Bayesian Analysis (2023) **18**, Number 3, pp. 989–1016

Gaussian Variational Approximations for High-dimensional State Space Models*

Matias Quiroz[†,‡], David J. Nott[§,¶] and Robert Kohn[‖,**]

Abstract. We consider a Gaussian variational approximation of the posterior density in high-dimensional state space models. The number of parameters in the covariance matrix of the variational approximation grows as the square of the number of model parameters, so it is necessary to find simple yet effective parametrisations of the covariance structure when the number of model parameters is large. We approximate the joint posterior density of the state vectors by a dynamic factor model, having Markovian time dependence and a factor covariance structure for the states. This gives a reduced description of the dependence structure for the states, as well as a temporal conditional independence structure similar to that in the true posterior. We illustrate the methodology on two examples. The first is a spatio-temporal model for the spread of the Eurasian collared-dove across North America. Our approach compares favorably to a recently proposed ensemble Kalman filter method for approximate inference in high-dimensional hierarchical spatio-temporal models. Our second example is a Wishart-based multivariate stochastic volatility model for financial returns, which is outside the class of models the ensemble Kalman filter method can handle.

Keywords: dynamic factor, stochastic gradient, spatio-temporal modelling.

1 Introduction

Variational approximation (VA) (Ormerod and Wand, 2010; Blei et al., 2017) replaces the true posterior density by a parametric density whose parameters optimise a measure of closeness to the true posterior. A frequent choice for the approximation is a multivariate Gaussian distribution, where the variational optimisation is over an unknown mean and covariance matrix. VA is an increasingly popular way to approximate the pos-

arXiv: 1801.07873

*Matias Quiroz and Robert Kohn were partially supported by Australian Research Council Center of Excellence grant CE140100049. David Nott was supported by a Singapore Ministry of Education Academic Research Fund Tier 2 grant (MOE2016-T2-2-135).

†Department of Statistics, Stockholm University, 106 91 Stockholm, Sweden, matias.quiroz@stat.su.se

‡ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), Melbourne VIC 3010, Australia, matias.quiroz@stat.su.se

§Department of Statistics and Data Science, National University of Singapore, Singapore 117546, standj@nus.edu.sg

¶Institute of Operations Research and Analytics, National University of Singapore, 21 Lower Kent Ridge Road, Singapore 119077

‖UNSW Business School, School of Economics, University of New South Wales, Sydney NSW 2052, Australia, r.kohn@unsw.edu.au

**ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), Melbourne VIC 3010, Australia, r.kohn@unsw.edu.au

 https://doi.org/10.1214/22-BA1332

terior because of its ability to handle large datasets and highly parametrised models. The accuracy of the VA depends on a number of factors, such as the flexibility of the approximating family, the model considered, and the sample size. There are now some theoretical results which show that the variational posterior converges to the true parameter value under suitable regularity conditions, and rates of convergence have been established for parametric models (Wang and Blei, 2019) and, more generally, for non-parametric and high-dimensional models (Zhang and Gao, 2020). However, for a finite number of observations, when the variational approximation does not collapse to a point mass, it is often observed that there is a practically meaningful discrepancy between the uncertainty quantification provided by the approximation and that of the true posterior distribution. This is especially the case when the variational family used is insufficiently flexible. Nevertheless, even in these cases, predictions and prediction intervals obtained from VA seem empirically to be usefully close to those obtained from the exact posterior. In the context of Gibbs posteriors (Zhang, 2006), i.e. when posteriors are formed via a loss function approach, Frazier et al. (2021b) show that the discrepancy between predictive inference based on the variational posterior and that of the true posterior is asymptotically zero. Frazier et al. (2021a) demonstrate empirically the accuracy of the predictive distribution for a range of variational approximation methods for state space models, including ours. As such, variational approximation methods provide a useful and fast alternative to Markov chain Monte Carlo (MCMC), especially when predictive inference is required.

Our article considers a Gaussian variational approximation (GVA) for a state space model when the state vector is high-dimensional. Such models are common in spatio-temporal applications (Cressie and Wikle, 2011), financial econometrics (Philipov and Glickman, 2006), and in other important applications. It is challenging to obtain the GVA when dealing with a high-dimensional model, because the number of variational parameters in the covariance matrix of the approximation grows quadratically with the number of model parameters. This makes it necessary to parametrise the variational covariance matrix parsimoniously, but still be able to capture the dependence structure of the posterior. This goal is best achieved by taking into account the dependence structure of the posterior itself. We do so by parametrising the variational posterior covariance matrix using a dynamic factor model, which reduces the dimension of the state vector. The Markovian time dependence for the low-dimensional factors provides the necessary sparsity in the precision matrix for the factors. We develop efficient computational methods for forming the approximations and illustrate the advantages of the approach in two example datasets with high-dimensional state vectors, where Bayesian inference by MCMC simulation is challenging for both models. The first is a spatio-temporal model for the spread of the Eurasian collared dove across North America (Wikle and Hooten, 2006); the second is a multivariate stochastic volatility model for a portfolio of assets (Philipov and Glickman, 2006). We derive GVAs for both models and demonstrate that they give useful predictive inference.

As noted above, Bayesian computation for complex state space models with a high-dimensional state vector is challenging. In this context, the use of ensemble Kalman filtering and smoothing methods (Evensen, 1994; Evensen and Van Leeuwen, 2000) provides an alternative and scalable approach to approximating the distribution of the

states. We refer to sampling algorithms using these methods as ensemble Kalman filter methods, by which we mean that an ensemble Kalman filter is used for filtering and an ensemble Kalman smoother for smoothing. Katzfuss et al. (2020) use ensemble Kalman filter methods to sample the posterior distribution of hierarchical dynamic spatio-temporal models. We later compare our method to their so-called Gibbs ensemble Kalman smoother in our spatio-temporal example and find that our method is a faster alternative with similar inferential performance. Importantly, the strength of our approach is that it is an all-purpose method that applies to general high-dimensional state space models. While ensemble Kalman filter methods can handle a non-Gaussian distribution of the measurement equation, the distribution of the state transition equation is assumed to be Gaussian. Hence it cannot be applied to our second example (multivariate stochastic volatility), where the distribution of the state transition is Wishart.

The paper has an appendix and a supplement, both are web-based (Quiroz et al., 2022). The appendix contains the equations to implement the gradients of the method, and the supplement contains proofs and other material. We refer to equations, sections, etc. in the main paper as (1.1), Section 1, etc., in the appendix as (A1.1), Appendix A, etc., and in the web-based supplement as (S1.1), Section S1, etc.

2 Methodology

To state some of the results below we use conventional matrix calculus notation. Section 6 defines the notation for readers unfamiliar with it.

2.1 Model, prior and posterior

Let $y = (y_1, \ldots, y_T)^\top$ be an observed time series generated by the state space model

$$y_t | X_t = x_t, \zeta \sim m_t(y|x_t, \zeta), \quad t = 1, \ldots, T, \tag{2.1a}$$

$$X_t | X_{t-1} = x_{t-1}, \zeta \sim s_t(x|x_{t-1}, \zeta), \quad t = 1, \ldots, T; \tag{2.1b}$$

the prior density for X_0 is $p(X_0|\zeta)$, ζ are the unknown fixed (non-time-varying) parameters in the model, and the elements of ζ in the measurement and the state equation are typically different, but the same symbol is used for brevity. The observations y_t are independent given ζ and the states $X = (X_0^\top, \ldots, X_T^\top)^\top$, and the prior distribution of X given ζ is

$$p(X|\zeta) = p(X_0|\zeta) \prod_{t=1}^{T} s_t(X_t|X_{t-1}, \zeta).$$

Let $\theta = (X^\top, \zeta^\top)^\top$ denote the full set of unknowns in the model. The posterior density of θ is $p(\theta|y) \propto p(\theta)p(y|\theta)$, with $p(\theta) = p(\zeta)p(X|\zeta)$, where $p(\zeta)$ is the prior density for ζ and $p(y|\theta) = \prod_{t=1}^{T} m_t(y_t|X_t, \zeta)$. Let p be the dimension of X_t and suppose p is large. Approximating the posterior distribution in this setting is difficult and we propose a method based on Gaussian variational approximation.

2.2 Example model: Multivariate stochastic volatility

To illustrate the discussion above, we consider the multivariate stochastic volatility model introduced by Philipov and Glickman (2006), which is useful for modeling the time-varying dependence of a portfolio of k assets over T time periods. Section 4 discusses the model in more detail.

Let $\mathbb{R}_+^{k\times k}$ denote the space of positive definite matrices of dimension $k \times k$. Philipov and Glickman (2006) assume that the mean-centred return at time period t, $t = 1, \ldots, T$, is the k dimensional vector y_t,

$$y_t \sim \mathcal{N}(0, \Sigma_t), \quad \Sigma_t \in \mathbb{R}_+^{k\times k}; \tag{2.2a}$$

$$\Sigma_t^{-1} \sim \text{Wishart}(\nu, S_{t-1}), \quad S_t = \frac{1}{\nu} H(\Sigma_t^{-1})^d H^\top, \; S_t \in \mathbb{R}_+^{k\times k}, \; \nu > k, \; 0 < d < 1; \tag{2.2b}$$

H is an unknown lower triangular Cholesky factor of $A = HH^\top \in \mathbb{R}_+^{k\times k}$, ν, d are unknown scalars and $\Sigma_0 \in \mathbb{R}_+^{k\times k}$ is known. Section 4 describes the priors for all the fixed parameters and states. The state vector $X_t = \text{vech}(\Sigma_t)$ is $k(k+1)/2$-dimensional. The dimensionality of X_t grows rapidly with k, e.g. it is 55 dimensional for $k = 10$. To carry out exact Bayesian inference based on Markov chain Monte Carlo it is necessary to use particle methods which do not scale to high-dimensional states (Katzfuss et al., 2020).

2.3 Stochastic gradient ascent variational methods

The main contribution of our article is to propose a parsimonious parametrisation of the covariance matrix of the approximating Gaussian variational density that captures dependence structures of the statistical model in (2.1); see Section 2.4. Before outlining this idea, we briefly describe how the optimal variational parameters are found given a parametrisation.

Variational approximation methods (Attias, 1999; Jordan et al., 1999; Winn and Bishop, 2005) express the problem of approximating an intractable posterior distribution as an optimisation problem. We consider a family of densities $\{q_\lambda(\theta)\}$, indexed by the variational parameter λ, to approximate $p(\theta|y)$. Our article takes the approximating family to be Gaussian so that λ consists of the mean vector and the distinct elements of the covariance matrix (as parametrised in the next subsection) in the approximating Gaussian density.

The optimisation problem is to find the variational parameters λ that minimise the Kullback-Leibler divergence between the variational approximation and the posterior density. This can be achieved by maximising the evidence lower bound (ELBO) (see e.g. Ormerod and Wand, 2010) given by

$$\mathcal{L}(\lambda) = \int \log \frac{p(\theta)p(y|\theta)}{q_\lambda(\theta)} q_\lambda(\theta) d\theta. \tag{2.3}$$

Since $\mathcal{L}(\lambda)$ is generally intractable, stochastic gradient methods (Robbins and Monro, 1951) are needed to perform the optimisation and there is now a large literature on

implementing them (Ji et al., 2010; Paisley et al., 2012; Nott et al., 2012; Salimans and Knowles, 2013; Kingma and Welling, 2014; Rezende et al., 2014; Hoffman et al., 2013; Ranganath et al., 2014; Titsias and Lázaro-Gredilla, 2015; Kucukelbir et al., 2017).

Stochastic gradient ascent methods start with an initial guess for the optimal value $\lambda^{(0)}$, which then gets updated according to the iterative scheme

$$\lambda^{(t+1)} = \lambda^{(t)} + a_t \widehat{\nabla_\lambda \mathcal{L}(\lambda^{(t)})}, \tag{2.4}$$

where a_t, $t \geq 0$, is a sequence of learning rates, $\nabla_\lambda \mathcal{L}(\lambda)$ is the gradient vector of $\mathcal{L}(\lambda)$ with respect to λ and $\widehat{\nabla_\lambda \mathcal{L}(\lambda)}$ denotes its unbiased estimate. The learning rate sequence is typically chosen to satisfy $\sum_t a_t = \infty$ and $\sum_t a_t^2 < \infty$, which ensures that the iterates $\lambda^{(t)}$ converge to a local optimum as $t \to \infty$ under suitable regularity conditions (Bottou, 2010). Various adaptive choices for the learning rates are also possible and we use the ADADELTA (Zeiler, 2012) approach in our applications in Sections 3 and 4.

Reducing the variance of the gradient estimates in (2.4) is important for both the stability of the algorithm and fast convergence. Our article uses gradient estimates based on the so-called reparametrisation trick (Kingma and Welling, 2014; Rezende et al., 2014) which is now outlined. The lower bound $\mathcal{L}(\lambda)$ in (2.3) is an expectation with respect to q_λ,

$$\mathcal{L}(\lambda) = E_q(\log h(\theta) - \log q_\lambda(\theta)), \tag{2.5}$$

where $E_q(\cdot)$ denotes expectation with respect to q_λ and $h(\theta) = p(\theta)p(y|\theta)$. Differentiating with respect to λ under the integral sign in (2.5), the resulting expression for the gradient can also be written as an expectation with respect to q_λ, which is easily estimated unbiasedly by Monte Carlo integration. However, this so-called score function method (Williams, 1992) typically results in a very large variance of the gradient estimator. The reparametrisation trick is often much more efficient (Xu et al., 2019) and we now illustrate the method using a full covariance matrix for simplicity. Suppose that we can write $\theta \sim q_\lambda(\theta)$ as $\theta = u(\lambda, \omega)$, where ω is a random vector with a density which does not depend on the variational parameters λ. For $q_\lambda(\theta) = \mathcal{N}(\mu, \Sigma)$, with $\Sigma = CC^\top$, where C is the (lower triangular) Cholesky factor of Σ, we can write $\theta = \mu + C\omega$, where $\omega \sim \mathcal{N}(0, I_d)$. Substituting $\theta = u(\lambda, \omega)$ into (2.5), we obtain

$$\mathcal{L}(\lambda) = E_\omega(\log h(u(\lambda, \omega)) - \log q_\lambda(u(\lambda, \omega))), \tag{2.6}$$

where E_ω is the expectation with respect to ω. Differentiating under the integral sign, we obtain

$$\nabla_\lambda \mathcal{L}(\lambda) = E_\omega(\nabla_\lambda \log h(u(\lambda, \omega)) - \nabla_\lambda \log q_\lambda(u(\lambda, \omega))), \tag{2.7}$$

which is easily estimated unbiasedly.

Section S2 discusses a further variance reduction technique proposed in Han et al. (2016), Tan and Nott (2018) and Roeder et al. (2017) that may be particularly effective when the variational family is flexible enough to accurately approximate the true posterior.

2.4 Structure of the proposed variational approximation

We derive the variational posterior density $q_\lambda(\theta)$ for $\theta = (X, \zeta)$ in three steps.

The first step approximates the posterior of X_t, and is based on a generative model which has the dynamic factor structure,

$$X_t = Bz_t + \epsilon_t \quad \epsilon_t \sim \mathcal{N}(0, D_t^2), \tag{2.8}$$

where B is a $p \times q$ matrix with $B_{ij} = 0$ for $i < j$, $q \ll p$, and D_t is a diagonal matrix with diagonal elements $\delta_t = (\delta_{t1}, \ldots, \delta_{tp})^\top$. Hence X_t of dimension p is approximated via a vector z_t of lower dimension q as in (2.8).

The second step approximates the joint posterior of the resulting low-dimensional state vector $z = (z_0^\top, \ldots, z_T^\top)^\top$ and ζ. We assume that $\rho = (z^\top, \zeta^\top) \sim \mathcal{N}(\mu, \Sigma)$, $\Sigma = C^{-\top}C^{-1}$ where C is the Cholesky factor of the precision matrix of ρ. For computational tractability, we further assume that

$$C = \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix},$$

is block diagonal, C_1 is the Cholesky factor of the precision matrix $\Omega_1 = C_1 C_1^\top$ for z, and C_2 is the Cholesky factor for the precision matrix of ζ. We note that this assumption implies independence between the two blocks. Let $\Sigma_1 = \Omega_1^{-1}$ denote the covariance matrix of z. We further assume that C_1 is lower triangular with a single band, implying that Ω_1 is band tridiagonal; see Section S3 for details. For a Gaussian distribution, zero elements in the precision matrix represent conditional independence relationships. In particular, the sparse structure imposed on C_1 means that in the generative distribution for ρ, the latent variable z_t, given z_{t-1} and z_{t+1}, is conditionally independent of the remaining elements of z; in other words, if we think of the variables z_t, $t = 1, \ldots, T$ as a time series, they have a Markovian dependence structure. Setting zero elements in C_2 follows the same principle and depends on the model; see Section 2.6.

The third step constructs the variational distribution for the full parameter vector θ through

$$\theta = \begin{bmatrix} X \\ \zeta \end{bmatrix} = \begin{bmatrix} I_{T+1} \otimes B & 0 \\ 0 & I_P \end{bmatrix} \rho + \begin{bmatrix} \epsilon \\ 0 \end{bmatrix},$$

P is the dimension of ζ, and $\epsilon = (\epsilon_0^\top, \ldots, \epsilon_T^\top)^\top$. We can apply the reparametrisation trick by writing $\rho = \mu + C^{-\top}\omega$, where $\omega \sim \mathcal{N}(0, I_{q(T+1)+P})$. Then,

$$\theta = W\rho + Ze = W\mu + WC^{-\top}\omega + Ze, \tag{2.9}$$

where

$$W = \begin{bmatrix} I_{T+1} \otimes B & 0_{p(T+1)\times P} \\ 0_{P\times q(T+1)} & I_P \end{bmatrix}, \quad Z = \begin{bmatrix} D & 0_{p(T+1)\times P} \\ 0_{P\times p(T+1)} & 0_{P\times P} \end{bmatrix}, \quad e = \begin{bmatrix} \epsilon \\ 0_{P\times 1} \end{bmatrix},$$

D is a diagonal matrix with diagonal entries $(\delta_0^\top, \ldots, \delta_T^\top)^\top$, and $u = (\omega^\top, \epsilon^\top)^\top \sim \mathcal{N}(0, I_{(p+q)(T+1)+P})$. We also write $\omega = (\omega_1^\top, \omega_2^\top)^\top$, where the blocks of this partition

follow those of $\rho = (z^\top, \zeta^\top)^\top$. We note that with the notation above, the variational approximation density has the explicit form

$$q_\lambda(\theta) = \mathcal{N}\left(\theta | W\mu, WC^{-\top}C^{-1}W^\top + Z^2\right). \tag{2.10}$$

The factor model above describes the covariance structure for the states, as well as for dimension reduction in the variational posterior mean of the states, since $E(X_t) = B\mu_t$, where $\mu_t = E(z_t)$. An alternative is to set $E(z_t) = 0$ and use

$$X_t = \mu_t + Bz_t + \epsilon_t, \tag{2.11}$$

where μ_t is now a p-dimensional vector specifying the variational posterior mean for X_t directly.

We call parametrisation (2.8) the low-dimensional state mean (LD-SM) parametrisation, and parametrisation (2.11) the high-dimensional state mean (HD-SM) parametrisation. In both parametrisations, B forms a basis for X_t, which is reweighted over time according to the latent weights (factors) z_t. The LD-SM parametrisation provides information on how these basis functions are reweighted over time to form the approximate posterior mean, since $E(X_t) = B\mu_t$ and we infer both B and μ_t in the variational optimisation. This is explored in Section S6 for the spatio-temporal example in Section 3.

It is well known that factor models have identifiability issues (Shapiro, 1985). The choice of identifying constraints in factor models can matter, particularly for interpretation. However, here the choice of any identifying constraints is not crucial as we do not interpret either the factors or the loadings, but only use them for modeling the covariance matrix and, in the LD-SM parametrisation, also the variational mean.

2.5 Related work

We note that the sparsity imposed above is very important for reducing the number of variational parameters that need to be optimised. This allows the Gaussian variational approximation method to be extended to high dimensions. Our method relies on the combination of two ideas.

The first idea is explored in Tan and Nott (2018), who consider an approach which parametrises the precision matrix $\Omega = \Sigma^{-1} = CC^\top$ in terms of its Cholesky factor C, and impose a sparse structure on C which comes from the conditional independence structure in the model. We apply this idea to the Cholesky factor C_1 of the precision matrix of the dynamic factors as described in Section 2.4. Archer et al. (2016) also consider parametrising a Gaussian variational approximation using the precision matrix, but they optimise directly with respect to the elements Ω.

While the method of Tan and Nott (2018) is an attractive way to reduce the number of variational parameters in problems with an exploitable conditional independence structure, there are models where no such structure is available. An alternative parsimonious parametrisation is to use a factor model structure (Geweke and Zhou, 1996;

Bartholomew et al., 2011) when modelling the covariance matrix of the variational posterior (Ong et al., 2018). This is the second idea, which we use to parametrise the covariance matrix of the high-dimensional state vector X_t.

2.6 Implementation

The objective function of interest is (2.5) with $h(\theta)$ being the posterior (up to a normalising constant) of the general state space model in (2.1), i.e.

$$h(\theta) = p(X_0|\zeta) \prod_{t=1}^{T} s_t(X_t|X_{t-1}, \zeta) m_t(y_t|x_t, \zeta), \tag{2.12}$$

and $q_\lambda(\theta)$ as in (2.10). With regards to the variational approximation $q_\lambda(\theta)$, the user only needs to impose zeroes for C_2, which is the Cholesky factor of the precision matrix of the vector ζ of fixed parameters. This depends on the model: if the parameters ζ_i and ζ_j are not connected in (2.12), then the i, jth element in C_2 is set to zero.

To apply the reparametrisation trick, we cast the objective function to a similar expectation as in (2.6) by using the generative model for θ in (2.9). Appendix A gives the resulting gradients. These gradients above are functions of $\nabla_\theta \log h(\theta)$ with $h(\theta)$ in (2.12). As such, the implementation of the method only requires the user to derive the gradient of the model structure, i.e. $\nabla_\theta \log h(\theta)$, either analytically or via automatic differentiation. It is hard to efficiently apply automatic differentiation to obtain the gradients of the variational structure, i.e. the gradient elements in Lemmas A1 and A2 beside the model specific $\nabla_\theta \log h(\theta)$, because they involve a combination of sparse and low rank matrix manipulations.

Algorithm 1 outlines the stochastic gradient ascent algorithm maximising (2.6). The gradients are unbiasedly estimated by generating one or more samples from (ω, ϵ). Lemma A1 (A2) contains the gradients corresponding to (2.7) ((S2.1)), which we refer to as the standard gradient (the Roeder et al., 2017 gradient). The gradients in Lemma A2 follow Han et al. (2016), and are discussed in a very general way in Roeder et al. (2017), and are preferred if the variational approximation is accurate; Section S2 provides a detailed discussion. However, since we consider massive dimension reduction with only a small numbers of factors, the approximation may be crude and we therefore investigate both approaches in later examples.

2.7 Efficient computation

The gradient estimates for the lower bound are efficiently computed using a combination of sparse matrix operations (for evaluating terms such as $C^{-\top}\omega$ and the high-dimensional matrix multiplications in the expressions) and, as in Ong et al. (2018), the Woodbury identity for dense matrices such as $(W\Sigma W^\top + Z^2)^{-1}$ and $(W_1\Sigma_1 W^\top + D^2)^{-1}$. The Woodbury identity is

$$(\Lambda\Gamma\Lambda^\top + \Psi)^{-1} = \Psi^{-1} - \Psi^{-1}\Lambda(\Lambda^\top\Psi^{-1}\Lambda + \Gamma^{-1})^{-1}\Lambda^\top\Psi^{-1},$$

Algorithm 1: Stochastic gradient ascent for optimising the variational objective $\mathcal{L}(\lambda)$ in (2.6).

Input: Starting values $\lambda_0 \leftarrow (\mu_0, B_0, \delta_0, C_0)$, learning rates $\eta_\mu, \eta_B, \eta_\delta, \eta_C$, number of iterations M.

for $m = 1$ **to** M **do**

$\mu_m \leftarrow \mu_{m-1} + \eta_\mu \odot \widehat{\nabla_\mu \mathcal{L}}(\lambda_{m-1})$ ▷ $\nabla_\mu \mathcal{L}$ in (A2) or (A6)

$\lambda_{m-1} \leftarrow (\mu_m, B_{m-1}, \delta_{m-1}, C_{m-1})$ ▷ Update μ

$B_m \leftarrow B_{m-1} + \eta_B \odot \widehat{\nabla_{\mathrm{vec}(B)} \mathcal{L}}(\lambda_{m-1})$ ▷ $\nabla_{\mathrm{vec}(B)} \mathcal{L}$ in (A3) or (A7)

$\lambda_{m-1} \leftarrow (\mu_m, B_m, \delta_{m-1}, C_{m-1})$ ▷ Update B

$\delta_m \leftarrow \delta_{m-1} + \eta_\delta \odot \widehat{\nabla_\delta \mathcal{L}}(\lambda_{m-1})$ ▷ $\nabla_\delta \mathcal{L}$ in (A4) or (A9)

$\lambda_{m-1} \leftarrow (\mu_m, B_m, \delta_m, C_{m-1})$ ▷ Update δ

$C_m \leftarrow C_{m-1} + \eta_C \odot \widehat{\nabla_C \mathcal{L}}(\lambda_{m-1})$ ▷ $\nabla_C \mathcal{L}$ in (A5) or (A10)

$\lambda_m \leftarrow (\mu_m, B_m, \delta_m, C_m)$ ▷ Update C

$\lambda_{m-1} \leftarrow \lambda_m$ ▷ Update λ

end

Output: λ_m

for conformable matrices Λ, Γ and diagonal Ψ. It reduces the required computations into a much lower dimensional space since $q \ll p$ and Ψ is diagonal.

3 Application 1: Spatio-temporal model

3.1 Eurasian collared-dove data

The first example considers the spatio-temporal model of Wikle and Hooten (2006) for a dataset on the spread of the Eurasian collared-dove across North America. The dataset consists of the number of doves $y_{s_i t}$ observed at location s_i (latitude, longitude) $i = 1, \ldots, p$, in year $t = 1, \ldots, T = 18$, corresponding to an observation period of 1986–2003. The spatial locations correspond to $p = 111$ grid points with the dove counts aggregated within each area. See Wikle and Hooten (2006) for details. The count observed at location s_i at time t depends on the number of times $N_{s_i t}$ that the location was sampled.

3.2 Model

The model in Wikle and Hooten (2006) is

$$y_t | v_t \sim \mathrm{Poisson}(\mathrm{diag}(N_t) \exp(v_t)), \quad y_t, N_t, v_t \in \mathbb{R}^p$$
$$v_t | u_t, \sigma_\epsilon^2 \sim \mathcal{N}(u_t, \sigma_\epsilon^2 I_p), \quad u_t \in \mathbb{R}^p, \sigma_\epsilon^2 \in \mathbb{R}^+$$

$$u_t|u_{t-1},\psi,\sigma_\eta^2 \sim \mathcal{N}(H(\psi)u_{t-1},\sigma_\eta^2 I_p), \quad \psi \in \mathbb{R}^p, H(\psi) \in \mathbb{R}^{p\times p}, \sigma_\eta^2 \in \mathbb{R}^+,$$

with priors $\sigma_\epsilon^2, \sigma_\psi^2, \sigma_\alpha^2 \sim \text{IG}(2.8, 0.28), \sigma_\eta^2 \sim \text{IG}(2.9, 0.175)$ and

$$\begin{aligned} u_0 &\sim \mathcal{N}(0, 10I_p) \\ \psi|\alpha,\sigma_\psi^2 &\sim \mathcal{N}(\Phi\alpha,\sigma_\psi^2 I_p), \quad \Phi \in \mathbb{R}^{p\times l}, \alpha \in \mathbb{R}^l, \sigma_\psi^2 \in \mathbb{R}^+ \\ \alpha &\sim \mathcal{N}(0,\sigma_\alpha^2 R_\alpha), \quad \alpha_0 \in \mathbb{R}^l, R_\alpha \in \mathbb{R}^{l\times l}, \sigma_\alpha^2 \in \mathbb{R}^+. \end{aligned}$$

Poisson$(\cdot)$ is the Poisson distribution for a (conditionally) independent response vector parametrised in terms of its expectation and IG$(\cdot)$ is the inverse-gamma distribution with shape and scale as arguments. The dynamic process u_t evolves according to a diffusion equation and $H(\Psi)$ (approximately) solves the diffusion equation for one time step. The spatial dependence is modeled via the prior mean $\Phi\alpha$ of the diffusion coefficients ψ, where Φ consists of the l orthonormal eigenvectors with the largest eigenvalues of the spatial correlation matrix $R(c) = \exp(-cd) \in \mathbb{R}^{p\times p}$, where d is the Euclidean distance between pairwise grid locations in s_i. Finally, R_α is a diagonal matrix with the l largest eigenvalues of $R(c)$. We follow Wikle and Hooten (2006) and set $l = 1$ and $c = 4$.

Let $u = (u_0^\top, \dots u_T^\top)^\top$, $v = (v_1^\top, \dots v_T^\top)^\top$ and denote the parameter vector

$$\theta = (u, v, \psi, \alpha, \log\sigma_\epsilon^2, \log\sigma_\eta^2, \log\sigma_\psi^2, \log\sigma_\alpha^2),$$

whose posterior distribution is

$$\begin{aligned} p(\theta|y) \quad \propto \quad & \sigma_\epsilon^2\sigma_\eta^2\sigma_\psi^2\sigma_\alpha^2 p(\sigma_\epsilon^2)p(\sigma_\eta^2)p(\sigma_\psi^2)p(\sigma_\alpha^2)p(\alpha|\sigma_\alpha^2)p(\psi|\alpha,\sigma_\psi^2) \\ & p(u_0)\prod_{t=1}^T p(u_t|u_{t-1},\psi,\sigma_\eta^2)p(v_t|u_t,\sigma_\epsilon^2)p(y_t|v_t). \end{aligned} \tag{3.1}$$

Section S4.2 derives the gradient of the log-posterior.

3.3 Variational approximations of the posterior distribution

Section 2 considers two different parametrisation of the low rank approximation, in which either the state vector X_t has mean $BE(z_t) = B\mu_t$, $\mu_t \in \mathbb{R}^q$ (low-dimensional state mean, LD-SM) or X_t has a separate mean $\mu_t \in \mathbb{R}^p$ and $E(z_t) = 0$ (high-dimensional state mean, HD-SM). In this application there is a third choice of parametrisation which we now consider. The model in Section 3.2 connects the data with the high-dimensional state vector u_t via a high-dimensional auxiliary variable v_t. In the notation of Section 2, we can include v in ζ, in which case the parametrisation of the variational posterior is the one described there. We refer to this parametrisation as a low-rank state (LR-S). Although (3.1) shows that there is posterior dependence between u_t and v_t, the variational approximation in Section 2 omits the dependence between z and ζ. Moreover, v_t is also high-dimensional, but the LR-S parametrisation does not reduce its dimension. An alternative parametrisation that deals with both considerations includes v in the z-block, which we refer to as the low-rank state and auxiliary variable (LR-SA) parametrisation. This comes at the expense of omitting dependence between v_t

and σ_ϵ^2, and also becomes more computationally costly because, while the total number of variational parameters is smaller (see Table S1 in Section S8), the dimension of the z-block increases (B and C_1) and the main computational effort lies here and not in the ζ-block. Table 1 shows the CPU times relative to the LR-S parametrisation. The LR-SA parametrisation requires a small modification of the derivations in Section 2, which are outlined in detail in Section S5 as they can be useful for other models with a high-dimensional auxiliary variable.

It is straightforward to deduce conditional independence relationships in (3.1) to build the Cholesky factor C_2 of the precision matrix Ω_2 of ζ, with

$$\zeta = \begin{cases} (v, \psi, \alpha, \log \sigma_\epsilon^2, \log \sigma_\eta^2, \log \sigma_\psi^2, \log \sigma_\alpha^2) & \text{(LR-S)} \\ (\psi, \alpha, \log \sigma_\epsilon^2, \log \sigma_\eta^2, \log \sigma_\psi^2, \log \sigma_\alpha^2) & \text{(LR-SA).} \end{cases}$$

Section 2 outlines the construction of the Cholesky factor C_1 of the precision matrix Ω_1 of z, whereas the minor modification needed for LR-SA is in Section S5. We note that, regardless of the parametrisation, we obtain massive parsimony (between 6,428–11,597 variational parameters) compared to the saturated Gaussian variational approximation which in this application has 8,923,199 parameters; see Section S8 for further details.

We consider four different variational parametrisations, combining each of LR-SA or LR-S with the different parametrisation of the means of X_t, i.e. LD-SM or HD-SM. In all cases, we let $q = 4$ and perform 10,000 iterations of a stochastic optimisation algorithm with learning rates chosen adaptively according to the ADADELTA approach (Zeiler, 2012). We use the gradient estimators in Roeder et al. (2017), i.e. (A6), (A7), (A9) and (A10), although we found no noticeable difference compared to (A2)–(A5); it is likely that this is due to the small number of factors as described in Sections 2.3 and 2.6. Our choice is motivated by computational efficiency as some terms cancel out using the approach in Roeder et al. (2017). We initialise B and C as unit diagonal matrices, and μ and D are chosen to match the starting values of the Gibbs sampler in Wikle and Hooten (2006).

Figure 1 monitors the convergence via the estimated value of $\mathcal{L}(\lambda)$ using a single Monte Carlo sample. Table 1 presents estimates of $\mathcal{L}(\lambda)$ at the final iteration using 100 Monte Carlo samples. The results suggest that the best GVA parametrisation in terms of ELBO is the low-rank state algorithm (LR-SA) with, importantly, a high-dimensional state-mean (HD-SM) (otherwise the poorest GVA is achieved; see Table 1). However, Table 1 also shows that this parametrisation is about three times as CPU intensive. The fastest GVA parametrisations are both Low-Rank State (LR-S) algorithms, and modeling the state mean separately for these does not seem to improve $\mathcal{L}(\lambda)$ (Table 1) and is also slightly more computationally expensive (Table 1). Taking these considerations into account, the final choice of GVA parametrisation used for this model is the low-rank state with low-dimensional state mean (LR-S + LD-SM). Section 3.6 shows that this parametrisation gives accurate approximations for our analysis. For the rest of this example, we benchmark the GVA posterior from LR-S + LD-SM against the MCMC approach in Wikle and Hooten (2006) and also a Gibbs sampler based on ensemble Kalman filter methods that is well suited for problems with a high-dimensional state vector.

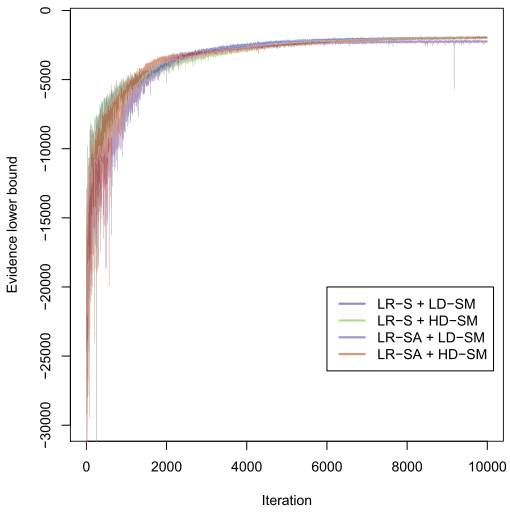

Figure 1: $\mathcal{L}(\lambda)$ for the variational approximations for the spatio-temporal example. The figure shows the estimated value of $\mathcal{L}(\lambda)$ vs iteration number for the four different parametrisations; see Section 3.3 or Table 1 for abbreviations.

3.4 MCMC settings

Before comparing the approximate methods (GVA and ensemble Kalman filter methods) to MCMC, it is necessary to determine a reasonable burn-in period and number of iterations for inference for the Gibbs sampler in Wikle and Hooten (2006). As it is infeasible to monitor convergence for every single parameter in such a large scale model as (3.1), we focus on ψ, u_{18} and v_{19}, which are among the variables considered in the analysis in Section 3.6.

Wikle and Hooten (2006) use 50,000 iterations, discarding the first 20,000 as burn-in. Four sampling chains are generated with these settings and inspected for convergence using the `coda` package (Plummer et al., 2006) in `R`. We compute the scale reduction factors (SRF) (Gelman and Rubin, 1992) for ψ, u_{18} and v_{19} as a function of the number of Gibbs iterations. The adequate number of iterations in MCMC depends on what functionals of the parameters are of interest; here we monitor convergence for these quantities since we report marginal posterior distributions for them later. The scale reduction factor of a parameter measures if there is a significant difference between the variance within the four chains and the variance between the four chains of that parameter. We use the rule of thumb that concludes convergence occurs when SRF $<$ 1.1, which gives a burn-in period of approximately 40,000 for these functionals here. After discarding these samples and applying a thinning of 10 we are left with 1,000 posterior samples for inference. However, as the draws are auto-correlated, this does not correspond to 1,000 independent draws used in the analysis in Section 3.6 (note that we obtain independent samples from our variational posterior). To decide how many Gibbs samples are equivalent to 1,000 independent samples for ψ, u_{18} and v_{19}, we

Parametrisation			
Spatio-temporal	R-CPU	$\mathcal{L}(\lambda_{\text{opt}})$	Confidence interval
LR-S + LD-SM	1	−1,996	[−2,004; −1,988]
LR-S + HD-SM	1.005	−2,024	[−2,032; −2,016]
LR-SA + LD-SM	3.189	−2,158	[−2,167; −2,148]
LR-SA + HD-SM	3.017	−1,909	[−1,918; −1,900]
Wishart process			
LR-S + LD-SM	1	−1,121	[−1,126; −1,115]
LR-S + HD-SM	1.0004	−1,040	[−1,046; −1,035]

Table 1: $\mathcal{L}(\lambda)$ and CPU time for the GVA parametrisations in the spatio-temporal and Wishart process example. The table shows the estimated value of $\mathcal{L}(\lambda)$ for the different GVA parametrisations by combining low-rank state / low-rank state and auxiliary (LR-S / LR-SA) with either of low-dimensional state mean / high-dimensional state mean (LD-SM / HD-SM). The estimate and its 95% confidence interval are computed at the final iteration using 100 Monte Carlo samples. The table also shows the relative CPU (R-CPU) times, where the reference is LD-SM.

compute the effective sample size (ESS) which takes into account the auto-correlation of the samples. We find that the smallest is ESS = 5 and hence we require 200,000 iterations after a thinning of 10, which makes a total of 2,000,000 Gibbs iterations, excluding the burn-in of 40,000. Thinning is advisable here due to memory issues — it is impractical to store 2,000,000 iterations for each parameter (which may be used, for example, to estimate kernel densities) in high-dimensional models.

3.5 The Gibbs ensemble Kalman smoother

Katzfuss et al. (2020) develop scalable Markov chain Monte Carlo inference for high-dimensional state space models based on ensemble Kalman filter methods for filtering (Evensen, 1994) and smoothing (Evensen and Van Leeuwen, 2000). The filtering density in the standard Kalman filter is costly to compute when the state is high-dimensional. The idea of the ensemble Kalman filter is to represent the filtering density via an initial ensemble from the prior of the state vector, and then i) forecast the ensemble, and ii) update the forecast; see Katzfuss et al. (2020) for details. Katzfuss et al. (2020) propose the Gibbs ensemble Kalman smoother (GEnKS), which is a Gibbs algorithm that uses the ensemble Kalman smoother (which requires the ensemble Kalman filter) when sampling from the full conditional distribution of the states. The updates of the rest of the parameters are then carried out conditional on this sample following the usual Gibbs sampling procedure. An important assumption in ensemble Kalman filter methods is that the distribution of the state transition equation is Gaussian. While this is the case for the model in this section (see Section 3.2), it does not hold for the model in Section 4.1.

We implement the GEnKS algorithm with 100 ensembles using a burn-in period of 5,000 iterations and with 33,000 post burn-in samples, which gives an average effective

sample size for ψ similar to that of the MCMC algorithm with the settings described in Section 3.4.

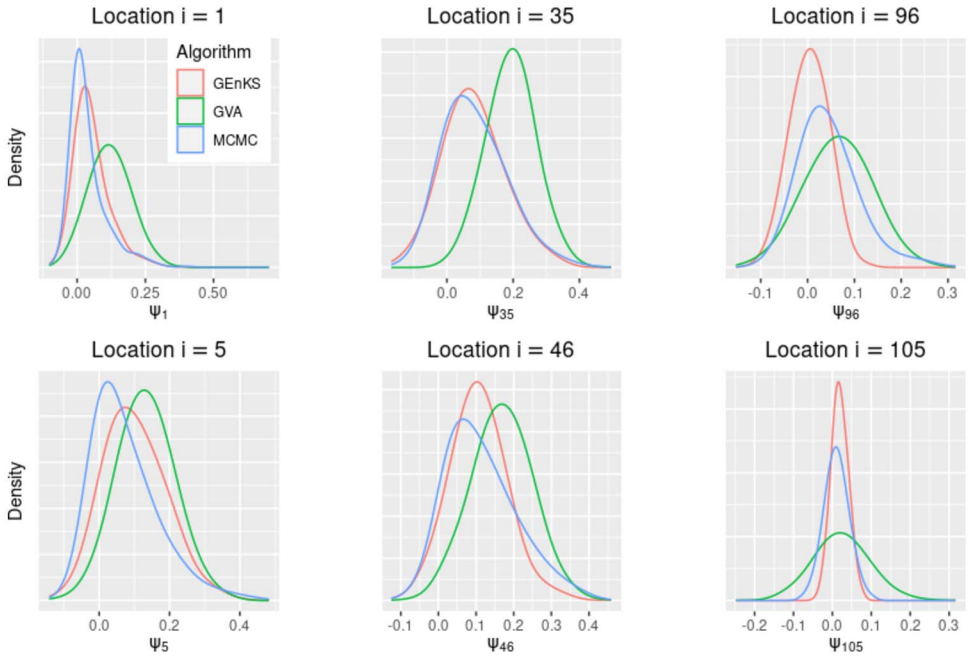

Figure 2: Distribution of the diffusion coefficients. The figure shows the posterior distribution of ψ_i obtained by MCMC, GVA and GEnKS. The locations are divided into three categories (total doves over time within brackets): zero count locations (Idaho, $i = 1\,[0]$, Arizona $i = 5\,[0]$, left panels), low count locations (Texas, $i = 35\,[16], 46\,[21]$, middle panels) and high count locations (Florida, $i = 96\,[1566], 105\,[1453]$, right panels).

3.6 Analysis and results

We first consider inference on the diffusion coefficient ψ_i for location i. Figure 2 plots the "true" posterior (represented by MCMC) together with the variational and GEnKS approximations for six locations described in the caption of the figure. The figure shows that the posterior distribution is highly skewed for locations with zero dove counts and approaches normality as the dove counts increase. Consequently, the accuracy of the variational posterior (which is Gaussian) improves with increasing dove counts. We note that the GEnKS provides a better approximation than the GVA approach; however, some notable discrepancies to the true posterior densities also exist for GEnKS.

Regarding the inferential performance of our method, it is well-known that inadequate approximations to the posterior distribution of the states can lead to inconsistent estimation of the fixed parameters (Wang and Titterington, 2004; Frazier et al., 2021a), although the impact on predictive performance is limited in many cases (Frazier et al.,

2021a). Frazier et al. (2021a) show that integrating out the states within a variational approximation by unbiased estimation of the likelihood (Tran et al., 2017), or exact sampling from the conditional distribution of the states if possible (Loaiza-Maya et al., 2021), is preferable to using a parametrised approximation to the states as in our approach. However, they also acknowledge that with a high-dimensional state, these methods are infeasible in practice. The state-of-the-art in this setting uses ensemble Kalman filter methods and our approach is competitive in terms of accuracy, but has a much smaller computational time as discussed above.

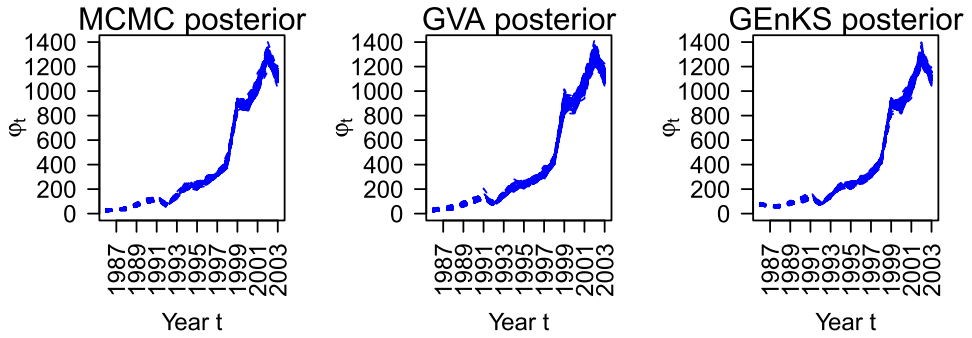

Figure 3: Samples from the posterior sum of dove intensity over the spatial grid for each year. The figure shows 100 samples from the posterior distribution of $\varphi_t = \sum_i \exp(v_{it})$ obtained by MCMC (left panel), GVA (middle panel) and GEnKS (right panel).

Figure 3 shows 100 GVA, GEnKS and MCMC posterior samples of the dove intensity for each year summed over the spatial locations, i.e. $\varphi_t = \sum_i \exp(v_{it})$. The three posteriors are similar and show an exponential increase of doves until the year 2002 followed by a steep decline for 2003. Section S6 shows some spatial properties of the model, confirming that GVA gives accurate location estimates of the spatial process.

Figure 4 illustrates the posterior distribution of the log-intensity for selected locations for 2003, as well as the out-of-sample posterior predictive distribution for 2004. The posterior distributions for the GVA, GEnK and MCMC are similar for the in-sample prediction year 2003. For the out-of-sample prediction year 2004, GEnKS does not give an accurate result because of a large error in estimating σ_ϵ^2; the point estimate with GEnKS is 0.059 compared to 4.45819 with MCMC. The corresponding point estimate for GVA is 4.781454, and thus GVA and MCMC give similar results for this out-of-sample prediction. It is evident that using this large scale model for forecasting future values is associated with a large uncertainty. We note that when the ensemble size and the number of MCMC samples tend to infinity, GEnKS will under mild regularity conditions produce samples from the exact posterior if the state evolution is linear and a consistent estimator of the forecast covariance matrix is used (Katzfuss et al., 2020). However, a large ensemble size is computationally infeasible in this example. With ensemble size 100, GEnKS was about 12.3 times slower than GVA. GVA took about 6.5 hours to run. We also note that GVA is 7.3 times faster than MCMC.

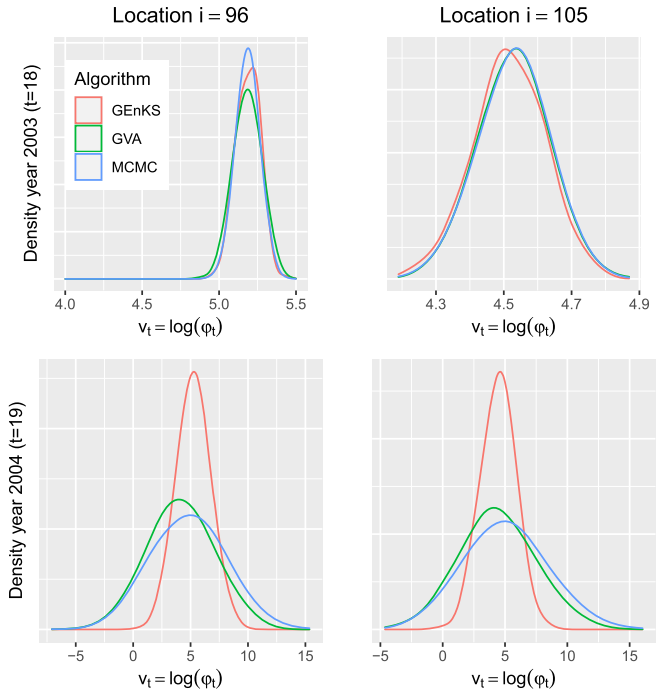

Figure 4: Forecasting the log-intensity of the spatial process. The figure shows an in-sample forecast density of the log-intensity v_{it} for 2003 ($t = 18$, upper panels) and out-of-sample forecast density for 2004 ($t = 19$, lower panels) for Central Florida ($i = 96$, left panels) and South East Florida ($i = 105$, right panels).

4 Application 2: Stochastic volatility modeling

4.1 Model

The second example considers the Wishart based multivariate stochastic volatility model proposed in Philipov and Glickman (2006) who used it to model the time-varying dependence of a portfolio of k assets over T time periods. Section 2.2 briefly discusses this model.

The model of the mean-centred returns in Philipov and Glickman (2006) is

$$y_t \sim \mathcal{N}(0, \Sigma_t); \quad \Sigma_t \in \mathbb{R}_+^{k\times k};$$
$$\Sigma_t^{-1} \sim \text{Wishart}(\nu, S_{t-1}); \quad S_t = \frac{1}{\nu} H(\Sigma_t^{-1})^d H^\top; \; S_t \in \mathbb{R}_+^{k\times k}; \; \nu > k; \; 0 < d < 1;$$

H is a lower triangular Cholesky factor of $A = HH^\top \in \mathbb{R}_+^{k\times k}$ and Σ_0 is assumed known. Philipov and Glickman (2006) use the following priors: the prior for A is inverse Wishart, i.e. $A^{-1} \sim \text{Wishart}(\gamma_0, Q_0)$, with $\gamma_0 = k+1$ and $Q_0 = I$; d has a uniform prior

on $[0,1]$, i.e. $d \sim U[0,1]$; ν has a shifted gamma prior, i.e. $\nu - k \sim \text{Gamma}(\alpha_0, \beta_0)$. The joint posterior density for $(\Sigma, A, \nu - k, d)$ is

$$p(\Sigma, A, \nu - k, d|y) \propto p(A, d, \nu - k) \prod_{t=1}^{T} p(\Sigma_t|\nu, S_{t-1}) p(y_t|\Sigma_t); \tag{4.1}$$

$p(A, d, \nu - k)$ is the joint prior density for $(A, d, \nu - k)$; $p(\Sigma_t|\nu, S_{t-1}, d)$ is the conditional inverse Wishart prior for Σ_t given ν, S_{t-1} and d, and $p(y_t|\Sigma_t)$ is the normal density for y_t given Σ_t.

We write C_t for the Cholesky factor of Σ_t and reparametrise the posterior in terms of the unconstrained parameter vector

$$\theta = (\text{vech}(H')^\top, d', \nu', \text{vech}(C_1')^\top, \ldots, \text{vech}(C_T')^\top)^\top;$$

where

$$\begin{aligned} &C_t' \in \mathbb{R}^{k \times k}; && C_{t,ij}' = C_{t,ij};\ i \neq j, \text{ and } C_{t,ii}' = \log C_{t,ii}; \\ &H' \in \mathbb{R}^{k \times k}; && H_{ij}' = H_{ij};\ i \neq j, \text{ and } H_{ii} = \log H_{ii}; \end{aligned}$$

with $d' = \log d/(1-d)$ and $\nu' = \log(\nu - k)$. Section S4.3 shows that

$$\begin{aligned} p(\theta|y) \propto\ & |L_k(I_{k^2} + K_{k,k})(H \otimes I_k)L_k^\top| \times \left\{ \prod_{t=1}^{T} |L_k(I_{k^2} + K_{k,k})(C_t \otimes I_k)L_k^\top| \right\} \times (\nu - k) \\ & \times d(1-d) \times \left\{ \prod_i H_{ii} \right\} \left\{ \prod_{t=1}^{T} \prod_{i=1}^{k} C_{t,ii} \right\} \times p(A, d, \nu - k) \\ & \times \left\{ \prod_{t=1}^{T} p(\Sigma_t|\nu, S_{t-1}, d) p(y_t|\Sigma_t) \right\}; \end{aligned} \tag{4.2}$$

Section S1 defines the elimination matrix L_k and the commutation matrix $K_{k,k}$; Section S4.4 shows how to evaluate the gradient of the log posterior.

4.2 Evaluating the predictive performance of the variational approximation

Philipov and Glickman (2006) develop an MCMC algorithm to estimate their Wishart based multivariate stochastic volatility model. Rinnergschwentner et al. (2012) point out that the Gibbs sampler developed by Philipov and Glickman (2006) contains a mistake which affects all the full conditionals. We find that implementing the corrected version of their algorithm results in a very inefficient sampler even for the $k = 5$ portfolios used by Philipov and Glickman (2006) in their empirical example. This means that the corrected Philipov and Glickman (2006) algorithm cannot be used as a 'gold standard' to compare against the variational approximation results. Although it may be possible to estimate the posterior of the Philipov and Glickman (2006) model using particle

methods, we do not pursue this here as particle methods do not scale well to high-dimensional states (Katzfuss et al., 2020). Section S9 illustrates the inefficiency of the corrected Philipov and Glickman (2006) sampler and explains its problems.

We now show empirically (by simulation) that the variational posterior provides useful predictive inference. Since MCMC is unavailable, the GVA is benchmarked against an oracle predictive approach, which assumes that the fixed model parameters are known. We use a bootstrap particle filter (Gordon et al., 1993) to obtain the posterior density of the state-vector at $t = T$; it is then possible to obtain the one-step ahead oracle predictive density $p(y_{T+1}|y_{1:T}, \zeta^{\text{true}})$, where ζ^{true} denotes the true fixed model parameters. The variational predictive density is then benchmarked against the oracle predictive density; we note that the variational predictive density integrates over the variational posterior of all the parameters, including the fixed model parameters.

Section S7.1 shows how to simulate from the oracle predictive density. Section S7.2 shows how to simulate from the variational predictive density. The one-step ahead prediction is repeated for $H = 4$ horizons. At horizons $h = 1, \ldots, H$, both filtering densities are based on $y_{1:T+h-1}$ and the optimisation for finding the variational posterior for $h > 1$ is fast since the variational parameters are initialised (except the ones added at $T + h$) at their variational mode from the previous optimisation.

4.3 Variational approximations of the posterior distribution

Since this example does not include a high-dimensional auxiliary variable, we use the low-rank state (LR-S) parametrisation combined with both a low-dimensional state mean (LD-SM) and a high-dimensional state mean (HD-SM). As in the previous example, it is straightforward to deduce conditional independence relationships in (4.2) to build the Cholesky factor C_2 of the precision matrix Ω_2 of ζ in Section 2; this section also outlines how to construct the Cholesky factor C_1 of the precision matrix Ω_1 of z. Massive parsimony is achieved in this application. In particular, for $k = 12$ assets, the saturated Gaussian variational approximation has 31,059,020 parameters, while our parametrisation has 10,813. For $k = 5$, the saturated case has 1,152,920 parameters and our parametrisations has 4,009–5,109. See Section S8 for more details.

For all the variational approximations, we take $q = 4$ and perform 10,000 iterations of a stochastic optimisation algorithm with learning rates chosen adaptively according to the ADADELTA approach (Zeiler, 2012). We initialise B and C as unit diagonal matrices and choose μ and D randomly. Figure 5 monitors the estimated ELBO for both parametrisations, using both the gradient estimators in Roeder et al. (2017) and the alternative standard ones which do not cancel terms that have zero expectation. For $k = 5$, the figure shows that the different gradient estimators perform equally well. Moreover, slightly more variable estimates are observed in the beginning for the low-dimensional state mean parametrisation compared to that of the high-dimensional mean. Table 1 presents estimates of $\mathcal{L}(\lambda)$ at the final iteration using 100 Monte Carlo samples and also presents the relative CPU times of the algorithms. In this example, the separate state mean present in the high-dimensional state mean improves the ELBO considerably.

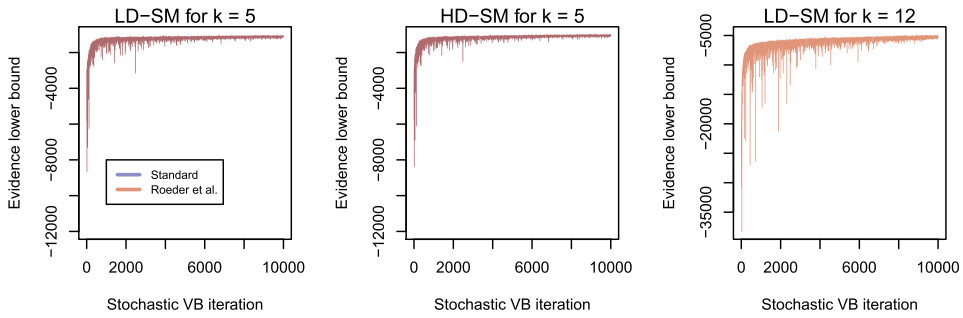

Figure 5: $\mathcal{L}(\lambda)$ for the variational approximations in the Wishart process example. The figure shows the estimated value of $\mathcal{L}(\lambda)$ vs iteration number using a low-dimensional state mean / high-dimensional state mean (LD-SM / HD-SM) with the gradient estimator in Roeder et al. (2017) or the standard estimator. The left and middle panels are for $k = 5$; the right panel is for the real data with $k = 12$.

4.4 Results for simulated data

We now assess the variational approximation by comparing its out-of-sample predictive properties against the oracle predictive density described in Section 4.2. The comparison is based on data generated by the multivariate stochastic volatility model with $d = 0.2$, $\nu = 20$ and A generated from $\sim$ Inv-Wishart$(5, \text{diag}(5))$. While the reported results are for a particular simulated dataset due to space restrictions, we have verified that the same performance is obtained when the random number seed is changed and d and ν are varied. Figure 6 shows the kernel density estimates for the marginals predictive densities of all five assets and bivariate kernel density estimates for all pairs of assets for the predictive $p(y_{T+1}|y_{1:T})$ (variational and oracle) for $T = 100$. The figure also shows the test observation (withheld when estimating the variational predictive and the oracle predictive). Figure 7 shows boxplots of draws from all marginals of the predictive densities $p(y_{T+h}|y_{1:T+h-1})$ (variational and oracle) for the horizons $h = 1, 2, 3, 4$. This figure also shows the withheld test observation which is within the prediction intervals of both methods. Figure 8 shows, for each of the $H = 4$ horizons, future predictions (variational and oracle) of an equally weighted portfolio $w_{T+h} = \sum_{k=1}^{5}(1/5)y_{(T+h)k}$ conditional on the posteriors using the data $y_{1:(T+h-1)}$. Section S7.3 gives plots that further confirm the accuracy of the variational predictive densities.

4.5 Real data results

The data consists of $T = 100$ monthly observations on all $k = 12$ value-weighted portfolios from the 201709 CRSP database, for the period 2009-06 to 2017-09. The portfolios are: consumer non-durables, consumer durables, manufacturing, energy, chemicals, business equipment, telecom, utilities, retail/wholesale, health care, finance, other. With $k = 12$ the dimension of the state vector is $p = 78$. We follow Philipov and Glickman (2006) and prefilter each series using an AR(1) process. Philipov and Glickman (2006)

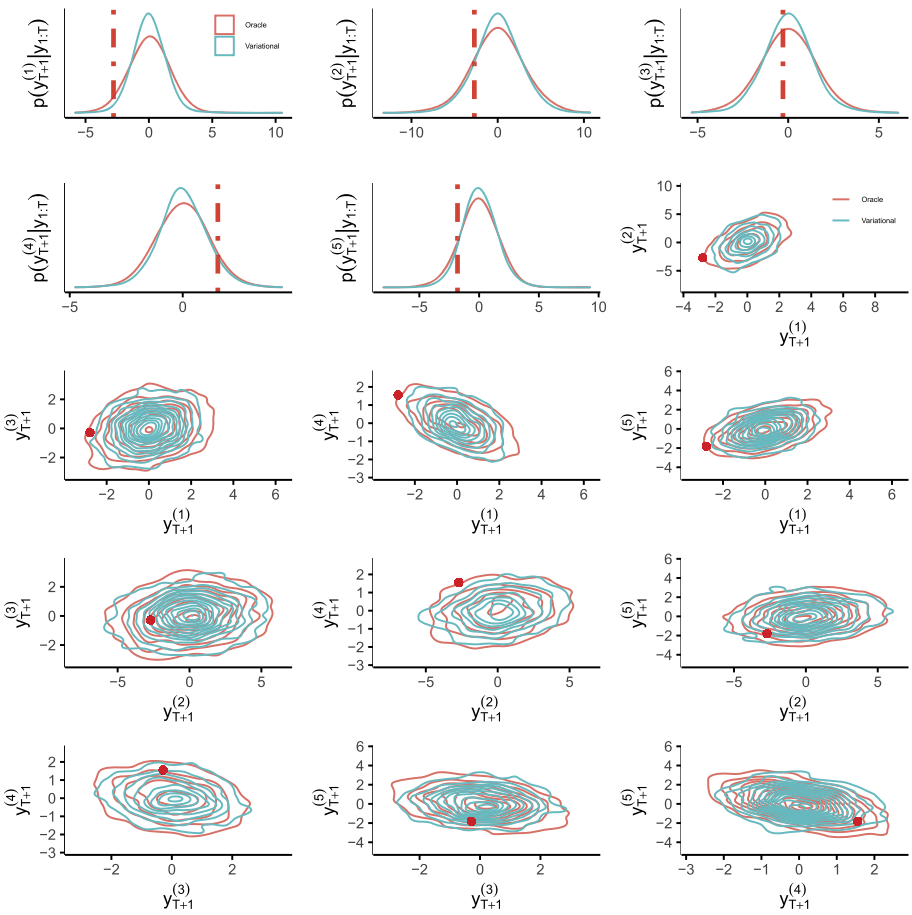

Figure 6: Multivariate stochastic volatility model with simulated data and $T = 100$. The top row and the two panels from the left of the second row show the marginal one-step-ahead kernel density estimates of the predictive density for each of the $k = 5$ variables for both the variational approximation and the oracle; the test observation is the red line. The right panel of the second row and the rest of the panels show the contour plots of the kernel density estimates of the one-step-ahead bivariate predictive densities for the variational approximation and the oracle; the red dot is the test observation.

only consider $k = 5$ assets and report an acceptance probability close to zero when $k = 12$ for their sampler.

The right panel in Figure 5 shows the estimated ELBO on a variational optimisation using the real dataset. While the estimated ELBO plot is more variable than for the $k = 5$ case, it settles down eventually. Figure 9 shows the in-sample prediction of $\tilde{y}_{100}$ given $y_{1:100}$, together with the observed data point, for some of the assets. The figure also shows an in-sample prediction of a portfolio consisting of equally weighted assets.

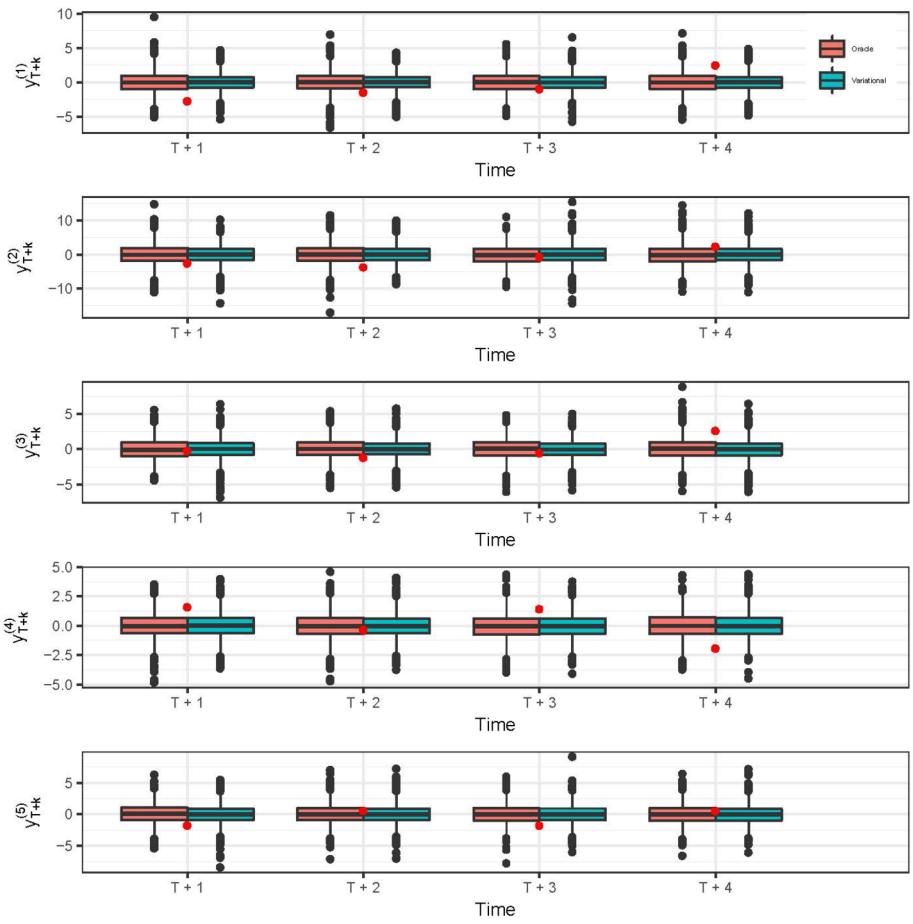

Figure 7: Boxplots of samples from the variational one-step ahead marginal predictive densities compared against the oracle predictive densities with $T = 100, 101, 102, 103$ using simulated data. The figure also shows the test observation (red) dot for each T and variable.

The variational posterior for the real data example uses the low-dimensional state mean parametrisation.

5 Discussion

We propose an all-purpose GVA method for high-dimensional state space models. Dimension reduction in the variational approximation is achieved through a dynamic factor structure for the variational covariance matrix. The factor structure reduces the dimension in the description of the states, whereas the Markovian dynamic structure for the

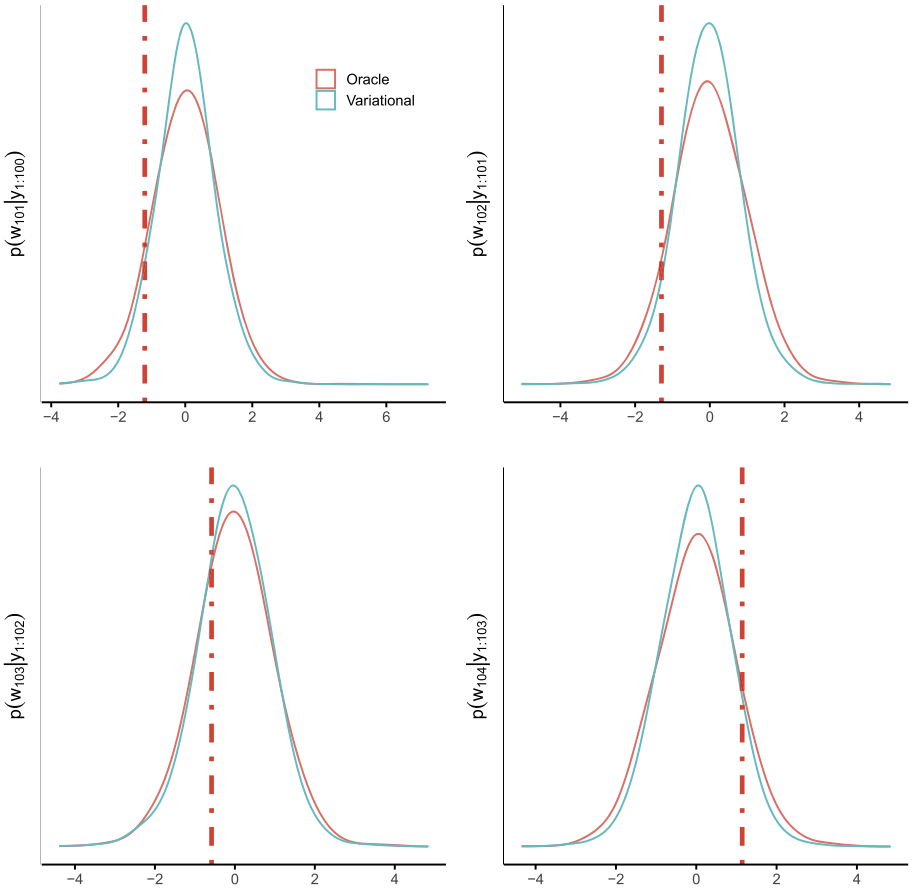

Figure 8: Kernel density estimates of the one-step ahead predictive densities of an equally weighted portfolio of assets for the simulated data. The results are for $T = 100, 101, 102, 103$. The figure also shows the test observation (red line) for each T.

factors achieves parsimony in describing the temporal dependence. We show that the method works well in two challenging models. The first is an extended example for a spatio-temporal data set describing the spread of the Eurasian collared-dove throughout North America. We benchmark our method against the Gibbs ensemble Kalman smoother with favorable outcomes. In particular, our method is 12 times faster in the demonstration example. Moreover, our method does not rely on the state distribution being Gaussian, and so applies more widely than ensemble Kalman filter methods. Thus, it applies to our second example which is a multivariate stochastic volatility model in which the state vector is high dimensional and follows a Wishart distribution.

The most obvious limitation of our current work is the restriction to a Gaussian approximation, which cannot capture skewness or heavy tails in the posterior distribution.

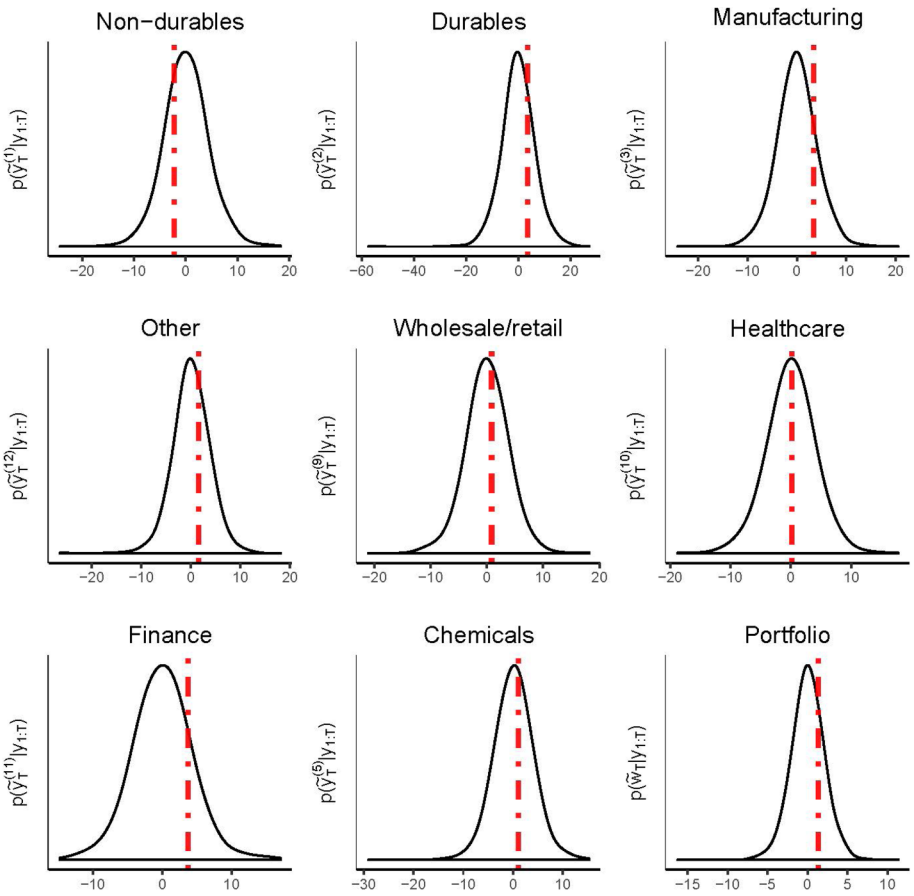

Figure 9: Kernel density estimates of the in-sample predictive density of the multivariate stochastic volatility model for some of the assets and an equally weighted portfolio of assets using real data. The figure also shows the in-sample observation (red line).

However, Gaussian variational approximations can be used as building blocks for more complex approximations based on normal mixtures, copulas or conditionally Gaussian families, for example Han et al. (2016); Miller et al. (2016); Smith et al. (2020); Tan et al. (2020) and these more complex variational families can overcome some of the limitations of the simple Gaussian approximation. We intend to consider this in future work.

6 Technical definitions

We consider any vector $x \in \mathbb{R}^n$ to be arranged as a column vector with n elements, i.e. $x = (x_1, \ldots, x_n)^\top$. Likewise, if g is a function whose range is vector valued, i.e.

$g(x) \in \mathbb{R}^m$, then $g(x) = (g_1(x), \ldots, g_m(x))^\top$. For a matrix A, $\text{vec}(A)$ is the vector obtained by stacking the columns of A from left to right.

Definition 1. *(i) Suppose that $g : \mathbb{R}^n \to \mathbb{R}$ is a scalar valued function of a vector valued argument x. Then $\nabla_x g$ is a column vector with ith element $\partial g/\partial x_i$.*

(ii) Suppose that $g : \mathbb{R}^n \to \mathbb{R}^m$ is a vector valued function of a vector valued argument x. Then dg/dx is a $m \times n$ matrix with (i, j)th element $\partial g_i/\partial x_j$.

(iii) Suppose that $g : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a scalar valued function of a $m\times n$ matrix $A = (a_{ij})$. Then $\nabla_A g$ is an $m \times n$ matrix with (i, j)th element $\partial g/\partial a_{ij}$.

(iv) Suppose that $G : \mathbb{R}^{m\times n} \to \mathbb{R}^{q\times r}$ is a matrix valued function of a matrix valued argument A. Then,

$$\frac{dG}{dA} = \frac{d\,\text{vec}(G)}{d\,\text{vec}(A)}$$

is an $mq \times nr$ matrix with (i, j)th element $\partial\,\text{vec}(G)_i/\partial\,\text{vec}(A)_j$.

Remark 1. *If g is a scalar function of a vector valued argument x, then Part (ii) (with $m = 1$) implies that dg/dx is a row vector. Hence, $\nabla_X g = (dg/dx)^\top$.*

We write I_n for the $n \times n$ identity matrix, $0_{m\times n}$ for the $m \times n$ matrix of zeros, $\otimes$ for the Kronecker product and $\odot$ for the Hadamard (elementwise) product which can be applied to two matrices of the same dimensions. We also write $K_{r,s}$ for the commutation matrix, of dimensions $rs \times rs$, which for an $r \times s$ matrix Z satisfies

$$K_{r,s}\,\text{vec}(Z) = \text{vec}(Z^\top).$$

Supplementary Material

Web-based supplementary materials (DOI: 10.1214/22-BA1332SUPP; .pdf). The supplementary material contains expression of the gradients with proofs, and additional material regarding results and modification of the methods.

References

Archer, E., Park, I. M., Buesing, L., Cunningham, J., and Paninski, L. (2016). "Black box variational inference for state space models." arXiv:1511.07367. 995

Attias, H. (1999). "Inferring parameters and structure of latent variable models by variational Bayes." In Laskey, K. and Prade, H. (eds.), *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 21–30. Morgan Kaufmann. 992

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach, 3rd edition*. John Wiley & Sons. MR2849614. doi: https://doi.org/10.1002/9781119970583. 995

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. MR3671776. doi: https://doi.org/10.1080/01621459.2017.1285773. 989

Bottou, L. (2010). "Large-scale machine learning with stochastic gradient descent." In Lechevallier, Y. and Saporta, G. (eds.), *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, 177–187. Springer. MR3362066. 993

Cressie, N. and Wikle, C. (2011). *Statistics for Spatio-Temporal Data*. Wiley. MR2848400. 990

Evensen, G. (1994). "Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics." *Journal of Geophysical Research: Oceans*, 99(C5): 10143–10162. 990, 1001

Evensen, G. and Van Leeuwen, P. J. (2000). "An ensemble Kalman smoother for nonlinear dynamics." *Monthly Weather Review*, 128(6): 1852–1867. 990, 1001

Frazier, D. T., Loaiza-Maya, R., and Martin, G. M. (2021a). "A note on the accuracy of variational Bayes in state space models: Inference and prediction." arXiv:2106.12262. 990, 1002, 1003

Frazier, D. T., Loaiza-Maya, R., Martin, G. M., and Koo, B. (2021b). "Loss-based variational Bayes prediction." arXiv:2104.14054. 990

Gelman, A. and Rubin, D. B. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science*, 7: 457–472. 1000

Geweke, J. and Zhou, G. (1996). "Measuring the pricing error of the arbitrage pricing theory." *Review of Financial Studies*, 9(2): 557–587. 995

Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). "Novel approach to nonlinear/non-Gaussian Bayesian state estimation." In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, 107–113. IET. 1006

Han, S., Liao, X., Dunson, D. B., and Carin, L. C. (2016). "Variational Gaussian copula inference." In Gretton, A. and Robert, C. C. (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, 829–838. Cadiz, Spain: JMLR Workshop and Conference Proceedings. 993, 996, 1011

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). "Stochastic variational inference." *Journal of Machine Learning Research*, 14: 1303–1347. MR3081926. 993

Ji, C., Shen, H., and West, M. (2010). "Bounded approximations for marginal likelihoods." Technical Report 10-05, Institute of Decision Sciences, Duke University. URL http://ftp.stat.duke.edu/WorkingPapers/10-05.html 993

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). "An introduction to variational methods for graphical models." *Machine Learning*, 37: 183–233. 992

Katzfuss, M., Stroud, J. R., and Wikle, C. K. (2020). "Ensemble Kalman methods for high-dimensional hierarchical dynamic space-time models." *Journal of the American*

Statistical Association, 115(530): 866–885. MR4107685. doi: https://doi.org/10.1080/01621459.2019.1592753. 991, 992, 1001, 1003, 1006

Kingma, D. P. and Welling, M. (2014). "Auto-encoding variational Bayes." In *Proceedings of the 2nd International Conference on Learning Representations (ICLR) 2014*. 993

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). "Automatic differentiation variational inference." *Journal of Machine Learning Research*, 18(14): 1–45. MR3634881. 993

Loaiza-Maya, R., Smith, M. S., Nott, D. J., and Danaher, P. J. (2021). "Fast and accurate variational inference for models with many latent variables." *Journal of Econometrics*. MR4466728. doi: https://doi.org/10.1016/j.jeconom.2021.05.002. 1003

Miller, A. C., Foti, N., and Adams, R. P. (2016). "Variational boosting: Iteratively refining posterior approximations." arXiv:1611.06585. 1011

Nott, D. J., Tan, S. L., Villani, M., and Kohn, R. (2012). "Regression density estimation with variational methods and stochastic approximation." *Journal of Computational and Graphical Statistics*, 21: 797–820. MR2970920. doi: https://doi.org/10.1080/10618600.2012.679897. 993

Ong, V. M.-H., Nott, D. J., and Smith, M. S. (2018). "Gaussian variational approximation with a factor covariance structure." *Journal of Computational and Graphical Statistics*, 27(3): 465–478. MR3863750. doi: https://doi.org/10.1080/10618600.2017.1390472. 996

Ormerod, J. T. and Wand, M. P. (2010). "Explaining variational approximations." *The American Statistician*, 64: 140–153. MR2757005. doi: https://doi.org/10.1198/tast.2010.09058. 989, 992

Paisley, J. W., Blei, D. M., and Jordan, M. I. (2012). "Variational Bayesian inference with stochastic search." In Langford, J. and Pineau, J. (eds.), *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*. 993

Philipov, A. and Glickman, M. E. (2006). "Multivariate stochastic volatility via Wishart processes." *Journal of Business & Economic Statistics*, 24(3): 313–328. MR2252482. doi: https://doi.org/10.1198/073500105000000306. 990, 992, 1004, 1005, 1006, 1007

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). "CODA: Convergence diagnosis and output analysis for MCMC." *R News*, 6(1): 7–11. 1000

Quiroz, M., Nott, D. J., and Kohn, R. (2022). "Supplementary Material for "Gaussian variational approximations for high-dimensional state space models"." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1332SUPP. 991

Ranganath, R., Gerrish, S., and Blei, D. M. (2014). "Black box variational inference." In Kaski, S. and Corander, J. (eds.), *Proceedings of the 17th International Conference*

on Artificial Intelligence and Statistics, volume 33, 814–822. JMLR Workshop and Conference Proceedings. 993

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). "Stochastic backpropagation and approximate inference in deep generative models." In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 29th International Conference on Machine Learning, ICML 2014*. 993

Rinnergschwentner, W., Tappeiner, G., and Walde, J. (2012). "Multivariate stochastic volatility via Wishart processes: A comment." *Journal of Business & Economic Statistics*, 30(1): 164–164. MR2899193. doi: https://doi.org/10.1080/07350015.2012.634358. 1005

Robbins, H. and Monro, S. (1951). "A stochastic approximation method." *The Annals of Mathematical Statistics*, 22: 400–407. MR0042668. doi: https://doi.org/10.1214/aoms/1177729586. 992

Roeder, G., Wu, Y., and Duvenaud, D. (2017). "Sticking the landing: Simple, lower-variance gradient estimators for variational inference." arXiv:1703.09194. 993, 996, 999, 1006, 1007

Salimans, T. and Knowles, D. A. (2013). "Fixed-form variational posterior approximation through stochastic linear regression." *Bayesian Analysis*, 8: 837–882. MR3150471. doi: https://doi.org/10.1214/13-BA858. 993

Shapiro, A. (1985). "Identifiability of factor analysis: Some results and open problems." *Linear Algebra and its Applications*, 70: 1–7. MR0808527. doi: https://doi.org/10.1016/0024-3795(85)90038-2. 995

Smith, M. S., Loaiza-Maya, R., and Nott, D. J. (2020). "High-dimensional copula variational approximation through transformation." *Journal of Computational and Graphical Statistics*, 29(4): 729–743. MR4191239. doi: https://doi.org/10.1080/10618600.2020.1740097. 1011

Tan, L. S. and Nott, D. J. (2018). "Gaussian variational approximation with sparse precision matrices." *Statistics and Computing*, 28(2): 259–275. MR3747562. doi: https://doi.org/10.1007/s11222-017-9729-7. 993, 995

Tan, L. S.-L., Bhaskaran, A., and Nott, D. J. (2020). "Conditionally structured variational Gaussian approximation with importance weights." *Statistics and Computing*, 30(5): 1255–1272. MR4137250. doi: https://doi.org/10.1007/s11222-020-09944-8. 1011

Titsias, M. and Lázaro-Gredilla, M. (2015). "Local expectation gradients for black box variational inference." In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 2638–2646. Curran Associates, Inc. 993

Tran, M.-N., Nott, D. J., and Kohn, R. (2017). "Variational Bayes with intractable likelihood." *Journal of Computational and Graphical Statistics*, 26(4): 873–882. MR3765351. doi: https://doi.org/10.1080/10618600.2017.1330205. 1003

Wang, B. and Titterington, D. (2004). "Lack of consistency of mean field and variational Bayes approximations for state space models." *Neural Processing Letters*, 20(3): 151–170. 1002

Wang, Y. and Blei, D. M. (2019). "Frequentist consistency of variational Bayes." *Journal of the American Statistical Association*, 114(527): 1147–1161. MR4011769. doi: https://doi.org/10.1080/01621459.2018.1473776. 990

Wikle, C. K. and Hooten, M. B. (2006). "Hierarchical Bayesian spatio-temporal models for population spread." In Clark, J. S. and Gelfand, A. (eds.), *Applications of Computational Statistics in the Environmental Sciences: Hierarchical Bayes and MCMC Methods*, 145–169. Oxford University Press: Oxford. 990, 997, 998, 999, 1000

Williams, R. J. (1992). "Simple statistical gradient-following algorithms for connectionist reinforcement learning." *Machine Learning*, 8(3): 229–256. 993

Winn, J. and Bishop, C. M. (2005). "Variational message passing." *Journal of Machine Learning Research*, 6: 661–694. MR2249835. 992

Xu, M., Quiroz, M., Kohn, R., and Sisson, S. A. (2019). "Variance reduction properties of the reparameterization trick." In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2711–2720. 993

Zeiler, M. D. (2012). "ADADELTA: An adaptive learning rate method." arXiv:1212.5701. 993, 999, 1006

Zhang, F. and Gao, C. (2020). "Convergence rates of variational posterior distributions." *The Annals of Statistics*, 48(4): 2180–2207. MR4134791. doi: https://doi.org/10.1214/19-AOS1883. 990

Zhang, T. (2006). "Information-theoretic upper and lower bounds for statistical estimation." *IEEE Transactions on Information Theory*, 52(4): 1307–1321. MR2241190. doi: https://doi.org/10.1109/TIT.2005.864439. 990

Acknowledgments

We thank the Editor, the Associate Editor and two anonymous referees for helping to improve both the content and the presentation of the article. We thank Mevin Hooten for his help with the Eurasian collared-dove data. We thank Linda Tan for her comments on an early version of this manuscript.