Fast and Accurate Variational Inference for Large Bayesian VARs with Stochastic Volatility

Joshua C.C. Chan
Purdue University

Xuewen Yu
Purdue University

First version: November 2020
This version: March 2021

Abstract

We propose a new variational approximation of the joint posterior distribution of the log-volatility in the context of large Bayesian VARs. In contrast to existing approaches that are based on *local* approximations, the new proposal provides a *global* approximation that takes into account the entire support of the joint distribution. In a Monte Carlo study we show that the new global approximation is over an order of magnitude more accurate than existing alternatives. We illustrate the proposed methodology with an application of a 96-variable VAR with stochastic volatility to measure global bank network connectedness. Our measure is able to detect the drastic increase in global bank network connectedness much earlier than rolling-window estimates from a homoscedastic VAR.

Keywords: large vector autoregression, stochastic volatility, Variational Bayes, volatility network, connectedness

JEL classifications: C11, C32, C55, G21

1 Introduction

Since the influential work of Banbura, Giannone, and Reichlin (2010), large Bayesian vector autoregressions (VARs) have been widely used to characterize the comovements of a large number of macroeconomic and financial variables.[1] Since a vast empirical literature has demonstrated the importance of allowing for time-varying volatility in small systems,[2] there is a lot of recent work that aims to develop stochastic volatility specifications for large VARs, including Koop and Korobilis (2013), Carriero, Clark, and Marcellino (2016, 2019), Chan (2020a), Chan, Eisenstat, and Strachan (2020) and Kastner and Huber (2020). Despite recent advances, estimating large VARs with flexible stochastic volatility specifications using conventional Markov chain Monte Carlo (MCMC) methods remains computationally intensive.

In view of the computational burden, some recent papers, such as Koop and Korobilis (2018) and Gefang, Koop, and Poon (2019), have adopted an alternative approach of using Variational Bayesian methods to approximate the posterior distributions of large VARs with stochastic volatility. The main advantage of these Variational Bayesian methods is that they are substantially faster than MCMC, especially for high-dimensional models such as large VARs, making estimation of very large systems possible. For example, fitting a 100-variable system takes only a few minutes compared to hours when MCMC is used. However, they are approximate methods — as opposed to MCMC that can be made arbitrarily accurate by increasing the simulation size — and the approximation accuracy depends on the Kullback-Leibler divergence of the approximating density to the posterior density.

Existing approximating densities of the joint distribution of the log-volatility are based on *local* approximations, such as a second-order Taylor expansion of the log target density around a point (e.g., the mode). As such, these approximations are guaranteed to approximate the target density well around the neighborhood of the point of expansion, but their accuracy typically deteriorates rapidly away from the approximation point. In contrast, we propose a *global* approximation of the joint distribution of the log-volatility

[1] Notable examples include Carriero, Kapetanios, and Marcellino (2009), Koop (2013), Banbura, Giannone, Modugno, and Reichlin (2013), Carriero, Clark, and Marcellino (2015), McCracken, Owyang, and Sekhposyan (2015), Ellahie and Ricco (2017) and Morley and Wong (2020).

[2] See, for example, Cogley and Sargent (2005), Primiceri (2005), Clark (2011), D'Agostino, Gambetti, and Giannone (2013), and Cross and Poon (2016).

that takes into account the entire support of the distribution. The key idea is to set up a formal optimization problem to locate the 'best' density within a class of multivariate Gaussian distributions. More specifically, we obtain the density within the family that is the closest to the target posterior distribution, measured by the Kullback-Leibler divergence between the two densities.

To implement the proposed approach, we first reduce the difficult functional optimization problem of finding the minimizer within a family of distributions to a standard vector optimization problem by parameterizing the family of Gaussian distributions. We then solve the associated optimization using the Newton-Raphson method. Since the optimization problem is high-dimensional, we carefully make use of fast band matrix routines to speed up computations. Since the class of Gaussian distributions we construct includes some of the existing Gaussian approximations, the optimal density located under this proposed approach is guaranteed to be a better approximation — in the sense of a smaller Kullback-Leibler divergence — than existing proposals.

We then demonstrate the superior approximation accuracy of the proposed approach relative to existing methods in a Monte Carlo study. In particular, the Monte Carlo results show that the mean squared errors under the proposed global approximation can be more than an order of magnitude smaller than those under existing local approximations.

The proposed methodology is illustrated using an application of a large VAR with stochastic volatility to measure global bank network connectedness. More specifically, we revisit the global bank network application in Demirer, Diebold, Liu, and Yilmaz (2018) — they consider a 96-variable homoscedastic VAR to measure global bank network connectedness. Since the connectedness measures are functions of the error covariance matrix, how it is modeled is likely to be important for the analysis. We therefore extend their analysis by allowing the error covariance matrix to vary over time via a stochastic volatility process. Using data on bank returns volatility, we find qualitatively similar results. In particular, our results show that North America and Europe are the two largest net transmitters of future volatility uncertainty to the rest of the world, whereas Asia is a large net receiver of future volatility uncertainty from the rest of the world.

While we also find a substantial increase in bank system-wide connectedness at the start of the Great Recession in late 2007, our connectedness measure from the VAR with stochastic volatility shows more pronounced movements. In particular, it is able to detect

the drastic increase in global bank network connectedness much earlier than the measure constructed from a homoscedastic VAR estimated using a fixed rolling-window sample. These results highlight the empirical relevance of using time-varying models in the context of detecting sudden breaks or drastic changes.

The rest of the paper is organized as follows. We first present a reparameterization of the reduced-form VAR with stochastic volatility in Section 2, followed by some discussion of an adaptive Minnesota prior. Section 3 provides an overview of Variational Bayesian methods. We then introduce a new global Gaussian approximation of the joint distribution of the log-volatility in Section 4. Section 5 conducts a Monte Carlo study to provide evidence of superior approximation accuracy of the proposed method relative to existing approaches. It is followed by an application of using a VAR with stochastic volatility to measure global bank network connectedness in Section 6. Lastly, Section 7 concludes and briefly discusses some future research directions.

2 A VAR with Stochastic Volatility

In this section we describe a large VAR with stochastic volatility and then outline an adaptive Minnesota prior. More specifically, we consider the following reparameterization of the standard reduced-form VAR:

$$\mathbf{B}_0\mathbf{y}_t = \mathbf{b} + \mathbf{B}_1\mathbf{y}_{t-1} + \cdots + \mathbf{B}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t^y, \quad \boldsymbol{\varepsilon}_t^y \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t), \tag{1}$$

where $\boldsymbol{\Sigma}_t = \text{diag}(\text{e}^{h_{1,t}}, \ldots, \text{e}^{h_{n,t}})$ is a diagonal matrix and $\mathbf{B}_0$ is a lower triangular matrix with ones on the main diagonal. Each log-volatility $h_{i,t}, i = 1, \ldots, n$, evolves as an independent random walk:

$$h_{i,t} = h_{i,t-1} + \varepsilon_{i,t}^h, \quad \varepsilon_{i,t}^h \sim \mathcal{N}(0, \sigma_{h,i}^2), \tag{2}$$

for $t = 1, \ldots, T$, where the initial condition $h_{i,0}$ is treated an unknown parameter. It is straightforward to check that one can recover the reduced-form intercepts and VAR coefficients by computing $\widetilde{\mathbf{b}} = \mathbf{B}_0^{-1}\mathbf{b}$ and $\widetilde{\mathbf{B}}_j = \mathbf{B}_0^{-1}\mathbf{B}_j, j = 1, \ldots, p$. In addition, the implied reduced-form inverse covariance matrix, or precision matrix, is $\widetilde{\boldsymbol{\Sigma}}_t^{-1} = \mathbf{B}_0'\boldsymbol{\Sigma}_t^{-1}\mathbf{B}_0$, as considered in Cogley and Sargent (2005) and Carriero, Clark, and Marcellino (2019).

Since the covariance matrix $\mathbf{\Sigma}_t$ in (1) is diagonal, we can estimate this recursive system equation by equation without loss of efficiency.[3] Below we first rewrite (1) as n separate univariate regressions. For notational convenience, let b_i denote the i-th element of $\mathbf{b}$ and let $\mathbf{b}_{j,i}$ represent the i-th row of $\mathbf{B}_j$. Then, $\boldsymbol{\beta}_i = (b_i, \mathbf{b}_{1,i}, \ldots, \mathbf{b}_{p,i})'$ is the intercept and VAR coefficients for the i-th equation. Furthermore, let $\boldsymbol{\alpha}_i$ denote the free elements in the i-th row of the impact matrix $\mathbf{B}_0$. Then, the i-th equation of the system in (1) can be rewritten as:

$$y_{i,t} = \widetilde{\mathbf{w}}_{i,t}\boldsymbol{\alpha}_i + \widetilde{\mathbf{x}}_t\boldsymbol{\beta}_i + \varepsilon^y_{i,t}, \quad \varepsilon^y_{i,t} \sim \mathcal{N}(0, \mathrm{e}^{h_{i,t}}),$$

where $\widetilde{\mathbf{w}}_{i,t} = (-y_{1,t}, \ldots, -y_{i-1,t})$ and $\widetilde{\mathbf{x}}_t = (1, \mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p})$. Note that this is a recursive system in which $y_{i,t}$ depends on the contemporaneous variables $y_{1,t}, \ldots, y_{i-1,t}$. Since the system is recursive, the Jacobian of the change of variables from $\boldsymbol{\varepsilon}^y_t$ to $\mathbf{y}_t$ has unit determinant. Hence, the likelihood function has the usual Gaussian form.

Letting $\mathbf{x}_{i,t} = (\widetilde{\mathbf{w}}_{i,t}, \widetilde{\mathbf{x}}_t)$, we can further simplify the i-th equation as:

$$y_{i,t} = \mathbf{x}_{i,t}\boldsymbol{\theta}_i + \varepsilon^y_{i,t}, \quad \varepsilon^y_{i,t} \sim \mathcal{N}(0, \mathrm{e}^{h_{i,t}}), \tag{3}$$

where $\boldsymbol{\theta}_i = (\boldsymbol{\alpha}'_i, \boldsymbol{\beta}'_i)'$ is of dimension $k_i = np + i$. We have therefore rewritten the VAR in (1) as a system of n separate univariate regressions. This representation facilitates equation-by-equation estimation, which substantially speeds up the computations.

To complete the model specification, we assume the following priors on the parameters $\boldsymbol{\theta}_i, h_{i,0}$ and $\sigma^2_{h,i}, i = 1, \ldots, n$:

$$\boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\theta}_{0,i}, \mathbf{V}_{\theta_i}), \quad h_{i,0} \sim \mathcal{N}(0, V_{h_{i,0}}), \quad \sigma^2_{h,i} \sim \mathcal{IG}(\nu_i, S_i),$$

where $\mathcal{IG}(a, b)$ denotes the inverse-gamma distribution with mean $b/(a-1)$. Since large VARs have a lot of parameters, a suitable shrinkage prior on $\boldsymbol{\theta}_i$ is vital. Below we describe a version of the Minnesota prior to elicit $\boldsymbol{\theta}_{0,i}$ and $\mathbf{V}_{\theta_i}$.

We emphasize that even if other hierarchical shrinkage priors on $\boldsymbol{\theta}_i$ are used, such as those in Bhattacharya, Pati, Pillai, and Dunson (2015) and Griffin and Brown (2017), the proposed variational Bayes method can be directly applied. Here we focus on the Minnesota prior for two reasons. First, the Minnesota prior remains the most popular

[3] This recursive estimation approach has been widely used in the literature; see, e.g., Baumeister and Hamilton (2015), Eisenstat, Chan, and Strachan (2016) and Carriero, Clark, and Marcellino (2019).

shrinkage prior for large Bayesian VARs. Second, there is a growing body of empirical evidence to suggest that it is more suitable for macroeconomic data than other hierarchical shrinkage priors; see, for example, Giannone, Lenza, and Primiceri (2017) and Cross, Hou, and Poon (2020).

For a general discussion of the Minnesota prior, we refer the readers to Koop and Korobilis (2010), Karlsson (2013) or Chan (2020b). Below we outline how we elicit $\boldsymbol{\theta}_{0,i}$ and $\mathbf{V}_{\theta_i}$. For growth rates data, we set $\boldsymbol{\theta}_{0,i} = \mathbf{0}$ to shrink the VAR coefficients to zero. For level data, $\boldsymbol{\theta}_{0,i}$ is also set to be zero except for the coefficient associated with the first own lag, which is set to be one. Next, for $\mathbf{V}_{\theta_i}$, we specify it to be diagonal with the k-th diagonal element $V_{\theta_i,k}$ set to be:

$$V_{\theta_i,k} = \begin{cases} \frac{\kappa_1}{l^2}, & \text{for the coefficient on the } l\text{-th lag of variable } i, \\ \frac{\kappa_2 s_i^2}{l^2 s_j^2}, & \text{for the coefficient on the } l\text{-th lag of variable } j, j \neq i, \\ \frac{s_i^2}{s_j^2}, & \text{for the } j\text{-th element of } \boldsymbol{\alpha}_i, \\ 100 s_i^2, & \text{for the intercept,} \end{cases}$$

where s_r^2 denotes the sample variance of the residuals from an AR(4) model for the variable $r, r = 1, \ldots, n$.

Here the prior covariance matrix $\mathbf{V}_{\theta_i}$ depends on two key hyperparameters: κ_1 and κ_2. The hyperparameter κ_1 controls the overall shrinkage strength of the coefficients on their own lags, while κ_2 controls those on lags of other variables. In the application we select them by maximizing the variational lower bound of the Bayes factor on a two-dimensional grid.[4] This is motivated by papers such as Carriero, Clark, and Marcellino (2015) and Giannone, Lenza, and Primiceri (2015), which show that one can substantially improve model fit and forecast performance by selecting shrinkage hyperparameters in a data-based fashion.

3 Overview of Variational Bayes

The primary goal of Bayesian analysis is to characterize the posterior distribution of the model parameters given the data, denoted as $p(\boldsymbol{\theta} \mid \mathbf{y})$. Since this posterior distribution is

[4] Alternatively, one can also treat them as parameters to be estimated.

intractable for most econometric models, one often requires stochastic simulation methods such as MCMC to characterize the posterior distribution.

In contrast, Variational Bayes is a collection of deterministic algorithms for approximating the posterior distribution using a more tractable density. This is done by first fixing a family of tractable densities. Then, we locate the optimal density within this family by minimizing the Kullback-Leibler divergence of the approximating density to the posterior density $p(\boldsymbol{\theta} \,|\, \mathbf{y})$. Below we give a general overview of this approach. For a more detailed discussion on Variational Bayesian methods, see, e.g., Jordan, Ghahramani, Jaakkola, and Saul (1999), Chap. 10 of Bishop (2006) and Ormerod and Wand (2010). Recent applications in econometrics include Hajargasht and Woźniak (2018), Koop and Korobilis (2018), Gefang, Koop, and Poon (2019) and Loaiza-Maya, Smith, Nott, and Danaher (2020).

Let $\mathcal{Q}$ denote a family of tractable densities within which to locate the optimal approximating density. Recall that the Kullback-Leibler divergence from a density p_1 to another density p_2 is defined as

$$D_{KL}(p_1||p_2) = \int p_1(\mathbf{x}) \log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} \mathrm{d}\mathbf{x}.$$

Now, we locate the optimal approximating density $q^*(\boldsymbol{\theta})$ as the density in $\mathcal{Q}$ that minimizes the Kullback-Leibler divergence to the posterior distribution $p(\boldsymbol{\theta} \,|\, \mathbf{y})$. More precisely, $q^*(\boldsymbol{\theta})$ is the minimizer of the following minimization problem:

$$\min_{q \in \mathcal{Q}} D_{KL}(q||p(\boldsymbol{\theta} \,|\, \mathbf{y})) = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \,|\, \mathbf{y})} \mathrm{d}\boldsymbol{\theta}.$$

It turns out that minimizing the Kullback-Leibler divergence is equivalent to maximizing the q-dependent lower bound on the marginal likelihood:

$$\underline{p}(\mathbf{y}; q) \equiv \exp \int q(\boldsymbol{\theta}) \log \frac{p(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \leqslant p(\mathbf{y}).$$

Hence, this gives an alternative interpretation of the optimal density $q^*(\boldsymbol{\theta})$ as the density in $\mathcal{Q}$ that has the largest lower bound on the marginal likelihood $p(\mathbf{y})$. Moreover, the lower bound is attained, i.e., $\underline{p}(\mathbf{y}; q) = p(\mathbf{y})$ if and only if $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \,|\, \mathbf{y})$.

In general both optimization problems are hard to solve as they involve a typically high-dimensional integral. Nevertheless, one can substantially simply the computations if the parameters $\boldsymbol{\theta}$ can be naturally divided into m blocks, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$, and the approximating density q is assumed to take the form

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{m} q_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i).$$

This amounts to breaking up a high-dimensional optimization problem into m lower-dimensional ones. In particular, it can be shown that the optimal densities satisify:

$$q^*_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \propto \exp\left[\mathbb{E}_{-\boldsymbol{\theta}_i} \log p(\mathbf{y}, \boldsymbol{\theta})\right], \quad 1 \leqslant i \leqslant m, \tag{4}$$

where $\mathbb{E}_{-\boldsymbol{\theta}_i}$ denotes the expectation taken with respect to the density $\prod_{j \neq i} q_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_j)$. This leads to an iterative scheme that cycles through $i = 1, \ldots, m$ via (4), until the increase in the variational lower bound $\underline{p}(\mathbf{y}; q)$ is negligible.

For our VAR with stochastic volatility, the parameters are $(\boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i}, \mathbf{h}_i), i, \ldots, n$. We approximate the posterior density $p(\boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i}, \mathbf{h}_i \,|\, \mathbf{y}_i)$ using the approximating density of the form:

$$q(\boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i}, \mathbf{h}_i) = q_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) q_{h_{i,0}}(h_{i,0}) q_{\sigma^2_{h,i}}(\sigma^2_{h,i}) q_{\mathbf{h}_i}(\mathbf{h}_i),$$

where the marginal densities $q_{\boldsymbol{\theta}_i}, q_{h_{i,0}}$ and $q_{\sigma^2_{h,i}}$ are unrestricted, whereas $q_{\mathbf{h}_i}$ is assumed to be Gaussian. In the next section we propose a new global Gaussian approximation $q_{\mathbf{h}_i}$ for $\mathbf{h}_i$. Detailed derivations of the other densities and the variational lower bound are provided in Appendix A.

4 A New Approximating Density of the Stochastic Volatility

In this section we introduce a *global* Gaussian approximation of the joint distribution of the log-volatility $\mathbf{h}_i = (h_{i,1}, \ldots, h_{i,T})'$ in contrast to *local* approximations that have been considered in the literature. For comparison, we consider two local Gaussian approximations that have been recently used in variational Bayes inference. The first Gaussian

approximation is based on the well-known approximation of the log-χ_1^2 distribution using the $\mathcal{N}(-1.27, \pi^2/4)$ distribution. More specifically, let $y_{i,t}^* = \log(y_{i,t} - \mathbf{x}_{i,t}\boldsymbol{\theta}_i)^2$. Then, one can rewrite (3) as

$$y_{i,t}^* = h_{i,t} + \varepsilon_{i,t}^{y*},$$

where $\varepsilon_{i,t}^{y*}$ follows the log-χ_1^2 distribution. Given this transformation, Koop and Korobilis (2018) then replace the log-χ_1^2 distribution with the $\mathcal{N}(-1.27, \pi^2/4)$ distribution. Consequently, the stochastic volatility model becomes (approximately) a linear Gaussian state space model.[5] One potential problem of this approach is that the log-χ_1^2 distribution is skewed and far from Gaussian, especially at the tails. In particular, the log-χ_1^2 distribution has a heavier left tail but a much thinner right tail compared to the Gaussian approximation. As a result, the stochastic volatility estimates obtained using this approximation could be substantially distorted.

In view of this problem, Gefang, Koop, and Poon (2019) suggest another Gaussian distribution that is expected to provide a more accurate approximation. More specifically, they approximate the optimal distribution of $\mathbf{h}_i$ by a second-order Taylor approximation expanded around the mode.[6] They find that this Gaussian approximation works better in their forecasting application than the one considered in Koop and Korobilis (2018). Despite this improvement, Taylor expansion is a local approximation that is guaranteed to work well only around the neighborhood of the point of expansion (the mode of the optimal density in this case). Next, we introduce a global Gaussian approximation that takes into account the entire support of the distribution.

To set the stage, first note that the unrestricted optimal density of $\mathbf{h}_i$ — i.e., not restricted to the class of Gaussian densities — has the form:

$$\widetilde{q}_{\mathbf{h}_i}^*(\mathbf{h}_i) \propto \exp\left\{\mathbb{E}_{-\mathbf{h}_i}\left[\log p(\mathbf{h}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}_i, h_{i,0}, \sigma_{h,i}^2)\right]\right\},$$

[5] This approach of approximating the stochastic volatility model can be traced back to Harvey, Ruiz, and Shephard (1994), who suggest a quasi-maximum likelihood method to estimate the linearized model based on the Kalman filter.

[6] This type of local Gaussian approximation is first introduced to estimate stochastic volatility models in the seminal papers by Durbin and Koopman (1997) and Shephard and Pitt (1997). Specifically, they use a linear Gaussian state space model to approximate the stochastic volatility model. The distribution implied by the linear state space model is of course Gaussian, which is then used as an importance sampling density (coupled with the Kalman filter) to estimate the stochastic volatility model. Chan and Grant (2016) and Chan and Eisenstat (2018) improve upon this approach by directly computing the Gaussian approximating density using Newton-Raphson method based on fast band matrix routines.

where the expectation is taken with respect to the marginal density $q_{-\mathbf{h}_i}(\boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i}) = q_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) q_{h_{i,0}}(h_{i,0}) q_{\sigma^2_{h,i}}(\sigma^2_{h,i})$. In particular, it can be shown that the log-density of $\widetilde{q}^*_{\mathbf{h}_i}$ has the following explicit expression:

$$\begin{aligned} \log \widetilde{q}^*_{\mathbf{h}_i}(\mathbf{h}_i) = \widetilde{c}_{\mathbf{h}_i} - \frac{1}{2}\sum_{t=1}^{T} h_{i,t} - \frac{1}{2}\sum_{t=1}^{T} \mathrm{e}^{-h_{i,t}} \widehat{s}^2_t \\ - \frac{1}{2}\mathbb{E}_{\sigma^2_{h,i}}\left[\frac{1}{\sigma^2_{h,i}}\right]\left(\sum_{t=2}^{T}(h_{i,t} - h_{i,t-1})^2 + (h_{i,1} - \widehat{h}_{i,0})^2\right), \end{aligned} \tag{5}$$

where $\widetilde{c}_{\mathbf{h}_i}$ and $\widehat{s}^2_t$ are constants independent on $\mathbf{h}_i$ ($\widehat{s}^2_t$ depends on the expectation and variance with respect to the density $q_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i)$; its definition is given in Appendix A). Unfortunately, $\widetilde{q}^*_{\mathbf{h}_i}$ given in (5) is a high-dimensional non-standard density that we cannot directly use. To proceed, we approximate $\widetilde{q}^*_{\mathbf{h}_i}$ using a Gaussian distribution that is optimal in a well-defined sense.

The idea is to set up a formal optimization problem to locate the 'best' density within a parameterized class of Gaussian distributions. To that end, consider the following family of Gaussian densities:

$$\mathcal{G} = \left\{ f_{\mathcal{N}}(\cdot; \mathbf{m}, \widehat{\mathbf{K}}^{-1}_{\mathbf{h}_i}) : \mathbf{m} \in \mathbb{R}^T \right\},$$

where $f_{\mathcal{N}}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the Gaussian density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\widehat{\mathbf{K}}_{\mathbf{h}_i}$ is the negative Hessian of $\log \widetilde{q}^*_{\mathbf{h}_i}(\mathbf{h}_i)$ evaluated at the mode of $\log \widetilde{q}^*_{\mathbf{h}_i}(\mathbf{h}_i)$.[7] We then locate the member in $\mathcal{G}$ that minimizes the Kullback-Leibler divergence to $\widetilde{q}^*_{\mathbf{h}_i}$, say, $f_{\mathcal{N}}(\cdot; \widehat{\mathbf{h}}_i, \widehat{\mathbf{K}}^{-1}_{\mathbf{h}_i})$. In other words, the optimal density for $\mathbf{h}_i$ we use, denoted as $q^*_{\mathbf{h}_i}$, is then the $\mathcal{N}(\widehat{\mathbf{h}}_i, \widehat{\mathbf{K}}^{-1}_{\mathbf{h}_i})$ distribution. Since the class $\mathcal{G}$ includes the Gaussian approximation proposed in Gefang, Koop, and Poon (2019), the optimal density located under this proposed approach is guaranteed to be a better approximation — in the sense of a smaller Kullback-Leibler divergence — than the former approximation.

Now, to obtain the best Gaussian approximation within the class $\mathcal{G}$, we consider the

[7] One could expand the class of Gaussian distributions by allowing the covariance matrix to vary as well. By enlarging the class of distributions, the optimal density located is expected to be better approximation to $\widetilde{q}^*_{\mathbf{h}_i}(\mathbf{h}_i)$. On the other hand, this expanded class of distributions is much more challenging to handle due to the complex nonlinear restrictions on the covariance matrix (i.e., symmetry and positive-definiteness). We leave this possibility for future research.

optimization problem:

$$\min_{f\in\mathcal{G}} D_{KL}(f||\widetilde{q}^*_{\mathbf{h}_i}) = \min_{\mathbf{m}\in\mathbb{R}^T} \mathbb{E}\left[\log \frac{f_{\mathcal{N}}(\mathbf{h}_i;\mathbf{m},\widehat{\mathbf{K}}^{-1}_{\mathbf{h}_i})}{\widetilde{q}^*_{\mathbf{h}_i}(\mathbf{h}_i)}\right], \tag{6}$$

where the expectation is taken with respect to the density $f_{\mathcal{N}}(\mathbf{h}_i;\mathbf{m},\widehat{\mathbf{K}}^{-1}_{\mathbf{h}_i})$. It turns out that the problem in (6) is a convex optimization problem with a unique minimizer. In addition, we are able to derive analytical expressions for the gradient and Hessian of the objective function, and therefore the minimization problem can be quickly solved using the Newton-Raphson method. Furthermore, the Hessian is banded, which further speeds up computations by implementing fast band matrix routines. We provide the computational details in Appendix A. Finally, we use the minimizer as the optimal density $q^*_{\mathbf{h}_i}$.

5 Monte Carlo Experiments

In this section we conduct a Monte Carlo study to assess the accuracy of approximating the posterior distribution of $\mathbf{h}$ using the proposed global Gaussian approximation. We also document the runtimes of the Variational Bayesian methods compared to MCMC for estimating VARs of different dimensions.

First, we assess the accuracy of proposed variational approximation. As a comparison, we include two local Gaussian approximations: 1) the Gaussian distribution obtained by replacing the log-χ^2_1 distribution in the transformed observation equation with the $\mathcal{N}(-1.27,\pi^2/4)$ distribution; 2) a second-order Taylor approximation expanded around the mode of the target posterior distribution of $\mathbf{h}$.

More specifically, we generate R datasets from the following univariate stochastic volatility model:

$$\begin{aligned} z_t &= \mathrm{e}^{\frac{1}{2}h_t}u_t, & u_t &\sim \mathcal{N}(0,1),\\ h_t &= h_{t-1}+v_t, & v_t &\sim \mathcal{N}(0,0.1) \end{aligned}$$

for $t=1,\ldots,T$, and we set $h_0=0$. For each dataset $\mathbf{z}^{(i)}=(z_1^{(i)},\ldots,z_T^{(i)})', i=1,\ldots,R$

and each Gaussian approximation with mean vector $\widehat{\mathbf{h}}^j = (\widehat{h}_1^j, \ldots, \widehat{h}_T^j)', j = 1, \ldots, 3$, we compute the mean squared error (MSE), $\text{MSE}_i(\widehat{\mathbf{h}}^j) = \sum_{t=1}^T (\widehat{h}_t^j - \bar{h}_t)^2/T$, where $\bar{h}_1, \ldots, \bar{h}_T$ are the posterior means obtained via MCMC.

We first set the sample size to be $T = 300$ and the number of Monte Carlo replications to be $R = 500$. The left panel in Figure 1 presents boxplots of the MSEs for the three Gaussian approximations. Due to the differences in scale, the right panel excludes the first approximation for better clarity. As it is clear from the left panel, the first local approximation based on the $\mathcal{N}(-1.27, \pi^2/4)$ approximation of the log-χ_1^2 distribution (approx1) is typically an order of magnitude worse than the other two Gaussian approximations. This approximation also performs poorly in absolute terms for a substantial number of datasets.

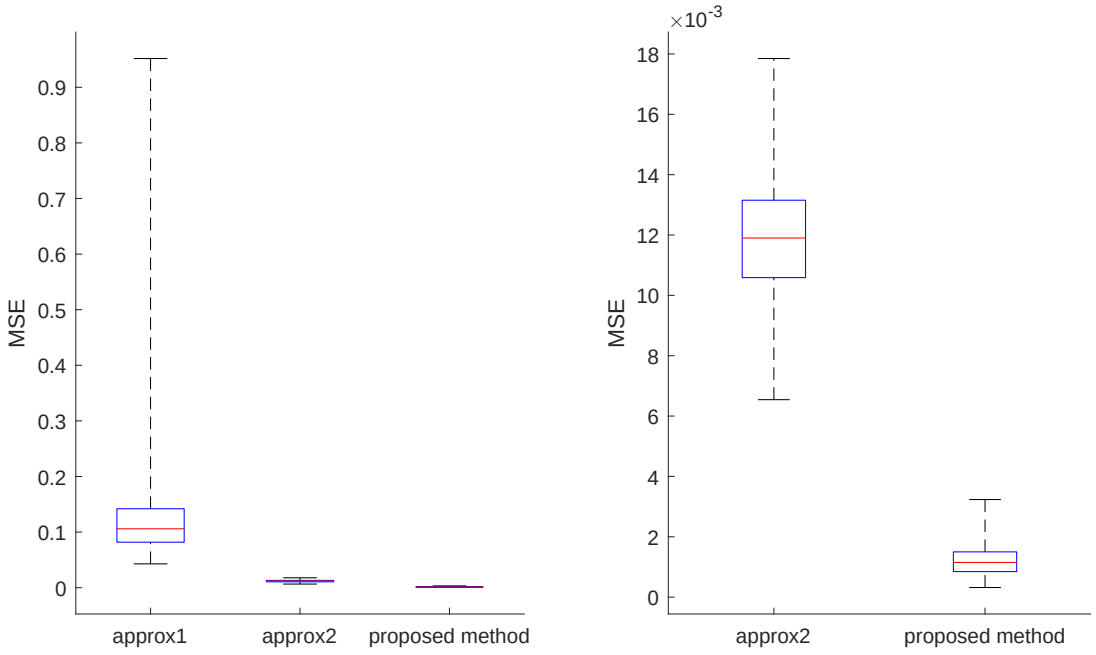

Figure 1: Boxplots of the mean squared errors for the three Gaussian approximations for $T = 300$. Approx1 is the Gaussian distribution obtained by the $\mathcal{N}(-1.27, \pi^2/4)$ approximation of the log-χ_1^2 distribution and approx2 is the Gaussian distribution based on a second-order Taylor approximation expanded around the mode of the target posterior distribution. The central mark indicates the median, whereas the bottom and top edges of the box indicate the 25-th and 75-th percentiles, respectively. The whiskers extend to the minimum and the maximum.

Next, the right panel shows that the proposed global approximation is substantially better

than the second local approximation based on a second-order Taylor expansion (approx2) — it provides another order of magnitude reduction in MSE. For example, the median MSE of the proposed approximation is only about 0.001, compared to about 0.012 for the second local approximation. Moreover, for all datasets the proposed method provides a better approximation — in terms of the lowest MSE — compared to the two alternatives.

We repeat the exercise with a longer sample of $T = 2,000$ and the results are reported in Figure 2. The main conclusion remains the same: the proposed global approximation method is able to obtain a much more accurate approximation of the log-volatility than existing local approximation methods.

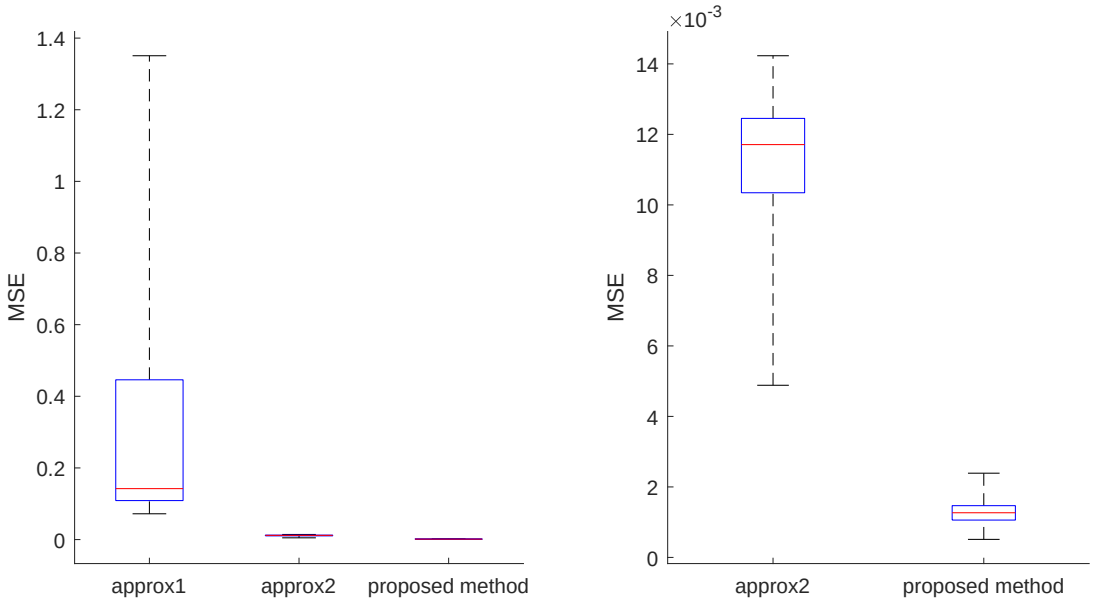

Figure 2: Boxplots of the mean squared errors for the three Gaussian approximations for $T = 2,000$. Approx1 is the Gaussian distribution obtained by the $\mathcal{N}(-1.27, \pi^2/4)$ approximation of the log-χ^2_1 distribution and approx2 is the Gaussian distribution based on a second-order Taylor approximation expanded around the mode of the target posterior distribution. The central mark indicates the median, whereas the bottom and top edges of the box indicate the 25-th and 75-th percentiles, respectively. The whiskers extend to the minimum and the maximum.

Next, we document the runtimes of estimating VARs of of different dimensions using the proposed Variational Bayesian methods versus MCMC. More specifically, Table 1 reports the computation times (in minutes) to fit VARs of dimensions $n = 25, 50, 100$ and sample sizes $T = 300, 2,000$ using the two approaches. The algorithms are implemented using

MATLAB on a desktop with an Intel Core i5-9600 @3.10 GHz processor and 16GB memory (when we implement the Variational Bayesian methods, we do not use parallel computing for a fair comparison with MCMC). As it is evident from the table, the proposed approach is much faster than conventional MCMC. For example, fitting a 100-variable VAR with $T = 300$ using the proposed approach takes only about 7 minutes, as opposed to about 70 minutes when MCMC is used.

Table 1: The computation times (in minutes) to fit an n-variable VAR with a sample size T using Variational Bayesian methods (VB) and MCMC (to obtain 10,000 posterior draws). All VARs have $p = 4$ lags.

	$T = 300$			$T = 2,000$		
	$n = 25$	$n = 50$	$n = 100$	$n = 25$	$n = 50$	$n = 100$
MCMC	2.16	7.30	69.4	15.2	61.1	257.5
VB	0.06	0.22	6.99	0.21	0.93	6.78

6 Application: Bank Network Connectedness

In this section we illustrate the proposed methodology by using a large VAR with stochastic volatility to measure global bank network connectedness with the connectedness measures developed in the series of papers by Diebold and Yilmaz (2009, 2014) and Demirer, Diebold, Liu, and Yilmaz (2018, DDLY henceforth). In particular, we revisit the application in DDLY, who consider a 96-variable homoscedastic VAR regularized by the adaptive elastic net (Zou and Zhang, 2009) to measure global bank network connectedness. Since the connectedness measures are functions of the error covariance matrix, how it is modeled is likely to be crucial. We therefore extend the analysis in DDLY by allowing the error covariance matrix to vary over time via a stochastic volatility process.[8]

Another interesting aspect of the analysis is the impact of shrinkage methods on the connectedness measures. While different shrinkage methods are expected to have varying effects on the VAR coefficient estimates, their role on the connectedness measures is less

[8]Korobilis and Yilmaz (2018) also use a time-varying VAR to measure bank network connectedness. However, they adopt the approach in Koop and Korobilis (2013), which gives only filtered estimates (estimates at time t are based only on data up to time t) but not smoothed estimates (estimates obtained using the whole sample). Moreover, in their application they use a 35-variable VAR. In contrast, using our fast Variational Bayes approach, it is feasible to estimates a 96-variable or even larger systems.

obvious. To regularize the large VAR, DDLY use an adaptive elastic net — an average of Lasso and ridge penalties, where the weights are inverses of the least squares estimates. The value of the overall penalty weight is selected by 10-fold cross validation. In contrast, our shrinkage method can be viewed as a combination of an adaptive and a subjective ridge — the weights on individual VAR coefficients are subjectively elicited according to the Minnesota prior, but the overall shrinkage parameter is obtained by maximizing the marginal likelihood of the approximate model (variational lower bound). It turns out that these two very different shrinkage methods give quite similar results, provided that the error covariance matrix of the VAR is kept to be time-invariant.

In the analysis we use data of daily stock prices of 96 banks from 29 developed and emerging economies, and the sample period is from September 12, 2003 to February 7, 2014. These 96 banks are those in the world's top 150 by assets that were publicly traded throughout the sample.[9] To measure the connectedness in the global bank stock return volatility network, raw daily stock prices (high, low, opening and closing prices) are used to compute daily range-based realized volatility as proposed in Garman and Klass (1980). This daily bank stock return volatility measure is then used as the dependent variable. We refer the reader to DDLY for more details on the data.

In what follows, we first define the connectedness measures as functions of the VAR parameters. Then, we present results based on a homoscedastic VAR with an adaptive Minnesota prior, followed by results obtained from a VAR with stochastic volatility.

6.1 Connectedness Measures

In this section we define the connectedness measures that we use to characterize bank network connectedness. These measures are based on variance decompositions and are specifically designed to quantify how much individual bank's future uncertainty can be attributed to another specific bank or all other banks as a whole.[10] They can be computed from the estimates of the VAR given in (1)-(2).

[9]We thank Laura Liu for providing us with the data and the associated R code. The dataset can also be downloaded from the *Journal of Applied Econometrics Data Archive*.

[10]The connectedness measures developed in Diebold and Yilmaz (2009, 2014) and Demirer, Diebold, Liu, and Yilmaz (2018) are based on a homoscedastic VAR. Here we extend these measures to a VAR with stochastic volatility.

More specifically, given the estimates (e.g., posterior means) of the structural-form VAR in (1), we can recover the reduced-form estimates by computing $\widetilde{\mathbf{b}} = \mathbf{B}_0^{-1}\mathbf{b}, \widetilde{\mathbf{B}}_j = \mathbf{B}_0^{-1}\mathbf{B}_j, j = 1, \ldots, p$ and $\widetilde{\mathbf{\Sigma}}_t = \mathbf{B}_0^{-1}\mathbf{\Sigma}_t(\mathbf{B}_0^{-1})'$. Then, using these reduced-form estimates, we construct the corresponding vector moving average matrices $\mathbf{A}_h, h = 0, 1, 2, \ldots$, with the convention that $\mathbf{A}_0 = \mathbf{I}_n$.

Next, following Diebold and Yilmaz (2014) we define the most granular directional connectedness from one bank to another. More specifically, bank j's contribution to bank i's H-step-ahead generalized forecast error variance is defined as

$$\theta^g_{ij,t}(H) = \frac{\widetilde{\sigma}^{-1}_{jj,t}\sum_{h=0}^{H-1}(\mathbf{e}_i'\mathbf{A}_h\widetilde{\mathbf{\Sigma}}_t\mathbf{e}_j)^2}{\sum_{h=0}^{H-1}(\mathbf{e}_i'\mathbf{A}_h\widetilde{\mathbf{\Sigma}}_t\mathbf{A}_h'\mathbf{e}_i)^2},$$

where $\widetilde{\sigma}_{jj,t}$ is the j-th diagonal elements of $\widetilde{\mathbf{\Sigma}}_t$ and $\mathbf{e}_i$ is the selection vector with one on the i-th position and and zeros otherwise. In contrast to the static measure in DDLY, here the connectedness measure is time-varying.

Since these generalized forecast error variances might not sum to one, we normalize them as follows:

$$C^H_{i\leftarrow j,t} = \frac{\theta^g_{ij,t}(H)}{\sum_{k=1}^n \theta^g_{ik,t}(H)}. \tag{7}$$

Next, we aggregate these pairwise directional connectedness measures to form total directional connectedness measures. The total directional connectedness to bank i from all other banks is:

$$C^H_{i\leftarrow \bullet,t} = \frac{1}{n}\sum_{j=1, j\neq i}^{n} C^H_{i\leftarrow j,t}. \tag{8}$$

The total directional connectedness from bank i to all other banks is similarly defined as

$$C^H_{\bullet\leftarrow i,t} = \frac{1}{n}\sum_{j=1, j\neq i}^{n} C^H_{j\leftarrow i,t}. \tag{9}$$

Finally, we can measure the total directional connectedness as

$$C^H_t = \frac{1}{n}\sum_{i,j=1, i\neq j}^{n} C^H_{i\leftarrow j,t}. \tag{10}$$

This measure is referred to as system-wide connectedness, as it aggregates the total

directional connectedness, both 'to' and 'from'.

6.2 Results from a Large Homoscedastic VAR

We first report a range of bank connectedness measures using a large homoscedastic VAR with the Minnesota prior described in Section 2. In our implementation, the two shrinkage priors κ_1 and κ_2 are selected by maximizing the variational lower bound over a 2-dimensional grid. The optimal values obtained are $\kappa_1 = 0.04$ and $\kappa_2 = 0.001$, indicating much stronger shrinkage toward zero for coefficients on 'other' lags than those on 'own' lags.

Following DDLY, instead of reporting individual bank connectedness measures, we aggregate the network connectedness measures into six regions: Africa, Asia, Europe, North America, Oceania and South America. The results are reported in Table 2. These are the connectedness measures defined in (7) with $H = 10$, where each unit is a region rather than an individual bank.[11] The row sums labeled 'from others' are the total directional connectedness from others defined in (8), and the column sums labeled 'to others' are the total directional connectedness to others defined in equation (9). Lastly, the lower right element is the system-wide connectedness defined in (10).

Table 2: Bank network connectedness for the six-group aggregation, 2003-2014, from a homoscedastic VAR.

	Africa	Asia	Europe	N. America	Oceania	S. America	From others
Africa	0.00	7.54	22.45	22.11	2.12	2.39	56.62
Asia	3.79	0.00	217.88	283.35	30.51	21.14	556.68
Europe	4.93	67.35	0.00	734.41	32.60	33.35	872.64
N. America	2.99	52.16	583.26	0.00	27.88	26.43	692.73
Oceania	1.72	26.28	116.65	134.80	0.00	6.51	285.97
S. America	1.22	12.07	50.02	54.00	2.83	0.00	120.14
To others	14.66	165.40	990.26	1228.67	95.94	89.83	2584.77

Overall our results are qualitatively similar to those reported in DDLY, despite their use of a different penalty (an adaptive elastic net). In particular, our results also suggest

[11] Note that here we use a homoscedastic VAR, and consequently these connectedness measures are time-invariant, i.e., $C^H_{i\leftarrow j,1} = \cdots = C^H_{i\leftarrow j,T}$.

that North America and Europe are the two largest net transmitters of future volatility uncertainty (the difference between 'to others' and 'from others') to the rest of the world. Moreover, Asia has substantial total directional connectedness in both directions (i.e., total directional connectedness 'to others' and 'from others'), and it is a net receiver of future volatility uncertainty.

6.3 Results from a Large VAR with SV

Next, we report bank connectedness measures using a large VAR with stochastic volatility. Since in our VAR the error covariance matrix is time-varying, the connectedness measures defined in (7)–(10) are also time-varying. In contrast, DDLY use rolling estimation with a 150-day window to characterize the global banking network dynamically. In this section we compare the connectedness measures under our VAR with stochastic volatility with their rolling-window results.

We use the Minnesota prior described in Section 2, where the two shrinkage priors κ_1 and κ_2 are selected by maximizing the variational lower bound over a 2-dimensional grid. The optimal hyperparameter values obtained are $\kappa_1 = 0.04$ and $\kappa_2 = 0.001$, again showing much stronger shrinkage for coefficients on 'other' lags than on 'own' lags. To see how well the VAR with stochastic volatility fits the data compared to a standard homoscedastic VAR, we first obtain the variational lower bound and the variational Bayesian Information Criterion (BIC) (see, e.g., You, Ormerod, and Mueller, 2014) for both models, and the results are reported in Table 3.

Table 3: Variational lower bound and variational BIC for the standard homoscedastic VAR (VAR) and the VAR with stochastic volatility (VAR-SV). For variational lower bound, a larger value indicates a better model; for variational BIC, a smaller value indicates a better model.

	VAR	VAR-SV
Variational lower bound	−286,050	−283,490
Variational BIC	536,690	523,530

Both criteria suggest that the VAR with stochastic volatility is strongly preferred by the data relative to its homoscedastic counterpart. Since both criteria have a built-in penalty

for model complexity, the results indicate that the increase in model-fit by allowing for time-varying variances outweighs the cost of additional model complexity.

We first report the bank network connectedness for the six-group aggregation in Table 4. Since the network connectedness measures are time-varying under the VAR with stochastic volatility, the values in the table are averages over the whole sample 2003–2014.

The values of these connectedness measures are very similar to those under the homoscedastic VAR in Table 2, even though now the model allows for stochastic volatility. Consequently, the main message — that North America and Europe are the two largest net transmitters of future volatility uncertainty and Asia is a large net receiver of future volatility uncertainty from the rest of the world — remains the same.

Table 4: Bank network connectedness for the six-group aggregation, 2003-2014, from the VAR-SV.

	Africa	Asia	Europe	N. America	Oceania	S. America	From others
Africa	0.00	7.56	24.79	21.81	2.18	2.33	58.67
Asia	3.57	0.00	236.56	284.11	32.30	24.49	581.03
Europe	5.16	64.37	0.00	678.02	29.32	33.26	810.13
N. America	3.16	54.03	618.20	0.00	27.21	27.23	729.83
Oceania	1.78	25.00	126.38	129.62	0.00	6.58	289.36
S. America	1.12	13.35	51.39	52.32	2.94	0.00	121.12
To others	14.80	164.30	1057.32	1165.88	93.95	93.88	2590.13

Next, we plot the dynamic system-wide connectedness in Figure 3. Our results show an overall similar pattern compared to DDLY's estimates from a homoscedastic VAR obtained via a 150-day rolling window. In particular, our dynamic estimates of the system-wide connectedness tend to increase from the beginning of the sample to around 2008. The main difference is that our measure peaks earlier — the first peak coincides with the start of the liquidity crisis in August 2007, when the French bank BNP Paribas froze three investment funds because of losses related to US subprime securities. This caused the bond market to seize up, prompting the US Federal Reserve and the European Central Bank to inject liquidity into the money markets to keep interest rates down. Moreover, after this initial shock, volatility connectedness remains high through the two waves of European Debt Crisis in May 2010 and July-August 2011.

To verify the timing of the first peak, we follow DDLY to compute the dynamic system-

wide connectedness measure using a homoscedastic VAR with a 150-day rolling window. The results are reported in Appendix B. Despite the very different shrinkage methods employed, our dynamic measure is remarkably similar to that in DDLY. In particular, the dynamic system-wide measure obtained using the rolling window increases substantially in 2007, but it does not peak till around the collapse of Lehman Brothers in September 2008. We thus conclude that the differences in the timing of the peak can be attributed to the use of the stochastic volatility specification. As Korobilis and Yilmaz (2018) argue, rolling window estimates from a homoscedastic VAR have a built-in persistence — the model itself is time-invariant, and changes only occur slowly as observations change gradually in the fixed-length rolling-sample window. Consequently, estimates obtained tend to be slow in detecting sudden breaks or drastic changes.

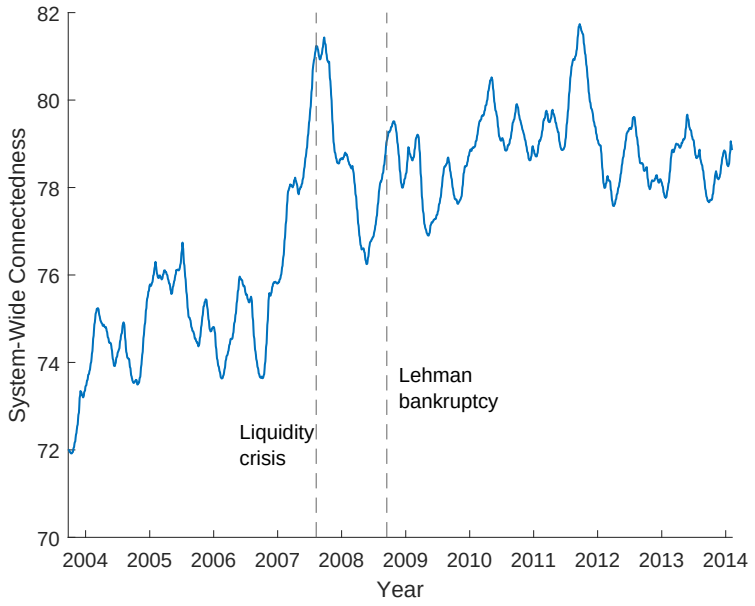

Figure 3: The dynamic system-wide connectedness measure from the VAR-SV.

Next, we decompose the dynamic system-wide connectedness into cross-country and within-country components. More specifically, cross-country system-wide connectedness is calculated as the sum of all pairwise connectedness across banks located in different countries. Similarly, within-country system-wide connectedness is the sum of pairwise connectedness across banks in the same country. This decomposition allows us to explore the country origins of volatility shocks and helps us better understand the dynamics of

global bank connectedness. The results of the decomposition are depicted in Figure 4.

Similar to the results reported in DDLY, we find that most variations in system-wide connectedness are due to variations in cross-country system-wide connectedness, whereas the within-country connectedness remains relatively stable throughout the sample period. Our results show that cross-country system-wide connectedness remains stable at around 50% from the beginning of the sample till early 2017, and it then begins to fluctuate significantly. Since the within-country connectedness remains relatively stable, cross-country system-wide connectedness shows similar patterns as the total system-wide connectedness depicted in Figure 3. In particular, cross-country system-wide connectedness increases substantially around the liquidity crisis of August 2007, and it remains high through the two waves of European Debt Crisis.

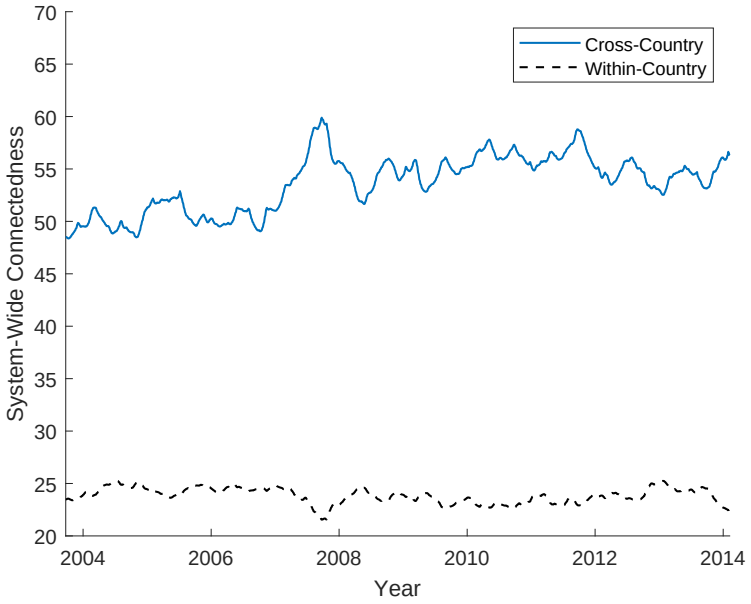

Figure 4: The decomposition of the dynamic system-wide connectedness measure from the VAR-SV into cross-country and within-country components.

7 Concluding Remarks and Future Research

We have developed a new variational approximation of the joint posterior distribution of the log-volatility in the context of large VARs. In contrast to existing approaches that

are based on local approximations around a point in the support of the distribution, the new method provides a global approximation that takes into account the entire support. We have provided evidence of superior approximation accuracy of the new proposal in a Monte Carlo study compared to existing approximations.

Econometricians have only recently begun to use Variational Bayesian methods as alternatives to MCMC for fitting high-dimensional models. In future research, it would be useful to explore fitting other high-dimensional models, such as large VARs with a factor stochastic volatility structure, using these methods. In addition, since the variational lower bound can be obtained quite quickly, it would also be interesting to explore using it to compare large stochastic volatility models or shrinkage priors.

Appendix A: Estimation Details

In this appendix we provide the technical details of the variational Bayes approximation of the posterior distribution. Recall that the model can be written as n unrelated regressions for $i = 1, \ldots, n$:

$$y_{i,t} = \mathbf{x}_{i,t}\boldsymbol{\theta}_i + \varepsilon^y_{i,t}, \quad \varepsilon^y_{i,t} \sim \mathcal{N}(0, \mathrm{e}^{h_{i,t}}),$$

where the log-volatility $h_{i,t}$ evolves as a random walk:

$$h_{i,t} = h_{i,t-1} + \varepsilon^h_{i,t}, \quad \varepsilon^h_{i,t} \sim \mathcal{N}(0, \sigma^2_{h,i}).$$

In addition, the priors on the parameters $\boldsymbol{\theta}_i, h_{i,0}$ and $\sigma^2_{h,i}$ are also independent across equations. Specifically, we assume for $i = 1, \ldots, n$:

$$\boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\theta}_{0,i}, \mathbf{V}_{\boldsymbol{\theta}_i}), \quad h_{i,0} \sim \mathcal{N}(0, V_{h_{i,0}}), \quad \sigma^2_{h,i} \sim \mathcal{IG}(\nu_i, S_i).$$

Therefore, the joint distribution of the parameters and log-volatilities are independent across equations. More specifically, the joint posterior distribution of $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_n)'$, $\mathbf{h}_0 = (h_{1,0}, \ldots, h_{n,0})'$, $\boldsymbol{\sigma}^2_h = (\sigma^2_{h,1}, \ldots, \sigma^2_{h,n})'$ and $\mathbf{h} = (\mathbf{h}'_1, \ldots, \mathbf{h}'_T)'$ can be decomposed as:

$$p(\boldsymbol{\theta}, \mathbf{h}_0, \boldsymbol{\sigma}^2_h, \mathbf{h} \,|\, \mathbf{y}) = \prod_{i=1}^{n} p(\boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i}, \mathbf{h}_i \,|\, \mathbf{y}_i),$$

where $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,T})'$. Hence, it suffices to obtain a variational approximation of each of the n components $p(\boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i}, \mathbf{h}_i \,|\, \mathbf{y}_i), i = 1, \ldots, n$.

Now, we approximate $p(\boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i}, \mathbf{h}_i \,|\, \mathbf{y}_i)$ using the product of densities:

$$q(\boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i}, \mathbf{h}_i) = q_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) q_{h_{i,0}}(h_{i,0}) q_{\sigma^2_{h,i}}(\sigma^2_{h,i}) q_{\mathbf{h}_i}(\mathbf{h}_i),$$

where the marginal densities $q_{\boldsymbol{\theta}_i}, q_{h_{i,0}}$ and $q_{\sigma^2_{h,i}}$ are unrestricted, whereas $q_{\mathbf{h}_i}$ is assumed to be Gaussian. Let $q^* = q^*_{\boldsymbol{\theta}_i} q^*_{h_{i,0}} q^*_{\sigma^2_{h,i}} q^*_{\mathbf{h}_i}$ denote the optimal density. In what follows, we derive the explicit forms of each of these optimal marginal densities and their associated parameters.

The Optimal Density $q^*_{\boldsymbol{\theta}_i}$

The optimal density $q^*_{\boldsymbol{\theta}_i}$ has the form

$$q^*_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \propto \exp\left\{\mathbb{E}_{-\boldsymbol{\theta}_i}\left[\log p(\boldsymbol{\theta}_i \,|\, \mathbf{y}_i, \mathbf{h}_i, h_{i,0}, \sigma^2_{h,i})\right]\right\},$$

where the expectation is taken with respect to the marginal density $q_{-\boldsymbol{\theta}_i}(h_{i,0}, \sigma^2_{h,i}, \mathbf{h}_i) = q_{h_{i,0}}(h_{i,0}) q_{\sigma^2_{h,i}}(\sigma^2_{h,i}) q_{\mathbf{h}_i}(\mathbf{h}_i)$. To derive an explicit expression of $q^*_{\boldsymbol{\theta}_i}$, first note that $\boldsymbol{\theta}_i$ is conditionally independent of $(h_{i,0}, \sigma^2_{h,i})$ given $(\mathbf{h}_i, \mathbf{y}_i)$. In particular, the log-density is given by

$$\log p(\boldsymbol{\theta}_i \,|\, \mathbf{y}_i, \mathbf{h}_i) = c_{\boldsymbol{\theta}_i} - \frac{1}{2}\sum_{t=1}^{T} \mathrm{e}^{-h_{i,t}}(y_{i,t} - \mathbf{x}_{i,t}\boldsymbol{\theta}_i)^2 - \frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0,i})'\mathbf{V}^{-1}_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0,i}),$$

where $c_{\boldsymbol{\theta}_i}$ is a constant not dependent on $\boldsymbol{\theta}_i$. Let $\widehat{\mathbf{h}}_i = (\widehat{h}_{i,1}, \ldots, \widehat{h}_{i,T})'$ and $\widehat{\mathbf{K}}_{\mathbf{h}_i}$ denote respectively the mean vector and precision matrix (i.e., inverse covariance matrix) of $\mathbf{h}_i$ with respect to the density $q_{\mathbf{h}_i}(\mathbf{h}_i)$. Then, taking expectation of $\log p(\boldsymbol{\theta}_i \,|\, \mathbf{y}_i, \mathbf{h}_i)$ gives

$$\mathbb{E}_{-\boldsymbol{\theta}_i}\left[\log p(\boldsymbol{\theta}_i \,|\, \mathbf{y}_i, \mathbf{h}_i)\right] = c_{\boldsymbol{\theta}_i} - \frac{1}{2}\sum_{t=1}^{T} \mathrm{e}^{-\widehat{h}_{i,t} + \frac{1}{2}\widehat{d}_{i,t}}(y_{i,t} - \mathbf{x}_{i,t}\boldsymbol{\theta}_i)^2 - \frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0,i})'\mathbf{V}^{-1}_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0,i}),$$

where $\widehat{d}_{i,t}$ is the t-th diagonal element of $\widehat{\mathbf{K}}^{-1}_{\mathbf{h}_i}$. Since $\mathbb{E}_{-\boldsymbol{\theta}_i}\left[\log p(\boldsymbol{\theta}_i \,|\, \mathbf{y}_i, \mathbf{h}_i)\right]$ is a (negative) quadratic form in $\boldsymbol{\theta}_i$, the optimal density $q^*_{\boldsymbol{\theta}_i}$ is Gaussian. To derive the corresponding mean vector and precision matrix, we rewrite $\mathbb{E}_{-\boldsymbol{\theta}_i}\left[\log p(\boldsymbol{\theta}_i \,|\, \mathbf{y}_i, \mathbf{h}_i)\right]$ in matrix form as

$$\mathbb{E}_{-\boldsymbol{\theta}_i}\left[\log p(\boldsymbol{\theta}_i \,|\, \mathbf{y}_i, \mathbf{h}_i)\right] = c_{\boldsymbol{\theta}_i} - \frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}_i)'\widehat{\mathbf{O}}_{\mathbf{h}_i}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}_i) - \frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0,i})'\mathbf{V}^{-1}_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0,i}),$$

where $\widehat{\mathbf{O}}_{\mathbf{h}_i} = \mathrm{diag}(\mathrm{e}^{-\widehat{h}_{i,1} + \frac{1}{2}\widehat{d}_{i,1}}, \ldots, \mathrm{e}^{-\widehat{h}_{i,T} + \frac{1}{2}\widehat{d}_{i,T}})$ and $\mathbf{X}_i = (\mathbf{x}'_{i,1}, \ldots, \mathbf{x}'_{i,T})'$. Now, using standard linear regression results (see, .e.g, Chan, Koop, Poirier, and Tobias, 2019, pp. 182-188), one can show that it is the $\mathcal{N}(\widehat{\boldsymbol{\theta}}_i, \widehat{\mathbf{K}}^{-1}_{\boldsymbol{\theta}_i})$ distribution, where

$$\widehat{\mathbf{K}}_{\boldsymbol{\theta}_i} = \mathbf{V}^{-1}_{\boldsymbol{\theta}_i} + \mathbf{X}'_i\widehat{\mathbf{O}}_{\mathbf{h}_i}\mathbf{X}_i, \quad \widehat{\boldsymbol{\theta}}_i = \widehat{\mathbf{K}}^{-1}_{\boldsymbol{\theta}_i}(\mathbf{V}^{-1}_{\boldsymbol{\theta}_i}\boldsymbol{\theta}_{0,i} + \mathbf{X}'_i\widehat{\mathbf{O}}_{\mathbf{h}_i}\mathbf{y}_i).$$

The Optimal Density $q^*_{h_{i,0}}$

Next, we derive the optimal density $q^*_{h_{i,0}}$, which takes the form

$$q^*_{h_{i,0}}(h_{i,0}) \propto \exp\left\{\mathbb{E}_{-h_{i,0}}\left[\log p(h_{i,0} \,|\, \mathbf{y}_i, \boldsymbol{\theta}_i, \mathbf{h}_i, \sigma^2_{h,i})\right]\right\},$$

where the expectation is taken with respect to the marginal density $q_{-h_{i,0}}(\boldsymbol{\theta}_i, \sigma^2_{h,i}, \mathbf{h}_i) = q_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) q_{\sigma^2_{h,i}}(\sigma^2_{h,i}) q_{\mathbf{h}_i}(\mathbf{h}_i)$. First write

$$\log p(h_{i,0} \,|\, \mathbf{y}_i, \boldsymbol{\theta}_i, \mathbf{h}_i, \sigma^2_{h,i}) = \log p(h_{i,0} \,|\, \mathbf{h}_i, \sigma^2_{h,i}) = c_{h_{i,0}} - \frac{1}{2\sigma^2_{h,i}}(h_{i,1} - h_{i,0})^2 - \frac{1}{2V_{h_{i,0}}} h^2_{i,0},$$

where $c_{h_{i,0}}$ is a constant not dependent on $h_{i,0}$. Then, taking expectation with respect to the marginal density $q_{-h_{i,0}}$, we obtain

$$\mathbb{E}_{-h_{i,0}}\left[\log p(h_{i,0} \,|\, \mathbf{h}_i, \sigma^2_{h,i})\right] = c_{h_{i,0}} - \frac{1}{2}\mathbb{E}_{\sigma^2_{h,i}}\left[\frac{1}{\sigma^2_{h,i}}\right]\left[(\widehat{h}_{i,1} - h_{i,0})^2 + \widehat{d}_{i,1}\right] - \frac{1}{2V_{h_{i,0}}} h^2_{i,0},$$

where $\widehat{d}_{i,1}$ is the first diagonal element of $\widehat{\mathbf{K}}^{-1}_{\mathbf{h}_i}$ and the expectation $\mathbb{E}_{\sigma^2_{h,i}}$ is taken with respect to the density $q_{\sigma^2_{h,i}}(\sigma^2_{h,i})$ — this expectation can be computed analytically as shown in the next subsection. Finally, using standard linear regression results, one can show that $q^*_{h_{i,0}}$ is the $\mathcal{N}(\widehat{h}_{i,0}, \widehat{K}^{-1}_{h_{i,0}})$ distribution, where

$$\widehat{K}_{h_{i,0}} = V^{-1}_{h_{i,0}} + \mathbb{E}_{\sigma^2_{h,i}}\left[\frac{1}{\sigma^2_{h,i}}\right], \quad \widehat{h}_{i,0} = \widehat{K}^{-1}_{h_{i,0}} \mathbb{E}_{\sigma^2_{h,i}}\left[\frac{1}{\sigma^2_{h,i}}\right] \widehat{h}_{i,1}.$$

The Optimal Density $q^*_{\sigma^2_{h,i}}$

The kernel of the optimal density $q^*_{\sigma^2_{h,i}}$ is given by

$$q^*_{\sigma^2_{h,i}} \propto \exp\left\{\mathbb{E}_{-\sigma^2_{h,i}}\left[\log p(\sigma^2_{h,i} \,|\, \mathbf{h}_i, h_{i,0})\right]\right\},$$

where the expectation is taken with respect to the marginal density $q_{-\sigma^2_{h,i}}(\boldsymbol{\theta}_i, h_{i,0}, \mathbf{h}_i) = q_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) q_{h_{i,0}}(h_{i,0}) q_{\mathbf{h}_i}(\mathbf{h}_i)$. To derive an explicit expression for $q^*_{\sigma^2_{h,i}}$, first note that

$$\log p(\sigma^2_{h,i} \,|\, \mathbf{h}_i, h_{i,0}) = c_{\sigma^2_{h,i}} - \frac{T}{2} \log \sigma^2_{h,i} - \frac{1}{2\sigma^2_{h,i}} (\mathbf{h}_i - h_{i,0}\mathbf{1}_T)' \mathbf{H}' \mathbf{H} (\mathbf{h}_i - h_{i,0}\mathbf{1}_T) - \nu_i \log \sigma^2_{h,i} - \frac{S_i}{\sigma^2_{h,i}},$$

where $c_{\sigma^2_{h,i}}$ is a constant not dependent on $\sigma^2_{h,i}$. Taking expectation with respect to the marginal density $q_{-\sigma^2_{h,i}}$ gives

$$\begin{aligned}\mathbb{E}_{-\sigma^2_{h,i}} \left[\log p(\sigma^2_{h,i} \,|\, \mathbf{h}_i, h_{i,0}) \right] = c_{\sigma^2_{h,i}} &- \left(\nu_i + \frac{T}{2} \right) \log \sigma^2_{h,i} - \frac{S_i}{\sigma^2_{h,i}} \\ &- \frac{1}{2\sigma^2_{h,i}} \left[(\widehat{\mathbf{h}}_i - \widehat{h}_{i,0}\mathbf{1}_T)' \mathbf{H}' \mathbf{H} (\widehat{\mathbf{h}}_i - \widehat{h}_{i,0}\mathbf{1}_T) + \mathrm{tr}(\mathbf{H}' \mathbf{H} \widehat{\mathbf{K}}^{-1}_{\mathbf{h}_i}) + \widehat{K}^{-1}_{h_{i,0}} \right].\end{aligned}$$

Exponentiating the term on the right-hand side, one recognizes that it is the kernel of the $\mathcal{IG}(\widehat{\nu}_i, \widehat{S}_i)$ distribution, where

$$\widehat{\nu}_i = \nu_i + \frac{T}{2}, \quad \widehat{S}_i = S_i + \frac{1}{2} \left[(\widehat{\mathbf{h}}_i - \widehat{h}_{i,0}\mathbf{1}_T)' \mathbf{H}' \mathbf{H} (\widehat{\mathbf{h}}_i - \widehat{h}_{i,0}\mathbf{1}_T) + \mathrm{tr}(\mathbf{H}' \mathbf{H} \widehat{\mathbf{K}}^{-1}_{\mathbf{h}_i}) + \widehat{K}^{-1}_{h_{i,0}} \right].$$

Since $q^*_{\sigma^2_{h,i}}$ is an inverse-gamma density, the expectation of $1/\sigma^2_{h,i}$ can be obtained analytically as:

$$\mathbb{E}_{\sigma^2_{h,i}} \left[\frac{1}{\sigma^2_{h,i}} \right] = \frac{\widehat{\nu}_i}{\widehat{S}_i}.$$

The Optimal Density $q^*_{\mathbf{h}_i}$

The unrestricted optimal density of $\mathbf{h}_i$ — i.e., not restricted to the class of Gaussian densities — has the form:

$$\widetilde{q}^*_{\mathbf{h}_i}(\mathbf{h}_i) \propto \exp \left\{ \mathbb{E}_{-\mathbf{h}_i} \left[\log p(\mathbf{h}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i}) \right] \right\}.$$

In what follows, we first derive an explicit expression for $\widetilde{q}^*_{\mathbf{h}_i}(\mathbf{h}_i)$. We then discuss how one can obtain the Gaussian density that is the closest, in the sense of Kullback-Leibler divergence, to $\widetilde{q}^*_{\mathbf{h}_i}$.

First, to derive an explicit expression for $\widetilde{q}^*_{\mathbf{h}_i}$, note that

$$\log p(\mathbf{h}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i}) = c_{\mathbf{h}_i} - \frac{1}{2}\sum_{t=1}^{T} h_{i,t} - \frac{1}{2}\sum_{t=1}^{T} \mathrm{e}^{-h_{i,t}}(y_{i,t} - \mathbf{x}_{i,t}\boldsymbol{\theta}_i)^2 - \frac{1}{2\sigma^2_{h,i}}\sum_{t=1}^{T}(h_{i,t} - h_{i,t-1})^2,$$

where $c_{\mathbf{h}_i}$ is a constant not dependent on $\mathbf{h}_i$. Hence, taking expectation with respect to the marginal density $q_{-\mathbf{h}_i}(\boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i})$ gives

$$\begin{aligned}\mathbb{E}_{-\mathbf{h}_i}\left[\log p(\mathbf{h}_i \,|\, \mathbf{y}_i, \boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i})\right] = c_{\mathbf{h}_i} - \frac{1}{2}\sum_{t=1}^{T} h_{i,t} - \frac{1}{2}\sum_{t=1}^{T} \mathrm{e}^{-h_{i,t}}\widehat{s}^2_t - \frac{1}{2}\mathbb{E}_{\sigma^2_{h,i}}\left[\frac{1}{\sigma^2_{h,i}}\right] \\ \times \left(\sum_{t=2}^{T}(h_{i,t} - h_{i,t-1})^2 + (h_{i,1} - \widehat{h}_{i,0})^2 + \widehat{K}^{-1}_{h_{i,0}}\right),\end{aligned}$$

where $\widehat{s}^2_t = (y_{i,t} - \mathbf{x}_{i,t}\widehat{\boldsymbol{\theta}}_i)^2 + \mathrm{tr}(\mathbf{x}'_{i,t}\mathbf{x}_{i,t}\widehat{\mathbf{K}}^{-1}_{\boldsymbol{\theta}_i})$. Therefore, the (log) kernel of $\widetilde{q}^*_{\mathbf{h}_i}$ has the following explicit expression:

$$\log \widetilde{q}^*_{\mathbf{h}_i}(\mathbf{h}_i) = \widetilde{c}_{\mathbf{h}_i} - \frac{1}{2}\sum_{t=1}^{T} h_{i,t} - \frac{1}{2}\sum_{t=1}^{T} \mathrm{e}^{-h_{i,t}}\widehat{s}^2_t - \frac{1}{2}\mathbb{E}_{\sigma^2_{h,i}}\left[\frac{1}{\sigma^2_{h,i}}\right]\left(\sum_{t=2}^{T}(h_{i,t} - h_{i,t-1})^2 + (h_{i,1} - \widehat{h}_{i,0})^2\right),$$

where $\widetilde{c}_{\mathbf{h}_i}$ is a constant independent on $\mathbf{h}_i$.

As discussed in the main text, to locate the optimal Gaussian density $q^*_{\mathbf{h}_i}$, we set up a parametric optimization problem. First, consider the following family of Gaussian densities parameterized by the mean vector $\mathbf{m}$:

$$\mathcal{G} = \left\{ f_{\mathcal{N}}(\cdot; \mathbf{m}, \widehat{\mathbf{K}}^{-1}_{\mathbf{h}_i}) : \mathbf{m} \in \mathbb{R}^T \right\},$$

where $\widehat{\mathbf{K}}_{\mathbf{h}_i}$ is the negative Hessian of $\log \widetilde{q}^*_{\mathbf{h}_i}(\mathbf{h}_i)$ evaluated at the mode of $\log \widetilde{q}^*_{\mathbf{h}_i}(\mathbf{h}_i)$. For notational convenience, we write $f_{\mathbf{m}}(\cdot) \equiv f_{\mathcal{N}}(\cdot; \mathbf{m}, \widehat{\mathbf{K}}^{-1}_{\mathbf{h}_i})$. Then, we consider the minimization problem defined in (6), which for convenience we reproduce below:

$$\min_{f_{\mathbf{m}} \in \mathcal{G}} D_{KL}(f_{\mathbf{m}} || \widetilde{q}^*_{\mathbf{h}_i}) = \min_{\mathbf{m} \in \mathbb{R}^T} \mathbb{E} \log \left[\frac{f_{\mathbf{m}}(\mathbf{h}_i)}{\widetilde{q}^*_{\mathbf{h}_i}(\mathbf{h}_i)}\right],$$

where the expectation is taken with respect to the density $f_{\mathbf{m}}(\mathbf{h}_i)$. Next, we derive an

explicit expression for the objective function $\mathbb{E}\log\left[\frac{f_{\mathbf{m}}(\mathbf{h}_i)}{\widetilde{q}^*_{\mathbf{h}_i}(\mathbf{h}_i)}\right]$. First, note that

$$\log\left[\frac{f_{\mathbf{m}}(\mathbf{h}_i)}{\widetilde{q}^*_{\mathbf{h}_i}(\mathbf{h}_i)}\right] = c_1 - \frac{1}{2}(\mathbf{h}_i-\mathbf{m})'\widehat{\mathbf{K}}_{\mathbf{h}_i}(\mathbf{h}_i-\mathbf{m}) + \frac{1}{2}\left[\mathbf{1}_T'\mathbf{h}_i + (\widehat{\mathbf{s}}^2)'\mathrm{e}^{-\mathbf{h}_i} + \mathbb{E}_{\sigma^2_{h,i}}\left[\frac{1}{\sigma^2_{h,i}}\right](\mathbf{h}_i-\widehat{h}_{i,0}\mathbf{1}_T)'\mathbf{H}'\mathbf{H}(\mathbf{h}_i-\widehat{h}_{i,0}\mathbf{1}_T)\right],$$

where c_1 is a constant independent of $\mathbf{h}_i$ and $\mathbf{m}$ and $\widehat{\mathbf{s}}^2 = (\widehat{s}^2_1,\ldots,\widehat{s}^2_T)'$. Then, taking expectation with respect to $f_{\mathbf{m}}$, we obtain:

$$\mathbb{E}\log\left[\frac{f_{\mathbf{m}}(\mathbf{h}_i)}{\widetilde{q}^*_{\mathbf{h}_i}(\mathbf{h}_i)}\right] = c_2 + \frac{1}{2}\left[\mathbf{1}_T'\mathbf{m} + (\widehat{\mathbf{s}}^2)'\mathrm{e}^{-\mathbf{m}+\frac{1}{2}\widehat{\mathbf{d}}_i} + \mathbb{E}_{\sigma^2_{h,i}}\left[\frac{1}{\sigma^2_{h,i}}\right](\mathbf{m}-\widehat{h}_{i,0}\mathbf{1}_T)'\mathbf{H}'\mathbf{H}(\mathbf{m}-\widehat{h}_{i,0}\mathbf{1}_T)\right],$$

where c_2 is a constant independent of $\mathbf{m}$ and $\widehat{\mathbf{d}}_i$ is a $T\times 1$ vector consisting of the diagonal elements of $\widehat{\mathbf{K}}^{-1}_{\mathbf{h}_i}$. Given this expression, it can be easily verify that $\mathbb{E}\log\left[\frac{f_{\mathbf{m}}(\mathbf{h}_i)}{\widetilde{q}^*_{\mathbf{h}_i}(\mathbf{h}_i)}\right]$ is convex in $\mathbf{m}$. Hence, the minimization problem in (6) can be solved readily. Furthermore, the gradient and the Hessian of the objective function with respect to $\mathbf{m}$ can be computed easily:

$$\begin{aligned}\text{grad} &= \mathbb{E}_{\sigma^2_{h,i}}\left[\frac{1}{\sigma^2_{h,i}}\right]\mathbf{H}'\mathbf{H}(\mathbf{m}-\widehat{h}_{i,0}\mathbf{1}_T) + \frac{1}{2}(\mathbf{1}_T - (\widehat{\mathbf{s}}^2)'\mathrm{e}^{-\mathbf{m}+\frac{1}{2}\widehat{\mathbf{d}}_i}),\\ \text{Hess} &= \mathbb{E}_{\sigma^2_{h,i}}\left[\frac{1}{\sigma^2_{h,i}}\right]\mathbf{H}'\mathbf{H} + \frac{1}{2}\mathrm{diag}(\widehat{\mathbf{s}}^2\odot\mathrm{e}^{-\mathbf{m}+\frac{1}{2}\widehat{\mathbf{d}}_i}),\end{aligned}$$

where $\odot$ denotes the component-wise product. Note that the Hessian is a positive-definite matrix for all $\mathbf{m}\in\mathbb{R}^T$. Hence, Newton-Raphson method can be used to quickly solve the minimization problem. Further, since the Hessian is also a band matrix, fast routines for band matrices can be used to drastically speed up computations (see, e.g. Chan, 2017). Let $\widehat{\mathbf{h}}_i$ denote the unique minimizer. Finally, we use $f_{\mathcal{N}}(\cdot;\widehat{\mathbf{h}}_i,\widehat{\mathbf{K}}^{-1}_{\mathbf{h}_i})$ as the optimal density $q^*_{\mathbf{h}_i}$.

The Variational Lower Bound

Next, we derive the variational lower bound $\underline{p}(\mathbf{y}_i; q)$. To that end, we first compute the log ratio of the joint posterior density and the variational approximation:

$$\begin{aligned}
\log\left[\frac{p(\mathbf{y}_i, \mathbf{h}_i \,|\, \boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i}) p(\boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i})}{q(\boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i}, \mathbf{h}_i)}\right] &= c_i - \frac{1}{2}\mathbf{1}_T'\mathbf{h}_i - \frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}_i)'\boldsymbol{\Omega}_{\mathbf{h}_i}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\theta}_i) \\
&- \frac{T}{2}\log\sigma^2_{h,i} - \frac{1}{2\sigma^2_{h,i}}(\mathbf{h}_i - h_{i,0}\mathbf{1}_T)'\mathbf{H}'\mathbf{H}(\mathbf{h}_i - h_{i,0}\mathbf{1}_T) - \frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0,i})'\mathbf{V}_{\boldsymbol{\theta}_i}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0,i}) \\
&- \frac{1}{2V_{h_{i,0}}}h_{i,0}^2 - (\nu_i + 1)\log\sigma^2_{h,i} - \frac{S_i}{\sigma^2_{h,i}} + \frac{1}{2}(\mathbf{h}_i - \widehat{\mathbf{h}}_i)'\widehat{\mathbf{K}}_{\mathbf{h}_i}(\mathbf{h}_i - \widehat{\mathbf{h}}_i) \\
&+ \frac{1}{2}(\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_i)'\widehat{\mathbf{K}}_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_i) + \frac{\widehat{K}_{h_{i,0}}}{2}(h_{i,0} - \widehat{h}_{i,0})^2 + (\widehat{\nu}_i + 1)\log\sigma^2_{h,i} + \frac{\widehat{S}_i}{\sigma^2_{h,i}},
\end{aligned}$$

where $\boldsymbol{\Omega}_{\mathbf{h}_i} = \mathrm{diag}(\mathrm{e}^{h_{i,1}}, \ldots, \mathrm{e}^{h_{i,T}})$ and $c_i = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\log V_{h_{i,0}} - \frac{1}{2}\log|\mathbf{V}_{\boldsymbol{\theta}_i}| + \nu_i \log S_i - \log\Gamma(\nu_i) - \frac{1}{2}\log|\widehat{\mathbf{K}}_{\mathbf{h}_i}| - \frac{1}{2}\log|\widehat{\mathbf{K}}_{\boldsymbol{\theta}_i}| - \frac{1}{2}\log\widehat{K}_{h_{i,0}} - \widehat{\nu}_i\log\widehat{S}_i + \log\Gamma(\widehat{\nu}_i)$. Taking expectation of the above log ratio with respect to q, we obtain the variational lower bound:

$$\begin{aligned}
\underline{p}(\mathbf{y}_i; q) &= \mathbb{E}_q\left\{\log\left[\frac{p(\mathbf{y}_i, \mathbf{h}_i \,|\, \boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i}) p(\boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i})}{q(\boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i}, \mathbf{h}_i)}\right]\right\} \\
&= c_i - \frac{1}{2}\mathbf{1}_T'\widehat{\mathbf{h}}_i - \frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\theta}}_i)'\widehat{\mathbf{O}}_{\mathbf{h}_i}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\theta}}_i) - \frac{1}{2}\mathrm{tr}(\mathbf{X}_i'\widehat{\mathbf{O}}_{\mathbf{h}_i}\mathbf{X}_i\widehat{\mathbf{K}}_{\boldsymbol{\theta}_i}^{-1}) - \frac{1}{2V_{h_{i,0}}}(\widehat{h}_{i,0}^2 + \widehat{K}_{h_{i,0}}^{-1}) \\
&\quad - \frac{1}{2}(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{i,0})'\mathbf{V}_{\boldsymbol{\theta}_i}^{-1}(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{i,0}) - \frac{1}{2}\mathrm{tr}(\mathbf{V}_{\boldsymbol{\theta}_i}^{-1}\widehat{\mathbf{K}}_{\boldsymbol{\theta}_i}^{-1}) + \frac{1}{2}(T + k_i + 1) \\
&= c_i - \frac{1}{2}\mathbf{1}_T'\widehat{\mathbf{h}}_i - \frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\theta}}_i)'\widehat{\mathbf{O}}_{\mathbf{h}_i}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\theta}}_i) - \frac{1}{2V_{h_{i,0}}}(\widehat{h}_{i,0}^2 + \widehat{K}_{h_{i,0}}^{-1}) \\
&\quad - \frac{1}{2}(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{i,0})'\mathbf{V}_{\boldsymbol{\theta}_i}^{-1}(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{i,0}) + \frac{1}{2}(T + 1),
\end{aligned}$$

where $\widehat{\mathbf{O}}_{\mathbf{h}_i} = \mathrm{diag}(\mathrm{e}^{-\widehat{h}_{i,1} + \frac{1}{2}d_{i,t}}, \ldots, \mathrm{e}^{-\widehat{h}_{i,T} + \frac{1}{2}d_{i,T}})$ and k_i is the dimension of $\boldsymbol{\theta}_i$. Note that the second equality in the above derivation follows from

$$-\frac{1}{2}\mathrm{tr}(\mathbf{X}_i'\widehat{\mathbf{O}}_{\mathbf{h}_i}\mathbf{X}_i\widehat{\mathbf{K}}_{\boldsymbol{\theta}_i}^{-1}) - \frac{1}{2}\mathrm{tr}(\mathbf{V}_{\boldsymbol{\theta}_i}^{-1}\widehat{\mathbf{K}}_{\boldsymbol{\theta}_i}^{-1}) = -\frac{1}{2}\mathrm{tr}((\mathbf{V}_{\boldsymbol{\theta}_i}^{-1} + \mathbf{X}_i'\widehat{\mathbf{O}}_{\mathbf{h}_i}\mathbf{X}_i)\widehat{\mathbf{K}}_{\boldsymbol{\theta}_i}^{-1}) = -\frac{1}{2}\mathrm{tr}(\mathbf{I}_{k_i}) = -\frac{1}{2}k_i.$$

We summarize the variational Bayes iterative scheme below.

Algorithm 1 Iterative scheme for obtaining the parameters in the optimal densities $q^*(\boldsymbol{\theta}_i, h_{i,0}, \sigma^2_{h,i}, \mathbf{h}_i)$.

Initialize: $\widehat{\mathbf{K}}_{\boldsymbol{\theta}_i}$, $\widehat{\boldsymbol{\theta}}_i$, $\widehat{\nu}_i$, $\widehat{S}_i$, $\widehat{h}_{i,0}$. Set $\widehat{\nu}_i = \nu_i + \frac{T}{2}$.
Cycle:

$$
\begin{aligned}
\widehat{\mathbf{K}}_{\mathbf{h}_i}, \widehat{\mathbf{h}}_i &\leftarrow \text{obtained by Newton-Raphson method} \\
\widehat{\mathbf{K}}_{\boldsymbol{\theta}_i} &\leftarrow \mathbf{V}_{\boldsymbol{\theta}_i}^{-1} + \mathbf{X}_i' \widehat{\mathbf{O}}_{\mathbf{h}_i} \mathbf{X}_i \\
\widehat{\boldsymbol{\theta}}_i &\leftarrow \widehat{\mathbf{K}}_{\boldsymbol{\theta}_i}^{-1} (\mathbf{V}_{\boldsymbol{\theta}_i}^{-1} \boldsymbol{\theta}_{0,i} + \mathbf{X}_i' \widehat{\mathbf{O}}_{\mathbf{h}_i} \mathbf{y}_i) \\
\widehat{S}_i &\leftarrow S_i + \frac{1}{2} \left[(\widehat{\mathbf{h}}_i - \widehat{h}_{i,0} \mathbf{1}_T)' \mathbf{H}' \mathbf{H} (\widehat{\mathbf{h}}_i - \widehat{h}_{i,0} \mathbf{1}_T) + \operatorname{tr}(\mathbf{H}' \mathbf{H} \widehat{\mathbf{K}}_{\mathbf{h}_i}^{-1}) + \widehat{K}_{h_{i,0}}^{-1} \right] \\
\widehat{K}_{h_{i,0}} &\leftarrow V_{h_{i,0}}^{-1} + \frac{\widehat{\nu}_i}{\widehat{S}_i} \\
\widehat{h}_{i,0} &\leftarrow \widehat{K}_{h_{i,0}}^{-1} \frac{\widehat{\nu}_i}{\widehat{S}_i} \widehat{h}_{i,1}
\end{aligned}
$$

until the increase in $\underline{p}(\mathbf{y}_i; q)$ is negligible.

Appendix B: Additional Results

In this appendix we provide additional empirical results from the bank network connectedness application.

First, Figure 5 plots the dynamic system-wide connectedness measure from the 96-variable homoscedastic VAR.

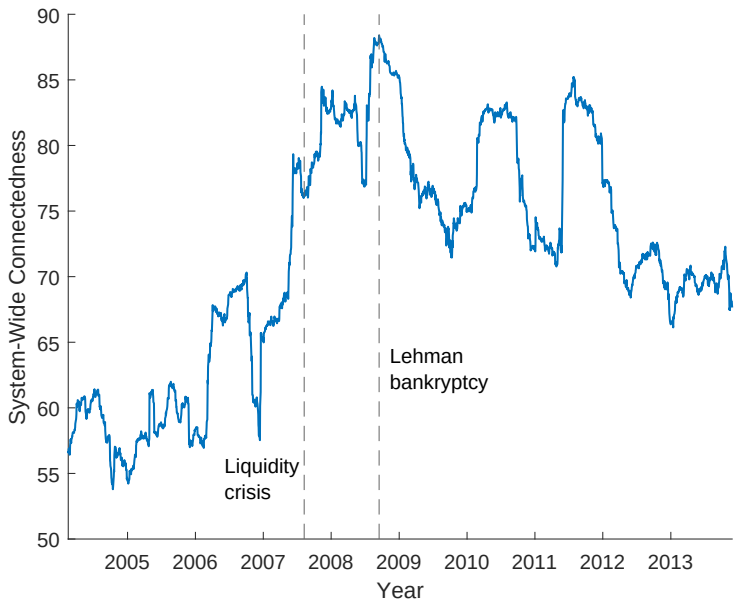

Figure 5: The dynamic system-wide connectedness measure from the homoscedastic VAR using a 150-day rolling window.

Next, we report in Figure 6 the decomposition of the dynamic system-wide connectedness measure into cross-country and within-country system-wide connectedness components.

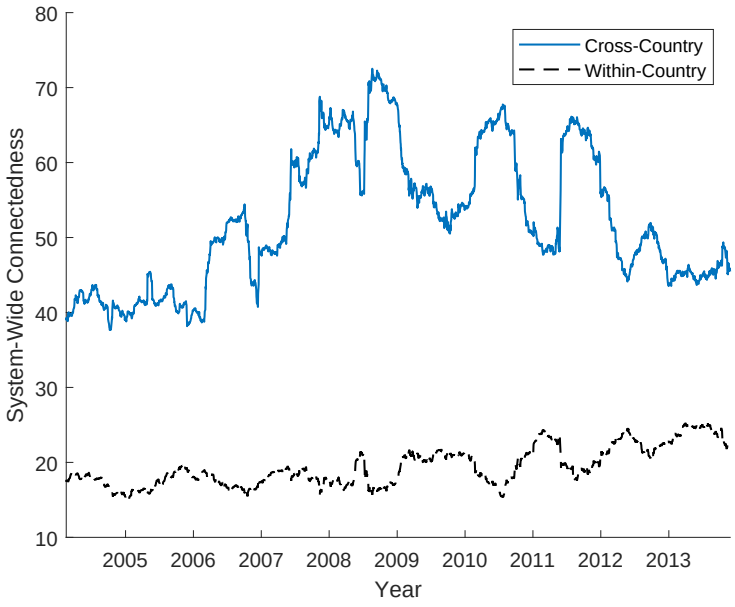

Figure 6: The decomposition of the dynamic system-wide connectedness measure from the homoscedastic VAR using a 150-day rolling window into cross-country and within-country components.

References

Banbura, M., D. Giannone, M. Modugno, and L. Reichlin (2013): "Now-casting and the real-time data flow," in *Handbook of Economic Forecasting*, vol. 2, pp. 195–237. Elsevier.

Banbura, M., D. Giannone, and L. Reichlin (2010): "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 25(1), 71–92.

Baumeister, C., and J. D. Hamilton (2015): "Sign restrictions, structural vector autoregressions, and useful prior information," *Econometrica*, 83(5), 1963–1999.

Bhattacharya, A., D. Pati, N. S. Pillai, and D. B. Dunson (2015): "Dirichlet–Laplace priors for optimal shrinkage," *Journal of the American Statistical Association*, 110(512), 1479–1490.

Bishop, C. M. (2006): *Pattern Recognition and Machine Learning*. springer.

Carriero, A., T. E. Clark, and M. G. Marcellino (2015): "Bayesian VARs: Specification Choices and Forecast Accuracy," *Journal of Applied Econometrics*, 30(1), 46–73.

——— (2016): "Common drifting volatility in large Bayesian VARs," *Journal of Business and Economic Statistics*, 34(3), 375–390.

——— (2019): "Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors," *Journal of Econometrics*, Forthcoming.

Carriero, A., G. Kapetanios, and M. Marcellino (2009): "Forecasting exchange rates with a large Bayesian VAR," *International Journal of Forecasting*, 25(2), 400–417.

Chan, J. C. C. (2017): "The Stochastic Volatility in Mean Model with Time-Varying Parameters: An Application to Inflation Modeling," *Journal of Business and Economic Statistics*, 35(1), 17–28.

——— (2020a): "Large Bayesian VARs: A Flexible Kronecker Error Covariance Structure," *Journal of Business and Economic Statistics*, 38(1), 68–79.

——— (2020b): "Large Bayesian Vector Autoregressions," in *Macroeconomic Forecasting in the Era of Big Data*, ed. by P. Fuleky, pp. 95–125. Springer.

Chan, J. C. C., and E. Eisenstat (2018): "Bayesian Model Comparison for Time-Varying Parameter VARs with Stochastic Volatility," *Journal of Applied Econometrics*, 33(4), 509–532.

Chan, J. C. C., E. Eisenstat, and R. W. Strachan (2020): "Reducing the State Space Dimension in a Large TVP-VAR," *Journal of Econometrics*, 218(1), 105–118.

Chan, J. C. C., and A. L. Grant (2016): "On the Observed-Data Deviance Information Criterion for Volatility Modeling," *Journal of Financial Econometrics*, 14(4), 772–802.

Chan, J. C. C., G. Koop, D. J. Poirier, and J. L. Tobias (2019): *Bayesian Econometric Methods*. Cambridge University Press, 2 edn.

Clark, T. E. (2011): "Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility," *Journal of Business and Economic Statistics*, 29(3), 327–341.

Cogley, T., and T. J. Sargent (2005): "Drifts and volatilities: Monetary policies and outcomes in the post WWII US," *Review of Economic Dynamics*, 8(2), 262–302.

Cross, J., C. Hou, and A. Poon (2020): "Macroeconomic forecasting with large Bayesian VARs: Global-local priors and the illusion of sparsity," *International Journal of Forecasting*, 36(3), 899–915.

Cross, J., and A. Poon (2016): "Forecasting structural change and fat-tailed events in Australian macroeconomic variables," *Economic Modelling*, 58, 34–51.

D'Agostino, A., L. Gambetti, and D. Giannone (2013): "Macroeconomic forecasting and structural change," *Journal of Applied Econometrics*, 28, 82–101.

Demirer, M., F. X. Diebold, L. Liu, and K. Yilmaz (2018): "Estimating global bank network connectedness," *Journal of Applied Econometrics*, 33(1), 1–15.

Diebold, F. X., and K. Yilmaz (2009): "Measuring financial asset return and volatility spillovers, with application to global equity markets," *The Economic Journal*, 119(534), 158–171.

Diebold, F. X., and K. Yilmaz (2014): "On the network topology of variance decompositions: Measuring the connectedness of financial firms," *Journal of Econometrics*, 182(1), 119–134.

Durbin, J., and S. J. Koopman (1997): "Monte Carlo maximum likelihood estimation for non-Gaussian state space models," *Biometrika*, 84, 669–684.

Eisenstat, E., J. C. C. Chan, and R. W. Strachan (2016): "Stochastic Model Specification Search for Time-Varying Parameter VARs," *Econometric Reviews*, 35(8-10), 1638–1665.

Ellahie, A., and G. Ricco (2017): "Government purchases reloaded: Informational insufficiency and heterogeneity in fiscal VARs," *Journal of Monetary Economics*, 90, 13–27.

Garman, M. B., and M. J. Klass (1980): "On the estimation of security price volatilities from historical data," *Journal of business*, pp. 67–78.

Gefang, D., G. Koop, and A. Poon (2019): "Variational Bayesian inference in large Vector Autoregressions with hierarchical shrinkage," *CAMA Working Paper*.

Giannone, D., M. Lenza, and G. Primiceri (2017): "Economic predictions with big data: The illusion of sparsity," *CEPR Discussion Paper No. DP12256*.

Giannone, D., M. Lenza, and G. E. Primiceri (2015): "Prior selection for vector autoregressions," *Review of Economics and Statistics*, 97(2), 436–451.

Griffin, J., and P. Brown (2017): "Hierarchical shrinkage priors for regression models," *Bayesian Analysis*, 12(1), 135–159.

Hajargasht, G., and T. Woźniak (2018): "Accurate computation of marginal data densities using variational Bayes," *arXiv preprint arXiv:1805.10036*.

Harvey, A., E. Ruiz, and N. Shephard (1994): "Multivariate stochastic variance models," *The Review of Economic Studies*, 61(2), 247–264.

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

Karlsson, S. (2013): "Forecasting with Bayesian vector autoregressions," in *Handbook of Economic Forecasting*, ed. by G. Elliott, and A. Timmermann, vol. 2 of *Handbook of Economic Forecasting*, pp. 791–897. Elsevier.

Kastner, G., and F. Huber (2020): "Sparse Bayesian vector autoregressions in huge dimensions," *Journal of Forecasting*, 37(7).

Koop, G. (2013): "Forecasting with medium and large Bayesian VARs," *Journal of Applied Econometrics*, 28(2), 177–203.

Koop, G., and D. Korobilis (2010): "Bayesian Multivariate Time Series Methods for Empirical Macroeconomics," *Foundations and Trends in Econometrics*, 3(4), 267–358.

——— (2013): "Large time-varying parameter VARs," *Journal of Econometrics*, 177(2), 185–198.

——— (2018): "Variational Bayes inference in high-dimensional time-varying parameter models," *Available at SSRN 3246472*.

Korobilis, D., and K. Yilmaz (2018): "Measuring dynamic connectedness with large Bayesian VAR models," *Available at SSRN 3099725*.

Loaiza-Maya, R., M. Smith, D. Nott, and P. Danaher (2020): "Fast and Accurate Variational Inference for Models with Many Latent Variables," *arXiv preprint arXiv:2005.07430*.

McCracken, M. W., M. Owyang, and T. Sekhposyan (2015): "Real-time forecasting with a large, mixed frequency, Bayesian VAR," *FRB St. Louis Working Paper*, 2015-30.

Morley, J., and B. Wong (2020): "Estimating and accounting for the output gap with large Bayesian vector autoregressions," *Journal of Applied Econometrics*, 35(1), 1–18.

Ormerod, J. T., and M. P. Wand (2010): "Explaining variational approximations," *The American Statistician*, 64(2), 140–153.

Primiceri, G. E. (2005): "Time Varying Structural Vector Autoregressions and Monetary Policy," *Review of Economic Studies*, 72(3), 821–852.

Shephard, N., and M. K. Pitt (1997): "Likelihood analysis of non-Gaussian measurement time series," *Biometrika*, 84, 653–667.

You, C., J. T. Ormerod, and S. Mueller (2014): "On variational Bayes estimation and variational information criteria for linear regression models," *Australian & New Zealand Journal of Statistics*, 56(1), 73–87.

Zou, H., and H. Zhang (2009): "On the adaptive elastic-net with a diverging number of parameters," *Annals of Statistics*, 37(4), 1733.