Bayesian forecasting for high-dimensional state-space models: A variational approach

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

- **Forecasting** for **high-dimensional time-varying parameter** models.
- **Demonstration** of the methodology in a **financial application**.
- **Bayesian** approach to inference

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

- **Bayes** accounts for **parameter uncertainty** in predictions:

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

What my talk is about

- **Forecasting** for **high-dimensional time-varying parameter** models.
- **Demonstration** of the methodology in a **financial application**.
- **Bayesian** approach to inference

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

- **Bayes** accounts for **parameter uncertainty** in predictions:

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

- Simulating from the **posterior predictive** $p(\tilde{y}|y)$ is trivial ...

$p(\tilde{y},\theta|y) = p(\tilde{y}|y,\theta)p(\theta|y)$ [discarding θ gives samples $\tilde{y}$ from $p(\tilde{y}|y)$]

What my talk is about

- **Forecasting** for **high-dimensional time-varying parameter** models.
- **Demonstration** of the methodology in a **financial application**.
- **Bayesian** approach to inference

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

- **Bayes** accounts for **parameter uncertainty** in predictions:

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

- Simulating from the **posterior predictive** $p(\tilde{y}|y)$ is trivial ...

$$p(\tilde{y},\theta|y) = p(\tilde{y}|y,\theta)p(\theta|y) \quad [\text{discarding } \theta \text{ gives samples } \tilde{y} \text{ from } p(\tilde{y}|y)]$$

- ... but only if we can **sample from** $p(\theta|y)$ in first place...

What my talk is about

- **Forecasting** for **high-dimensional time-varying parameter** models.
- **Demonstration** of the methodology in a **financial application**.
- **Bayesian** approach to inference

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

- **Bayes** accounts for **parameter uncertainty** in predictions:

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

- Simulating from the **posterior predictive** $p(\tilde{y}|y)$ is trivial ...

$$p(\tilde{y},\theta|y) = p(\tilde{y}|y,\theta)p(\theta|y) \quad \text{[discarding } \theta \text{ gives samples } \tilde{y} \text{ from } p(\tilde{y}|y)]$$

- ... but only if we can **sample from** $p(\theta|y)$ in first place...
- ... which is **VERY HARD** for **complex dynamic** models...

What is "complex + dynamic"?

- Let $y := (y_1, \dots, y_T)^\top$ be an (**observed**) time-series.
- Let $X := (X_0^\top, \dots, X_T^\top)^\top$ 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**

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

- **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**?

- **Usual complex setting**:
  $m_t()$ and $s_t()$ are non-Gaussian (**Kalman filtering not possible**).

What is "complex + dynamic"?, cont.

- **Usual complex setting**:
  $m_t()$ and $s_t()$ 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**.

- Develop a **Variational Inference** (VI) methodology in this **high-dimensional and complex setting**.

One slide of Variational Inference (VI)

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

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

- VI finds a λ such that $q_\lambda(\theta) \approx p(\theta|y)$ in **"some sense"**.
- **"Some sense"**: $q_{\lambda_{\text{opt}}}(\theta)$ minimizes the **Kullback-Leibler** (KL) divergence between $q_\lambda(\theta)$ and $p(\theta|y)$. Hard to compute **KL**.
- Easier (and equivalent) to maximize **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 $h(\theta) := p(y|\theta)p(\theta)$.

- **Stochastic optimization**: Monte Carlo to compute $\widehat{\nabla_\lambda} \mathcal{L}(\theta)$.

The problem of our difficult (high-dimensional) setting

- **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.

The problem of our difficult (high-dimensional) setting

- **Recall**

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

  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

- **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].

The problem of our difficult (high-dimensional) setting

- **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

- **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 VI

- 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** $X_t \sim \mathcal{N}(\mu_t, B\Sigma_{z_t}B^\top + D_t^2)$.

Our approach to parsimonious VI

- 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** $X_t \sim \mathcal{N}(\mu_t, B\Sigma_{z_t}B^\top + D_t^2)$.

- **Markovian** $\implies$ z_t depends only on z_{t+1} and z_{t-1} in the posterior...

Our approach to parsimonious VI

- 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** $X_t \sim \mathcal{N}(\mu_t, B\Sigma_{z_t}B^\top + D_t^2)$.

- **Markovian** $\implies$ z_t depends only on z_{t+1} and z_{t-1} in the posterior...

- ... **sparse precision matrix** $\Omega_z = \Sigma_z^{-1}$ for $z = (z_0^\top, \ldots, z_T^\top)^\top$!

Our approach to parsimonious VI

- 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** $X_t \sim \mathcal{N}(\mu_t, B\Sigma_{z_t}B^\top + D_t^2)$.

- **Markovian** $\implies$ z_t depends only on z_{t+1} and z_{t-1} in the posterior...
- ... **sparse precision matrix** $\Omega_z = \Sigma_z^{-1}$ for $z = (z_0^\top, \ldots, z_T^\top)^\top$!
- **Massive reduction** of in the number of variational parameters...
- ... while **capturing the dependencies** in the posterior $p(\theta|y)$.

Multivariate stochastic volatility via Wishart Processes

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

Multivariate stochastic volatility via Wishart Processes

- **Joint volatility model** for k assets [Philipov and Glickman, 2006].
- **Model**: $y_t = (y_{1t}, \ldots, y_{kt})^\top \in \mathbb{R}^k$, for $t = 1, \ldots, T$, 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: $\nu - k \sim \mathrm{Gamma}(\alpha_0, \beta_0)$, $d \sim \mathrm{Unif}(0,1)$, $A \sim \text{Inv-Wish}(\nu_0, Q_0^{-1})$.
- **Posterior** of interest $p(\theta | y_{1:T})$, $\theta = (\Sigma_{1:T}, A, \nu, d)$.

Multivariate stochastic volatility via Wishart Processes

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

Multivariate stochastic volatility via Wishart Processes

- **Joint volatility model** for k assets [Philipov and Glickman, 2006].
- **Model**: $y_t = (y_{1t}, \ldots, y_{kt})^\top \in \mathbb{R}^k$, for $t = 1, \ldots, T$, 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 $p = 5,050$ (!!!).

Multivariate stochastic volatility via Wishart Processes, cont.

- Transform θ to be **unrestricted** (support of Gaussian $\mathbb{R}^{d_\theta}$).
- **Sparsity** obtained using $q = 4$ factors in our method.
  1. For $k = 5$. $\dim(\theta) = 1,517$. **Saturated VI**: $1,152,920$. **Our VI**: $5,109$.
  2. For $k = 12$. $\dim(\theta) = 7,880$, **Saturated VI**: $31,059,020$. **Our VI**: $10,813$.

Validating our results

- **Error in the Gibbs sampler** of [Philipov and Glickman, 2006] (see [Rinnergschwentner et al., 2012]).
- How to validate our approach? We need a **"ground truth"**.
- A **"predictive oracle"** approach using simulated data.
- Oracle is the **"ground truth"**. Compare **VI predictions** to that of the oracle.

Predictive densities to compare

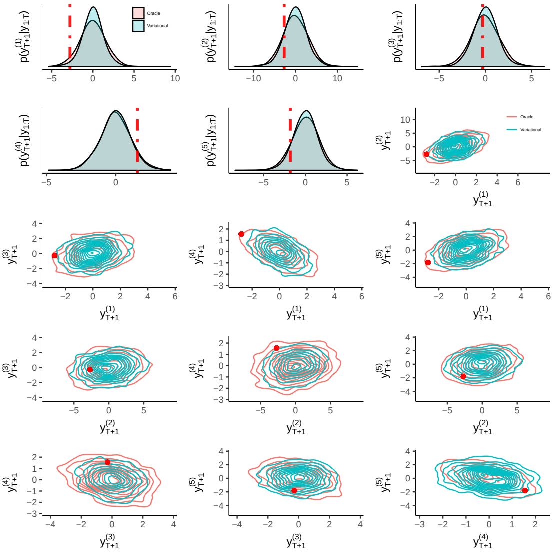
- The **one-step ahead** **oracle predictive** density

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

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

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

- 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}.
- The **variational predictive** similarly obtained but averages over:
  1. The variational posterior of **the static model**.
  2. The variational posterior of X_T.


Oracle
Variational
$p(y^{(1)}_{T+1}|y_{1:T})$
$p(y^{(2)}_{T+1}|y_{1:T})$
$p(y^{(3)}_{T+1}|y_{1:T})$
$p(y^{(4)}_{T+1}|y_{1:T})$
$p(y^{(5)}_{T+1}|y_{1:T})$
$y^{(1)}_{T+1}$
$y^{(2)}_{T+1}$
$y^{(3)}_{T+1}$
$y^{(4)}_{T+1}$
$y^{(5)}_{T+1}$

Concluding remarks and future research

- **VI** to obtain the posterior predictive in high-dimensional state space models.
- **Gaussian VI approximation** + **Sensible structure of Σ_λ** allows fitting **extremely high-dimensional models**.
- **Performs well** for predictions.
- Future work:
  - **More flexible** variational families.
  - **More applications**!

Thank you for listening!

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

Questions?

References I

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).

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.

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