Gaussian variational approximation for high-dimensional state space models

Matias Quiroz[1,2]

Collaborators: David Nott (NUS) and Robert Kohn (UNSW)

[1]School of Mathematical and Physical Sciences, University of Technology Sydney

[2]ARC Centre of Excellence for Mathematical & Statistical Frontiers

June 2019

What my talk is about

- **An approach** to "Fitting **complex dynamic** models to data".
- **Demonstration** of the methodology in (i) **spatio-temporal application** and (ii) **financial application**.
- **Bayesian** approach to inference

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y), \quad p(y) = \int p(y|\theta)p(\theta)d\theta.$$

Pros

- **Coherent** uncertainty quantification of unknown parameters θ.
- Accounts for **parameter uncertainty** in predictions:

$$p(\tilde{y}|y) = \int p(\tilde{y}|y,\theta)p(\theta|y)d\theta.$$

- Transparent use of **prior knowledge** $p(\theta)$.
- It is **the right thing to do**.

Con

- **Computationally hard** (unfortunately, **VERY hard**).

What my talk is about

- **An approach** to "Fitting **complex dynamic** models to data".
- **Demonstration** of the methodology in (i) **spatio-temporal application** and (ii) **financial application**.
- **Bayesian** approach to inference

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y), \quad p(y) = \int p(y|\theta)p(\theta)d\theta.$$

Pros

- **Coherent** uncertainty quantification of unknown parameters θ.
- Accounts for **parameter uncertainty** in predictions:

$$p(\tilde{y}|y) = \int p(\tilde{y}|y,\theta)p(\theta|y)d\theta.$$

- Transparent use of **prior knowledge** $p(\theta)$.
- It is **the right thing to do**.

Con

- **Computationally hard** (unfortunately, **VERY hard**).

- Why so **hard**?
  Use of **exact methods** such as Markov Chain Monte Carlo (**MCMC**).

What is "complex + dynamic"?

- Let $y := (y_1, \ldots, y_T)^\top$ be an (**observed**) "time"-series.
- Let $X := (X_0^\top, \ldots, X_T^\top)^\top$ be the latent (**unobserved**) state. X_t is **dynamic** (**time-varying**).
- Let $\zeta := (\zeta_y^\top, \zeta_X^\top)^\top$ be the vector of **static** parameters (**non time-varying**).
- The **data generating process**

$$\begin{aligned} y_t | X_t = x_t &\sim m_t(\cdot | x_t, \zeta_y), \quad t = 1, \ldots, T \\ X_t | X_{t-1} = x_{t-1} &\sim s_t(\cdot | x_{t-1}, \zeta_X), \quad t = 1, \ldots, T \\ X_0 &\sim p(\cdot | \zeta_X). \end{aligned}$$

- y_t, $t = 1, \ldots, T$ assumed **conditionally independent** given X.
- **Known**: y. **Unknown**: $\theta := (X, \zeta)^\top$. Crank the **Bayesian machine**:

$$p(\theta | y) \propto p(y | \theta) p(\theta) = p(y | X, \zeta) p(X | \zeta) p(\zeta).$$

- Obviously a **dynamic** model. But why **complex**?

What is "complex + dynamic"?, cont.

- **Usual complex setting**:
  $m_t()$ [density of **observation equation**] and $s_t()$ [density of **state equation**] are non-Gaussian (**Kalman filtering not possible**).
- **Our complex setting**:
  In addition to non-Gaussianity, X_t is **high-dimensional**. This makes $\dim(\theta)$ **HUGE** ($t = 0, \ldots, T$).
- In this **"complex + dynamic"** setting, **MCMC can be very hard**.
- **Our research**:
  Develops a **Variational Bayes methodology** for inference in this **high-dimensional and complex setting**.

Variational Inference (VI) - a crash course

- Finds an **approximate posterior** $q_\lambda(\theta)$ indexed by variational parameters λ.
- **In our research**:

$$q_\lambda(\theta) = \mathcal{N}\left(\theta | \mu_\lambda, \Sigma_\lambda\right), \quad \lambda = (\mu_\lambda, \text{vech}(\Sigma_\lambda))^\top.$$

- VB finds a λ such that $q_\lambda(\theta) \approx p(\theta|y)$ in **"some sense"**.
- Alternative "semi-parametric" approach: **Mean-Field** (MF) approximation.
- **MF** assumes $q_\lambda(\theta) = \prod_{i=1}^{\dim(\theta)} q_{\lambda_i}(\theta_i)$ [Ormerod and Wand, 2010].
- **MF** cannot capture **posterior dependencies**.

- **"Some sense"**: $q_{\lambda_{\text{opt}}}(\theta)$ minimizes the **Kullback-Leibler** (KL) divergence between $q_\lambda(\theta)$ and $p(\theta|y)$

$$\mathrm{KL}\left(q_\lambda(\theta)|p(\theta|y)\right) = \int \log\left(\frac{q_\lambda(\theta)}{p(\theta|y)}\right) q_\lambda(\theta)d\theta.$$

- Minimizing KL is equivalent to maximizing **Evidence Lower BOund** (ELBO)

$$\mathcal{L}(\lambda) = \mathrm{E}_{q_\lambda}\left[\log h(\theta) - \log q_\lambda(\theta)\right] = \int \left(\log h(\theta) - \log q_\lambda(\theta)\right) q_\lambda(\theta)d\theta,$$

with $h(\theta) := p(y|\theta)p(\theta)$. Name **ELBO** due to $\log p(y) \geq \mathcal{L}(\lambda)$.

- **VI** approximates probability densities **through optimization**.
- **Optimization** much easier than **simulation** in high-dimensional spaces.

Variational inference - a crash course, cont.

- **VB only needs to find** λ_{opt} [Instead of obtaining 100Ks MCMC samples].

VB gradient ascent optimization

While **ELBO** has **not converged** do

- $\lambda^{(j)} = \lambda^{(j-1)} + \eta_j \nabla_\lambda \mathcal{L}(\lambda^{(j-1)})$
- $j = j + 1$

- Looks straightforward... but **some problems** arise...
- **Three problems** (among others).
- **Problem #1**:
  Neither $\mathcal{L}(\lambda)$ nor $\nabla_\lambda \mathcal{L}(\lambda)$ are analytically tractable...
- **Remedy for Problem #1**: Monte Carlo integration.
  The ELBO (and its gradient) is just an expectation wrt. $q_\lambda(\theta)$. Unbiased estimates are **trivial to obtain**...
- ... gives rise to **Problem #2**: the gradient estimate has a **huge** variance.

- Remedy for **Problem #2**: The ReParameterization (RP) trick [Kingma and Welling, 2014].

  Reparameterize $\theta = u(\lambda, \omega)$, with $\omega \sim f$, f a density independent of λ.

  Example for Gaussian VB: $\theta \sim \mathcal{N}(\mu_\lambda, \Sigma_\lambda)$ can be obtained through

$$\theta = \mu_\lambda + C_\lambda \omega,\ \omega \sim \mathcal{N}\left(0, I_{\dim(\theta)}\right),\ \Sigma_\lambda = C_\lambda C_\lambda^\top,$$

  i.e. $u(\lambda, \omega) = \mu_\lambda + C_\lambda \omega$. **Then**,

$$\begin{aligned} \mathcal{L}(\lambda) &= \mathrm{E}_f\left[\log h(u(\lambda,\omega)) - \log q_\lambda(u(\lambda,\omega))\right] \\ \nabla_\lambda \mathcal{L}(\lambda) &= \mathrm{E}_f\left[\nabla_\lambda \log h(u(\lambda,\omega)) - \nabla_\lambda \log q_\lambda(u(\lambda,\omega))\right]. \end{aligned}$$

- **Straightforward** to estimate both $\nabla_\lambda \mathcal{L}(\lambda)$ (also $\mathcal{L}(\lambda)$) by MC integration.
- We **demystify** the success of the **RP trick** [Xu et al., 2019, AISTATS].

- The **stochastic version** of the optimization algorithm:

VB Stochastic gradient ascent optimization

While **ELBO** has **not converged** do

- $\lambda^{(j)} = \lambda^{(j-1)} + \eta_j \widehat{\nabla_\lambda \mathcal{L}}(\lambda^{(j-1)})$
- $j = j + 1$

The problem of our difficult (high-dimensional) setting

- **Problem #3** (our focus): Recall

  $q_\lambda(\theta) = \mathcal{N}(\theta|\mu_\lambda, \Sigma_\lambda)$, $\lambda = (\mu_\lambda, \mathrm{vech}(\Sigma_\lambda))^\top$, $\dim(\lambda) = O(d_\theta^2)$, $d_\theta = \dim\theta$

  Too many variational parameters (λ) to optimize over.
- **Idea**: Look for a **parsimonious yet flexible** Σ_λ.
- Assuming a diagonal Σ_λ gives $\lambda = O(d_\theta)$, but **NO** posterior dependence.
- Utilize **statistical ideas / properties of the statistical model** to **impose sparseness** on λ.

The problem of our difficult (high-dimensional) setting

- **Problem #3** (our focus): Recall

  $q_\lambda(\theta) = \mathcal{N}(\theta|\mu_\lambda, \Sigma_\lambda)$, $\lambda = (\mu_\lambda, \mathrm{vech}(\Sigma_\lambda))^\top$, $\dim(\lambda) = O(d_\theta^2)$, $d_\theta = \dim\theta$

  Too many variational parameters (λ) to optimize over.

- **Idea**: Look for a **parsimonious yet flexible** Σ_λ.
- Assuming a diagonal Σ_λ gives $\lambda = O(d_\theta)$, but **NO** posterior dependence.
- Utilize **statistical ideas / properties of the statistical model** to **impose sparseness** on λ.
  1. **Low rank approximation** of Σ_λ [Ong et al., 2018].

The problem of our difficult (high-dimensional) setting

- **Problem #3** (our focus): Recall

  $$q_\lambda(\theta) = \mathcal{N}(\theta|\mu_\lambda, \Sigma_\lambda), \; \lambda = (\mu_\lambda, \text{vech}(\Sigma_\lambda))^\top, \; \dim(\lambda) = O(d_\theta^2), \; d_\theta = \dim\theta$$

  Too many variational parameters (λ) to optimize over.

- **Idea**: Look for a **parsimonious yet flexible** Σ_λ.
- Assuming a diagonal Σ_λ gives $\lambda = O(d_\theta)$, but **NO** posterior dependence.
- Utilize **statistical ideas / properties of the statistical model** to **impose sparseness** on λ.
  1. **Low rank approximation** of Σ_λ [Ong et al., 2018].
  2. **Impose 0s** in $\Omega_\lambda = \Sigma_\lambda^{-1}$ for the pair of (θ_i, θ_j) that are conditionally independent in the posterior (property of the Gaussian) [Tan and Nott, 2018].

The problem of our difficult (high-dimensional) setting

- **Problem #3** (our focus): Recall

  $$q_\lambda(\theta) = \mathcal{N}(\theta|\mu_\lambda, \Sigma_\lambda), \ \lambda = (\mu_\lambda, \mathrm{vech}(\Sigma_\lambda))^\top, \ \dim(\lambda) = O(d_\theta^2), \ d_\theta = \dim\theta$$

  Too many variational parameters (λ) to optimize over.

- **Idea**: Look for a **parsimonious yet flexible** Σ_λ.
- Assuming a diagonal Σ_λ gives $\lambda = O(d_\theta)$, but **NO** posterior dependence.
- Utilize **statistical ideas / properties of the statistical model** to **impose sparseness** on λ.
  1. **Low rank approximation** of Σ_λ [Ong et al., 2018].
  2. **Impose 0s** in $\Omega_\lambda = \Sigma_\lambda^{-1}$ for the pair of (θ_i, θ_j) that are conditionally independent in the posterior (property of the Gaussian) [Tan and Nott, 2018].
- **Example** of 2.: A state space model (θ_t is the unobserved state at t)

  $$p(\theta_{0:T}|y) \propto p(\theta_0) \prod_{t=1}^{T} p(\theta_t|\theta_{t-1}) p(y_t|\theta_t)$$

  would have a **tridiagonal structure** of Ω_λ. Hence $\lambda = O(d_\theta)$.

Our approach to parsimonious VB

- **Suppose that** $\dim(X_t) = p$ and $\dim(\zeta) = P$,

$$\theta = (X_0^\top, \ldots, X_T^\top, \zeta)^\top \in \mathbb{R}^{p(T+1)+P},$$

- We assume a **dynamic factor model** for the **high-dimensional** state:

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

  $D_t = \mathrm{diag}(\delta_{1t}, \ldots, \delta_{pt})$, $B \in \mathbb{R}^{p\times q}$, q (# of factors) $\ll p$, $z_t \in \mathbb{R}^q$ with $\mathrm{E}[z_t] = 0$ and $\mathrm{V}[z_t] = \Sigma_{z_t}$ in $\mathbb{R}^{q\times q}$. **Implies** $X_t \sim \mathcal{N}(\mu_t, B\Sigma_{z_t}B^\top + D_t^2)$.
- Note that θ **has been reduced** to

$$\rho = (z_0^\top, \ldots, z_T^\top, \zeta)^\top \in \mathbb{R}^{q(T+1)+P}.$$

- **Markovian structure** for z_t: z_t depends only on z_{t+1} and z_{t-1}.
  Sparse precision matrix for $z = (z_0^\top, \ldots, z_T^\top)^\top$!

Our approach to parsimonious VB, cont.

- Let C_z denote the **Cholesky factor** of Ω_z (the precision matrix of z).
- C_z takes the form (each $C_{xx} \in \mathbb{R}^{q \times q}$ and **lower triangular**)

$$C_z = \begin{bmatrix} C_{00} & 0 & 0 & \dots & 0 & 0 \\ C_{10} & C_{11} & 0 & \dots & 0 & 0 \\ 0 & C_{21} & C_{22} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & C_{T-1,T-1} & 0 \\ 0 & 0 & \dots & \dots & C_{T,T-1} & C_{TT} \end{bmatrix}.$$

- The resulting **precision matrix** $\Omega_z = C_z C_z^\top$

$$\Omega_z = \begin{bmatrix} \Omega_{00} & \Omega_{10}^\top & 0 & \dots & 0 & 0 \\ \Omega_{10} & \Omega_{11} & \Omega_{21}^\top & \dots & 0 & 0 \\ 0 & \Omega_{21} & \Omega_{22} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \dots & \vdots \\ 0 & 0 & 0 & \dots & \Omega_{T-1,T-1} & \Omega_{T,T-1}^\top \\ 0 & 0 & 0 & \dots & \Omega_{T,T-1} & \Omega_{TT} \end{bmatrix}.$$

Application 1: Modeling invasive species

- **Dataset**: spread of the **Eurasian collared-dove** across North America [Wikle and Hooten, 2006].
- $y_{s_i t} =$ "Number of doves at location s_i (lat, lon) in year t".
- $i = 1, \ldots, p = 111$, and $t = 1, \ldots, T = 18$ (1986-2003).
- We use our **VB approximation** with $q = 4\,(\ll 111)$ factors.
- In total, this example has $\dim(\theta) = 4,223$.
  - **Gaussian VB** with **unrestricted** covariance matrix: $\dim(\lambda) = 8,923,199$.
  - **Our Gaussian VB** with $q = 4$: $\dim(\lambda) = 11,587$.

- **The model** (high-dimensional state vector in **red**)

$$
\begin{aligned}
y_t | v_t &\sim \text{Poisson}(N_t \exp(v_t)) \quad y_t, v_t \in \mathbb{R}^p, N_t \in \mathbb{N} \\
v_t | u_t, \sigma_\epsilon^2 &\sim \mathcal{N}(u_t, \sigma_\epsilon^2 I_p), \quad u_t \in \mathbb{R}^p, I_p \in \mathbb{R}^{p \times p}, \sigma_\epsilon^2 \in \mathbb{R}^+ \\
u_t | u_{t-1}, \psi, \sigma_\eta^2 &\sim \mathcal{N}(H(\psi) u_{t-1}, \sigma_\eta^2 I_p), \quad \psi \in \mathbb{R}^p, H(\psi) \in \mathbb{R}^{p \times p}, \sigma_\eta^2 \in \mathbb{R}^+,
\end{aligned}
$$

Application 1: Modeling invasive species, cont.

- **The model** (high-dimensional state vector in **red**)

$$
\begin{aligned}
y_t | v_t &\sim \text{Poisson}(N_t \exp(v_t)) \quad y_t, v_t \in \mathbb{R}^p, N_t \in \mathbb{N} \\
v_t | u_t, \sigma_\epsilon^2 &\sim \mathcal{N}(u_t, \sigma_\epsilon^2 I_p), \quad u_t \in \mathbb{R}^p, I_p \in \mathbb{R}^{p\times p}, \sigma_\epsilon^2 \in \mathbb{R}^+ \\
u_t | u_{t-1}, \psi, \sigma_\eta^2 &\sim \mathcal{N}(H(\psi) u_{t-1}, \sigma_\eta^2 I_p), \quad \psi \in \mathbb{R}^p, H(\psi) \in \mathbb{R}^{p\times p}, \sigma_\eta^2 \in \mathbb{R}^+,
\end{aligned}
$$

- **Priors** [Wikle and Hooten, 2006]

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

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

- **Key parameters**: Diffusion coefficients ψ.
  Spatial dependence modeled via $\Phi\alpha$, where Φ has l orthonormal eigenvectors with the largest eigenvalues of a **spatial correlation matrix**.

Application 1: Validating accuracy of VB, part 1

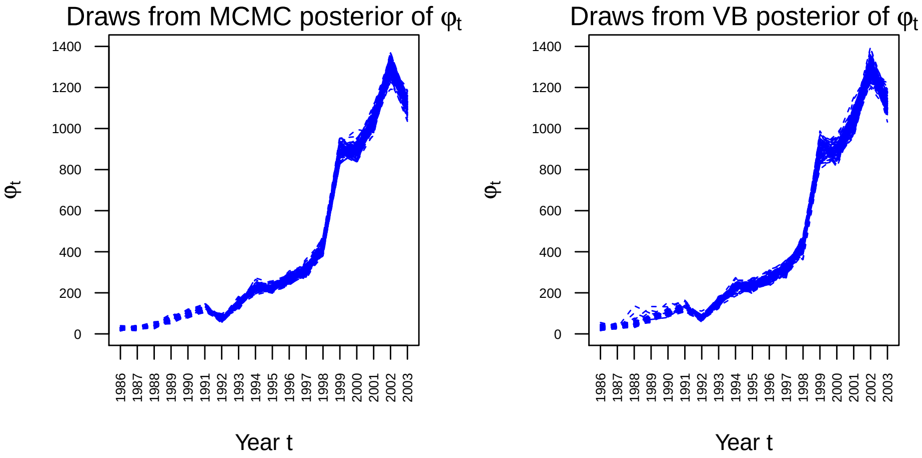


Figure 1 : 100 samples from the posterior distribution of the **sum of dove intensity over the spatial grid for each year** $\varphi_t = \sum_i \exp(v_{it})$.

Application 1: Validating accuracy of VB, part 2

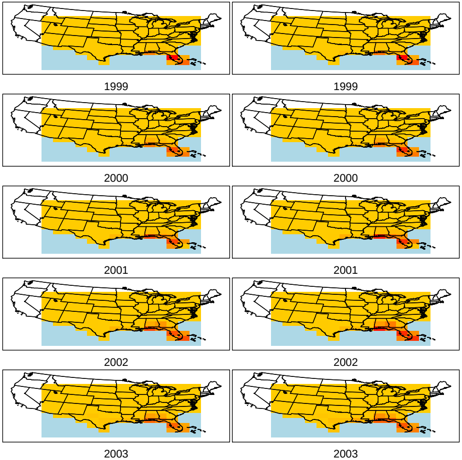


Figure 2 : Posterior mean of the dove intensity for $\varphi_{it} = \exp(v_{it})$ by MCMC (left panel) and VB (right panel) for the years 1999-2003 and for $i = 1, \ldots, p = 111$.

Application 1: Validating accuracy of VB, part 3

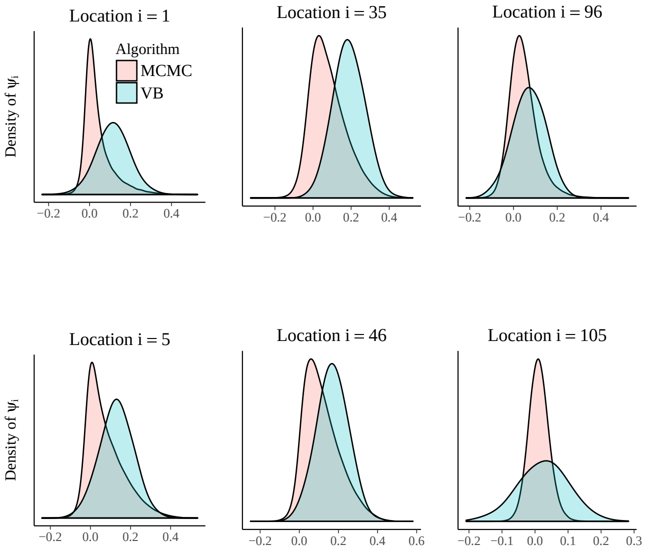


Figure 3 : Posterior (VB and MCMC) for some diffusion coefficients Ψ.

Application 2: Multivariate stochastic volatility via Wishart Processes

- **Joint volatility model** for k assets [Philipov and Glickman, 2006].
- **Model**: $y_t = (y_{1t}, \dots, y_{kt})^\top \in \mathbb{R}^k$, for $t = 1, \dots, T$, follows

$$
\begin{aligned}
y_t | \Sigma_t &\sim \mathcal{N}(0, \Sigma_t) \\
\Sigma_t^{-1} | \Sigma_{t-1}^{-1} &\sim \mathrm{Wish}(\nu, S_{t-1}), \quad S_{t-1} = \frac{1}{\nu} H \left(\Sigma_{t-1}^{-1} \right)^d H^\top,
\end{aligned}
$$

with $0 < d < 1, \nu > k$ and H is the **Cholesky factor** of $A = HH^\top \in \mathbb{R}_+^k$.

Application 2: Multivariate stochastic volatility via Wishart Processes

- **Joint volatility model** for k assets [Philipov and Glickman, 2006].
- **Model**: $y_t = (y_{1t}, \dots, y_{kt})^\top \in \mathbb{R}^k$, for $t = 1, \dots, T$, follows

$$
\begin{aligned}
y_t | \Sigma_t &\sim \mathcal{N}(0, \Sigma_t) \\
\Sigma_t^{-1} | \Sigma_{t-1}^{-1} &\sim \mathrm{Wish}(\nu, S_{t-1}), \quad S_{t-1} = \frac{1}{\nu} H \left(\Sigma_{t-1}^{-1} \right)^d H^\top,
\end{aligned}
$$

  with $0 < d < 1, \nu > k$ and H is the **Cholesky factor** of $A = HH^\top \in \mathbb{R}_+^k$.
  Priors: $\nu - k \sim \mathrm{Gamma}(\alpha_0, \beta_0)$, $d \sim \mathrm{Unif}(0, 1)$, $A \sim \mathrm{Inv\text{-}Wish}(\nu_0, Q_0^{-1})$.
- **Posterior** of interest $p(\theta | y_{1:T})$, $\theta = (\Sigma_{1:T}, A, \nu, d)$.
- **NOTE 1**: This is a **state space model**. Let $X_t = \mathrm{vech}(\Sigma_t)$.
  1. **Measurement equation** $y_t | X_t$ is Gaussian.
  2. **State transition** $m_t(X_t | X_{t-1})$ is inverse Wishart.
- **NOTE 2**: The state is **high-dimensional**!
  1. $k = 5$ assets gives $p = 15$ states.
  2. $k = 12$ assets gives $p = 78$ states.
  3. Suppose we had $k = 100$ assets. Then $p = 5,050$ (!!!).

Application 2: Multivariate stochastic volatility via Wishart Processes, cont.

- The Gibbs sampler in [Philipov and Glickman, 2006] (**Different sampling** of A^{-1} due to an error discovered by [Rinnergschwentner et al., 2012]).

Gibbs sampling for the model

At iteration i, cycle through the updates

- **For** $t = 1, \ldots, T-1$,

  $\Sigma_t^{-1}|\text{rest}$ - Independent M-H with Wishart proposal
- $\Sigma_T^{-1}|\text{rest}$ - Perfect sampling from Wishart
- $A^{-1}|\text{rest}$ - RW M-H with Wishart proposal with mean $A^{-1(i-1)}$
- $\nu|\text{rest}$ - Perfect sampling by inverse cdf (scalar)
- $d|\text{rest}$ - Perfect sampling by inverse cdf (scalar)

Application 2: Multivariate stochastic volatility via Wishart Processes, cont.

- Independent (and Random Walk) proposals **do not work** in high dimensions.
- **[Philipov and Glickman, 2006]** report acceptance probabilities of 0.005 for Σ_t $(t < T)$ when $k = 12$.
- **Sparsity** obtained using $q = 4$ factors
  1. For $k = 5$. $\dim(\theta) = 1,517$. **Saturated VB**: $1,152,920$. **Our VB**: $5,109$.
  2. For $k = 12$. $\dim(\theta) = 7,880$, **Saturated VB**: $31,059,020$. **Our VB**: $10,813$.
- The Gibbs sampler **does not convergence**. How to evaluate VB?
- A **"predictive oracle"** approach using simulated data.

Application 2: The oracle predictive density

- The **one-step ahead** oracle predictive density

$$
\begin{aligned}
p(y_{T+1}|y_{1:T}, \zeta^{\text{true}}) &= \int p(y_{T+1}, X_{T+1}|y_{1:T}, \zeta^{\text{true}}) dX_{T+1} \\
&= \int p(y_{T+1}|X_{T+1}) p(X_{T+1}|y_{1:T}, \zeta^{\text{true}}) dX_{T+1}.
\end{aligned}
$$

- **The posterior** of X_{T+1}

$$
\begin{aligned}
p(X_{T+1}|y_{1:T}, \zeta^{\text{true}}) &= \int p(X_{T+1}, X_T|y_{1:T}, \zeta^{\text{true}}) dX_T \\
&= \int p(X_{T+1}|X_T, \zeta^{\text{true}}) p(X_T|y_{1:T}, \zeta^{\text{true}}) dX_T,
\end{aligned}
$$

- Samples from $p(X_T|y_{1:T}, \zeta^{\text{true}})$ are obtained by **the particle filter**.
- The above provides a **"ground truth"** for predicting y_{T+1}.
- Find the variational predictive and **compare to** the oracle for different T.

Application 2: The VB predictive density

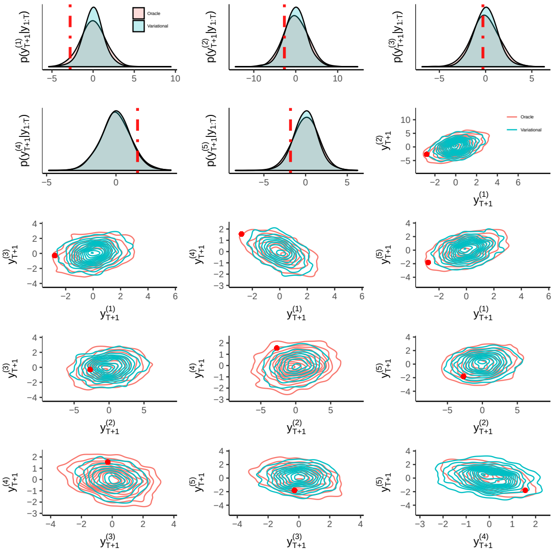
- The **one-step ahead** oracle predictive averages over:
  1. The variational posterior of **the static model**.
  2. The variational posterior of **the state parameters**.
- **Mathematically**

$$
\begin{aligned}
p(y_{T+1}|y_{1:T}) &= \int\int p(y_{T+1}, X_{T+1}, \zeta|y_{1:T})dX_{T+1}d\zeta \\
&= \int\int p(y_{T+1}|X_{T+1})p(X_{T+1}, \zeta|y_{1:T})dX_{T+1}d\zeta \\
&= \int\int p(y_{T+1}|X_{T+1})p(X_{T+1}|y_{1:T}, \zeta)p(\zeta|y_{1:T})dX_{T+1}d\zeta.
\end{aligned}
$$

- Setting $p(\zeta|y_{1:T}) = q(\zeta)$ and then

$$
\begin{aligned}
p(X_{T+1}|y_{1:T}, \zeta) &= \int p(X_{T+1}, X_T|y_{1:T}, \zeta)dX_T \\
&= \int p(X_{T+1}|X_T, \zeta)p(X_T|y_{1:T}, \zeta)dX_T,
\end{aligned}
$$

using the **variational approximation** $p(X_T|y_{1:T}, \zeta) = q(X_T)$.


Oracle
Variational

And what about computational gains?

- In the **spatio-temporal example**, **VB about 7 times faster** than MCMC. **MCMC still OK for this model**.
- In the **multivariate stochastic volatility via Wishart processes** with **5 assets**, **VB was about 30 times faster**...
- ... and with **12 assets**, **VB was infinitely many times faster** (**MCMC had 0 acceptance probability!**).

Concluding remarks and future research

- **VB is a useful alternative** to MCMC.
- **Gaussian VB approximation** + **RP trick** + **Sensible structure of Σ_λ** allows fitting **extremely high-dimensional models**.
- Future work:
  - **More flexible** variational families.
  - **More applications**!

Thank you for listening!

You can find our paper on
https://arxiv.org/abs/1801.07873

Slides uploaded on **www.matiasquiroz.com/news**

Questions?

References I

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

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, (To appear).

Ormerod, J. T. and Wand, M. P. (2010).
Explaining variational approximations.
The American Statistician, 64:140–153.

Philipov, A. and Glickman, M. E. (2006).
Multivariate stochastic volatility via Wishart processes.
Journal of Business & Economic Statistics, 24(3):313–328.

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.

References II

Tan, L. S. and Nott, D. J. (2018).
Gaussian variational approximation with sparse precision matrices.
Statistics and Computing, 28(2):259–275.

Wikle, C. K. and Hooten, M. B. (2006).
Hierarchical Bayesian spatio-temporal models for population spread.
In Clark, J. S. and Gelfand, A., editors, *Applications of computational statistics in the environmental sciences: hierarchical Bayes and MCMC methods*, pages 145–169. Oxford University Press: Oxford.

Xu, M., Quiroz, M., Kohn, R., and Sisson, S. A. (2019).
Variance reduction properties of the reparameterization trick.
In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2711–2720. PMLR.