistence ϕ and volatility σ of the log-variance in SVMs.

Acknowledgements We wish to thank the editor and reviewer for their time in reviewing this manuscript and for their constructive comments.

Appendix A: Derivation of stochastic gradient

Let $\otimes$ denote the Kronecker product between any two matrices. We have

$$r_\lambda(s) = \begin{bmatrix}\theta_G \\ \theta_L\end{bmatrix} = \begin{bmatrix}\mu_1 + C_1^{-T}s_1 \\ d + C_2^{-T}(s_2 - DC_1^{-T}s_1)\end{bmatrix},$$

where $v(C_2^*) = f + F(\mu_1 + C_1^{-T}s_1)$. Differentiating $r_\lambda(s)$ with respect to λ, $\nabla_\lambda r_\lambda(s)$ is given by

$$\begin{bmatrix} \nabla_{\mu_1}\theta_G & \nabla_{\mu_1}\theta_L \\ \nabla_{v(C_1^*)}\theta_G & \nabla_{v(C_1^*)}\theta_L \\ \nabla_d\theta_G & \nabla_d\theta_L \\ \nabla_{\text{vec}(D)}\theta_G & \nabla_{\text{vec}(D)}\theta_L \\ \nabla_f\theta_G & \nabla_f\theta_L \\ \nabla_{\text{vec}(F)}\theta_G & \nabla_{\text{vec}(F)}\theta_L \end{bmatrix}.$$

Since θ_G does not depend on d, D, f and F, we have

$$\nabla_d\theta_G = 0_{nL\times G}, \quad \nabla_{\text{vec}(D)}\theta_G = 0_{nLG\times G}$$
$$\nabla_f\theta_G = 0_{nL(nL+1)/2\times G}, \quad \nabla_{\text{vec}(F)}\theta_G = 0_{nLG(nL+1)/2\times G}.$$

It is easy to see that $\nabla_{\mu_1}\theta_G = I_G$ and $\nabla_d\theta_L = I_{nL}$. The rest of the terms are derived as follows.

Differentiating θ_G with respect to $v(C_1^*)$,

$$\begin{aligned} \mathrm{d}\theta_G &= -C_1^{-T}\mathrm{d}(C_1^T)C_1^{-T}s_1 \\ &= -(s_1^TC_1^{-1}\otimes C_1^{-T})K_GE_G^TD_1^*\mathrm{d}v(C_1^*) \\ &= -(C_1^{-T}\otimes s_1^TC_1^{-1})E_G^TD_1^*\mathrm{d}v(C_1^*). \\ \therefore\ \nabla_{v(C_1^*)}\theta_G &= -D_1^*E_G(C_1^{-1}\otimes C_1^{-T}s_1). \end{aligned}$$

Differentiating θ_L with respect to f,

$$\begin{aligned} \mathrm{d}\theta_L &= -C_2^{-T}\mathrm{d}(C_2^T)C_2^{-T}(s_2 - DC_1^{-T}s_1) \\ &= -\{(s_2 - DC_1^{-T}s_1)^TC_2^{-1}\otimes C_2^{-T}\} \\ &\quad \times K_{nL}E_{nL}^TD_2^*\mathrm{d}f \\ \therefore\ \nabla_f\theta_L &= -D_2^*E_{nL}\{C_2^{-1}\otimes C_2^{-T}(s_2 - DC_1^{-T}s_1)\}. \end{aligned}$$

Differentiating θ_L with respect to F,

$$\begin{aligned} \mathrm{d}\theta_L &= (\nabla_f\theta_L)^T\mathrm{d}F\theta_G \\ &= \{\theta_G^T\otimes(\nabla_f\theta_L)^T\}\mathrm{dvec}(F). \\ \therefore\ \nabla_{\text{vec}(F)}\theta_L &= \theta_G\otimes\nabla_f\theta_L. \end{aligned}$$

Differentiating θ_L with respect to D,

$$\begin{aligned} \mathrm{d}\theta_L &= -C_2^{-T}\mathrm{d}DC_1^{-T}s_1 \\ &= -(s_1^TC_1^{-1}\otimes C_2^{-T})\mathrm{dvec}(D). \\ \therefore\ \nabla_{\text{vec}(D)}\theta_L &= -(C_1^{-T}s_1\otimes C_2^{-1}). \end{aligned}$$

Differentiating θ_L with respect to μ_1,

$$\begin{aligned} \mathrm{d}\theta_L &= (\nabla_f\theta_L)^TF\mathrm{d}\mu_1 \\ \therefore\ \nabla_{\mu_1}\theta_L &= F^T(\nabla_f\theta_L). \end{aligned}$$

Differentiating θ_L with respect to $v(C_1)$,

$$\begin{aligned}
\mathrm{d}\theta_L &= -C_2^{-T}\mathrm{d}(C_2^T)C_2^{-T}(s_2 - DC_1^{-T}s_1) \\
&\quad - C_2^{-T}D\mathrm{d}(C_1^{-T})s_1 \\
&= (\nabla_f\theta_L)^T F\mathrm{d}(C_1^{-T})s_1 - C_2^{-T}D\mathrm{d}(C_1^{-T})s_1 \\
&= \{(\nabla_f\theta_L)^T F - C_2^{-T}D\}(\nabla_{v(C_1^*)}\theta_G)^T\mathrm{d}v(C_1^*) \\
\therefore\ \nabla_{v(C_1^*)}\theta_L &= \nabla_{v(C_1^*)}\theta_G\{F^T\nabla_f\theta_L - D^TC_2^{-1}\} \\
&= \nabla_{v(C_1^*)}\theta_G\{\nabla_{\mu_1}\theta_L - D^TC_2^{-1}\}.
\end{aligned}$$

Since $s_1 = C_1^T(\theta_G - \mu_1)$ and $s_2 = C_2^T(\theta_L - \mu_2)$, we have

$$\begin{aligned}
\log q_\lambda(\theta) &= \log q(\theta_G) + \log q(\theta_L|\theta_G) \\
&= -\frac{G}{2}\log(2\pi) + \log|C_1| \\
&\quad - \frac{1}{2}(\theta_G - \mu_1)^T C_1C_1^T(\theta_G - \mu_1) \\
&\quad - \frac{nL}{2}\log(2\pi) + \log|C_2| \\
&\quad - \frac{1}{2}(\theta_L - \mu_2)^T C_2C_2^T(\theta_L - \mu_2) \\
&= -\frac{nL+G}{2}\log(2\pi) + \log|C_1C_2| - \frac{1}{2}s^Ts.
\end{aligned}$$

As $\mu_2 = d + C_2^{-T}D(\mu_1 - \theta_G)$ and $v(C_2^*) = f + F\theta_G$, differentiating $\log q_\lambda(\theta)$ with respect to θ_G,

$$\begin{aligned}
\mathrm{d}\log q_\lambda(\theta) &= -(\theta_G - \mu_1)^T C_1C_1^T\mathrm{d}\theta_G - (\theta_L - \mu_2)^T C_2C_2^T(-\mathrm{d}\mu_2) \\
&\quad - (\theta_L - \mu_2)^T\mathrm{d}C_2s_2 + \mathrm{tr}(C_2^{-1}\mathrm{d}C_2) \\
&= -s_1^TC_1^T\mathrm{d}\theta_G + s_2^TC_2^T\{-C_2^{-T}D\mathrm{d}\theta_G \\
&\quad + \mathrm{d}(C_2^{-T})D(\mu_1 - \theta_G)\} \\
&\quad - \mathrm{vec}(C_2^{-T}s_2s_2^T)^T\mathrm{dvec}(C_2) + \mathrm{vec}(C_2^{-T})^T d\mathrm{vec}(C_2) \\
&= \mathrm{vec}(C_2^{-T} - \{C_2^{-T}s_2 + (\mu_2 - d)\}s_2^T)^T d\mathrm{vec}(C_2) \\
&\quad - s_1^TC_1^T\mathrm{d}\theta_G - s_2^TD\mathrm{d}\theta_G \\
&= \mathrm{vec}(C_2^{-T} - (\theta_L - d)s_2^T)^T E_{nL}^TD_2^*F\mathrm{d}\theta_G \\
&\quad - s_1^TC_1^T\mathrm{d}\theta_G - s_2^TD\mathrm{d}\theta_G.
\end{aligned}$$

Therefore

$$\begin{aligned}
\nabla_{\theta_G}\log q_\lambda(\theta) &= F^TD_2^*v(C_2^{-T} - (\theta_L - d)s_2^T) \\
&\quad - C_1s_1 - D^Ts_2.
\end{aligned}$$

Note that $D_2^*v(C_2^{-T}) = v(I_{nL})$ as C_2^{-T} is upper triangular and $v(C_2^{-T})$ only retains the diagonal elements of C_2^{-T}.

Differentiating $\log q_\lambda(\theta)$ with respect to θ_L,

$$\begin{aligned}
\mathrm{d}\log q_\lambda(\theta) &= -(\theta_L - \mu_2)^TC_2C_2^T\mathrm{d}\theta_L \\
&= -s_2^TC_2^T\mathrm{d}\theta_L. \\
\therefore\ \nabla_{\theta_L}\log q_\lambda(\theta) &= -C_2s_2.
\end{aligned}$$

Appendix B: Gradients for generalized linear mixed models

Since $\theta = [\beta^T, \omega^T, \tilde{b}_1^T, \ldots, \tilde{b}_n^T]^T$, we require

$$\begin{aligned}
\nabla_\theta\log p(y,\theta) &= [\nabla_\beta\log p(y,\theta), \nabla_\omega\log p(y,\theta), \\
&\quad \nabla_{\tilde{b}_1}\log p(y,\theta), \ldots, \nabla_{\tilde{b}_n}\log p(y,\theta)]^T.
\end{aligned}$$

For the centered parametrization, the components in $\nabla_\theta\log p(y,\theta)$ are given below. Note that $\beta = [\beta_{RG_1}^T, \beta_{G_2}^T]^T$.

$$\begin{aligned}
\nabla_{\beta_{G_2}}\log p(y,\theta) &= \sum_{i=1}^n X_i^{G_2T}\{y_i - h'(\eta_i)\} - \beta_{G_2}/\sigma_\beta^2. \\
\nabla_{\beta_{RG_1}}\log p(y,\theta) &= \sum_{i=1}^n C_i^TWW^T(\tilde{b}_i - C_i\beta_{RG_1}) \\
&\quad - \beta_{RG_1}/\sigma_\beta^2.
\end{aligned}$$

Differentiating $\log p(y,\theta)$ with respect to ω,

$$\begin{aligned}
\mathrm{d}\log p(y,\theta) &= -\sum_{i=1}^n(\tilde{b}_i - C_i\beta_{RG_1})^T\mathrm{d}WW^T(\tilde{b}_i - C_i\beta_{RG_1}) \\
&\quad + n\mathrm{tr}(W^{-1}\mathrm{d}W) - \omega^T\mathrm{d}\omega/\sigma_\omega^2 \\
&= \mathrm{vec}\Big\{-\sum_{i=1}^n(\tilde{b}_i - C_i\beta_{RG_1})(\tilde{b}_i - C_i\beta_{RG_1})^TW \\
&\quad + nW^{-T}\Big\}^T E_L^TD_L^*\mathrm{d}\omega - \omega^T\mathrm{d}\omega/\sigma_\omega^2,
\end{aligned}$$

where $\mathrm{d}v(W) = D_L^*\mathrm{d}\omega$ and $D_L^* = \mathrm{diag}\{v(\mathrm{dg}(W) + \mathbf{1}_L\mathbf{1}_L^T - I_L)\}$. Hence

$$\begin{aligned}
\nabla_\omega\log p(y,\theta) &= -D_L^*\sum_{i=1}^n v\{(\tilde{b}_i - C_i\beta_{RG_1})(\tilde{b}_i \\
&\quad - C_i\beta_{RG_1})^TW\} \\
&\quad + nv(I_L) - \omega/\sigma_\omega^2.
\end{aligned}$$

Note that $D_L^*v(W^{-T}) = v(I_L)$ because W^{-T} is upper triangular and $v(W^{-T})$ only retains the diagonal elements.

$$\nabla_{\tilde{b}_i}\log p(y,\theta) = Z_i^T\{y_i - h'(\eta_i)\} - WW^T(\tilde{b}_i - C_i\beta_{RG_1}).$$

Appendix C: Gradients for state space models

Since $\theta = [\alpha, \kappa, \psi, b_1^T, \ldots, b_n^T]^T$, we require

$$\nabla_\theta\log p(y,\theta) = [\nabla_\alpha\log p(y,\theta), \nabla_\kappa\log p(y,\theta),$$

$$\nabla_\psi \log p(y,\theta), \nabla_{b_1} \log p(y,\theta), \ldots, \nabla_{b_n} \log p(y,\theta)]^T.$$

The components in $\nabla_\theta \log p(y,\theta)$ are given below.

$$\nabla_\alpha \log p(y,\theta) = \frac{1}{2}\sum_{i=1}^{n}(b_i y_i^2 \mathrm{e}^{-\sigma b_i-\kappa} - b_i)(1-\mathrm{e}^{-\sigma}) - \frac{\alpha}{\sigma_\alpha^2}.$$

$$\nabla_\kappa \log p(y,\theta) = \frac{1}{2}\left(\sum_{i=1}^{n} y_i^2 \mathrm{e}^{-\sigma b_i-\kappa} - n\right) - \kappa/\sigma_\kappa^2.$$

$$\nabla_\psi \log p(y,\theta) = \left\{\sum_{i=2}^{n}(b_i - \phi b_{i-1})b_{i-1} + b_1^2\phi - \frac{\phi}{1-\phi^2}\right\} \times\phi(1-\phi) - \psi/\sigma_\psi^2.$$

$$\nabla_{b_1} \log p(y,\theta) = \frac{\sigma}{2}(y_1^2 \mathrm{e}^{-\sigma b_1-\kappa} - 1) + \phi(b_2 - \phi b_1) - b_1(1-\phi)^2.$$

For $2 \le i \le n-1$,

$$\nabla_{b_i} \log p(y,\theta) = \frac{\sigma}{2}(y_i^2 \mathrm{e}^{-\sigma b_i-\kappa} - 1) + \phi(b_{i+1} - \phi b_i) - (b_i - \phi b_{i-1}).$$

$$\nabla_{b_n} \log p(y,\theta) = \frac{\sigma}{2}(y_n^2 \mathrm{e}^{-\sigma b_n-\kappa} - 1) - (b_n - \phi b_{n-1}).$$

References

Archer, E., Park, I.M., Buesing, L., Cunningham, J., Paninski, L.: Black box variational inference for state space models (2016). arXiv:1511.07367

Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**, 859–877 (2017)

Breslow, N.E., Clayton, D.G.: Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. **88**, 9–25 (1993)

Burda, Y., Grosse, R., Salakhutdinov, R.: Importance weighted autoencoders. In: Proceedings of the 4th International Conference on Learning Representations (ICLR) (2016)

Diggle, P.J., Heagerty, P., Liang, K.Y., Zeger, S.L.: The Analysis of Longitudinal Data, 2nd edn. Oxford University Press, Oxford (2002)

Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real NVP. In: Proceedings of the 5th International Conference on Learning Representations (ICLR) (2017)

Domke, J., Sheldon, D.R.: Importance weighting and variational inference. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 4470–4479. Curran Associates, Inc., New York (2018)

Fitzmaurice, G.M., Laird, N.M.: A likelihood-based method for analysing longitudinal binary responses. Biometrika **80**, 141–151 (1993)

Germain, M., Gregor, K., Murray, I., Larochelle, H.: MADE: masked autoencoder for distribution estimation. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning, PMLR, Lille, France, Proceedings of Machine Learning Research, vol. 37, pp. 881–889 (2015)

Guo, F., Wang, X., Broderick, T., Dunson, D.B.: Boosting variational inference (2016). arXiv: 1611.05559

Han, S., Liao, X., Dunson, D.B., Carin, L.C.: Variational Gaussian copula inference. In: Gretton, A., Robert, C.C. (eds.) Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, vol. 51, pp. 829–838 (2016)

Hoffman, M., Blei, D.: Stochastic structured variational inference. In: Lebanon, G., Vishwanathan, S. (eds.) Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, vol. 38, pp. 361–369 (2015)

Huszár, F.: Variational inference using implicit distributions (2017). arXiv:1702.08235

Jaakkola, T.S., Jordan, M.I.: Improving the mean field approximation via the use of mixture distributions, pp. 163–173. Springer, Dordrecht (1998)

Kastner, G., Frühwirth-Schnatter, S.: Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. Comput. Stat. Data Anal. **76**, 408–423 (2014)

Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR) (2015)

Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: Proceedings of the 2nd International Conference on Learning Representations (ICLR) (2014)

Kingma, D.P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., Welling, M.: Improved variational inference with inverse autoregressive flow. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29, pp. 4743–4751. Curran Associates, Inc., New York (2016)

Li, Y., Turner, R.E.: Rényi divergence variational inference. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, Curran Associates Inc., USA, NIPS'16, pp. 1081–1089 (2016)

Maddison, C.J., Lawson, J., Tucker, G., Heess, N., Norouzi, M., Mnih, A., Doucet, A., Teh, Y.: Filtering variational objectives. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 6573–6583. Curran Associates, Inc, New York (2017)

Magnus, J.R., Neudecker, H.: The elimination matrix: some lemmas and applications. SIAM J. Algebr. Discrete Methods **1**, 422–449 (1980)

Magnus, J.R., Neudecker, H.: Matrix Differential Calculus with Applications in Statistics and Econometrics, 3rd edn. Wiley, New York (1999)

Miller, A.C., Foti, N., Adams, R.P.: Variational boosting: iteratively refining posterior approximations (2016). arXiv: 1611.06585

Minka, T.: Divergence measures and message passing. Technical report (2005)

Ormerod, J.T., Wand, M.P.: Explaining variational approximations. Am. Stat. **64**, 140–153 (2010)

Papaspiliopoulos, O., Roberts, G.O., Sköld, M.: Non-centered parameterisations for hierarchical models and data augmentation. In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M. (eds.) Bayesian Statistics 7, pp. 307–326. Oxford University Press, New York (2003)

Papaspiliopoulos, O., Roberts, G.O., Sköld, M.: A general framework for the parametrization of hierarchical models. Stat Sci **22**, 59–73 (2007)

Ranganath, R., Tran, D., Blei, D.M.: Hierarchical variational models. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning, JMLR Workshop and Conference Proceedings, vol. 37, pp. 324–333 (2016)

Regli, J.B., Silva, R.: Alpha-beta divergence for variational inference (2018). arXiv: 1805.01045

Rényi, A.: On measures of entropy and information. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics

and Probability, Volume 1: Contributions to the Theory of Statistics, University of California Press, Berkeley, Calif., pp. 547–561 (1961)

Rezende, D.J., Mohamed, S.: Variational inference with normalizing flows. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning, JMLR Workshop and Conference Proceedings, vol. 37, pp. 1530–1538 (2015)

Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: Xing, E.P., Jebara, T. (eds.) Proceedings of the 31st International Conference on Machine Learning, JMLR Workshop and Conference Proceedings, vol. 32, pp. 1278–1286 (2014)

Roeder, G., Wu, Y., Duvenaud, D.K.: Sticking the landing: simple, lower-variance gradient estimators for variational inference. In: Guyon, I., Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30 (2017)

Roeder, G., Grant, P.K., Phillips, A., Dalchau, N., Meeds, E.: Efficient amortised bayesian inference for hierarchical and nonlinear dynamical systems. Proc. Mach. Learn. Res. **97**, 4445–4455 (2019)

Rothman, A.J., Levina, E., Zhu, J.: A new approach to Cholesky-based covariance regularization in high dimensions. Biometrika **97**, 539–550 (2010)

Salimans, T., Knowles, D.A.: Fixed-form variational posterior approximation through stochastic linear regression. Bayesian Anal. **8**, 837–882 (2013)

Smith, M.S., Loaiza-Maya, R., Nott, D.J.: High-dimensional copula variational approximation through transformation (2019). arXiv:1904.07495

Spall, J.C.: Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control. Wiley, New York (2003)

Spantini, A., Bigoni, D., Marzouk, Y.: Inference via low-dimensional couplings. J. Mach. Learn. Res. **19**, 1–71 (2018)

Tan, L.S.L.: Efficient data augmentation techniques for Gaussian state space models (2017). arXiv:1712.08887

Tan, L.S.L.: Use of model reparametrization to improve variational Bayes (2018). arXiv:1805.07267

Tan, L.S.L., Nott, D.J.: Variational inference for generalized linear mixed models using partially non-centered parametrizations. Stat. Sci. **28**, 168–188 (2013)

Tan, L.S.L., Nott, D.J.: Gaussian variational approximation with sparse precision matrices. Stat. Comput. **28**, 259–275 (2018)

Thall, P.F., Vail, S.C.: Some covariance models for longitudinal count data with overdispersion. Biometrics **46**, 657–671 (1990)

Titsias, M., Lázaro-Gredilla, M.: Doubly stochastic variational Bayes for non-conjugate inference. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 1971–1979 (2014)

Tran, D., Blei, D.M., Airoldi, E.M.: Copula variational inference. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7–12, 2015, Montreal, Quebec, Canada, pp. 3564–3572 (2015)

Tucker, G., Lawson, D., Gu, S., Maddison, C.J.: Doubly reparametrized gradient estimators for Monte Carlo objectives (2018). arXiv:1810.04152

van Erven, T., Harremos, P.: Rnyi divergence and Kullback–Leibler divergence. IEEE Trans. Inf. Theory **60**, 3797–3820 (2014)

Yang, Y., Pati, D., Bhattacharya, A.: α-variational inference with statistical guarantees. Ann. Stat. (2019) (to appear)