Large Bayesian VARs for Forecasting:

Shrinkage Priors, Stochastic Volatility and Computation

Joshua Chan
Purdue University

11th ECB Conference on Forecasting Techniques

Why Large VARs?

VARs are the main workhorse in empirical macroeconomics

Growing need to include more information/variables

Large VARs are increasingly used in applications:

- lots of variables for a single country (Banbura et al., 2010)
- few variables for many countries (Koop and Korobilis, 2016)
- mixed frequency data (McCracken et al., 2016)
- disaggregate data (Giannone et al., 2014; Ellahie and Ricco, 2017)
- firm-level data (Demirer et al., 2018)

VARs VS Factor Models

Viable alternative to factor models

There are a wide variety of VARs:

- steady-state, regime-switching, smooth transition, panel, factor-augmented, time-vary parameter...

Can use all the machinery developed for VARs:

- many identification schemes, impulse-responses, forecast error variance decompositions, historical decompositions...

Three Themes for the Talk

1. Hierarchical shrinkage priors for VAR coefficients
   - VARs have lots of coefficients, large VARs especially so
   - appropriate shrinkage/regularization is key
   - shrinkage is necessary, but computation can be intensive

2. Comparing SV specifications for large VARs
   - for small VARs, time-varying volatility is empirically important
   - a few SV specifications designed for large systems
   - adding SV makes estimation even more time consuming

3. Fast algorithms for estimation and model comparison

Minnesota Priors

Many versions:

- original; fixed covariance matrix (Doan, Litterman, and Sims, 1984; Litterman, 1986)
- unknown covariance matrix (Kadiyala and Karlsson, 1993, 1997)
- data-based hyperparameters (Giannone, Lenza, and Primiceri, 2015)

Have enjoyed great success, but recently been criticized for not being adaptive enough

(Shrink both ‘large’ and ‘small’ coefficients)

Two Most Relevant Features for Minnesota Priors

Cross-variable shrinkage:

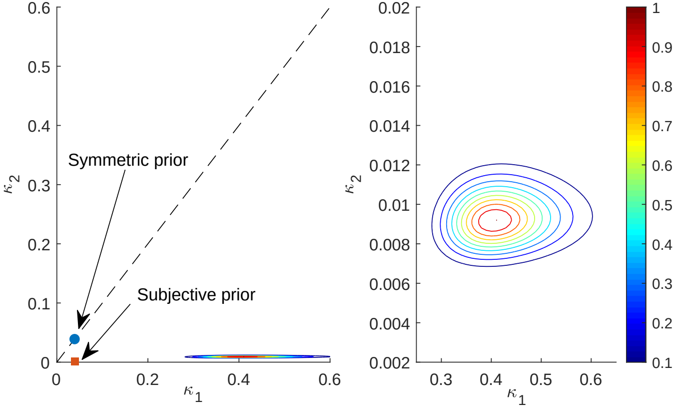
- shrinking coefficients on 'own' lags more aggressively than on 'other' lags
- different hyperparameters, κ_1 and κ_2, to control shrinkage strength for own vs other lags
- present in the original version, but less common in large VARs

Data-based hyperparameters:

- estimate κ_1 and κ_2 from the data
- κ_1 and κ_2 are expected to vary across types of variables, sample periods, countries, frequency...

Joint Posterior Density of κ_1 and κ_2

Results from a 21-variable VAR without SV
(see Asymmetric Conjugate Priors for Large Bayesian VARs)



Adaptive Hierarchical Priors

Many proposals:

- normal-gamma (Lasso as a special case) (Griffin and Brown, 2010; Huber and Feldkircher, 2019)
- horseshoe (Carvalho, Polson and Scott, 2010; Follett and Yu, 2019)
- Dirichlet-Laplace (Bhattacharya et al., 2015; Kastner and Huber, 2018)

Heavy tails with substantial mass at 0 — tend to shrink only ‘small’ coefficients

But don’t seem to forecast better than a data-based Minnesota prior (Cross, Hou and Poon, 2019)

Minnesota VS Adaptive Hierarchical Priors

While adaptive hierarchical priors have good theoretical properties, they treat all variables identically

In contrast, Minnesota priors incorporate richer prior beliefs:

- cross-variable shrinkage
- shrinking coefficients on higher lags more aggressively
- adjust coefficient prior variances by the variability of the variables

Best of Both Worlds

(see Minnesota-Type Adaptive Hierarchical Priors for Large Bayesian VARs)

New priors that capture the best features of both families

Like adaptive hierarchical priors: heavy tails, substantial mass at 0, good theoretical properties

Similar to Minnesota priors: richer prior beliefs about the coefficients

Forecast better than both families in the context of large VARs with SV

Some Details

For the j-th coefficient in the i-th equation:

$$(\theta_{i,j} \,|\, \kappa_1, \kappa_2, \psi_{i,j}) \sim \mathcal{N}(m_{i,j}, \kappa_{i,j}\psi_{i,j}C_{i,j})$$

- $\kappa_{i,j} = \kappa_1$ or κ_2 is a global variance component common to many coefficients
- $\psi_{i,j} \sim F_\psi(\psi_{i,j})$ is a local variance component
- $C_{i,j}$ is a constant that incorporates richer prior beliefs

This setup includes both the Minnesota and global-local priors

Empirical Application

Focus on the Minnesota-type normal-gamma prior:

$$(\theta_{i,j} \,|\, \kappa_1, \kappa_2, \psi_{i,j}) \sim \mathcal{N}(m_{i,j}, 2\kappa_{i,j}\psi_{i,j}C_{i,j}),$$
$$\psi_{i,j} \sim \mathcal{G}(\nu_\psi, \nu_\psi/2)$$

If $\nu_\psi = 1$, this reduces to Lasso

Compare to the data-based Minnesota prior and the normal-gamma prior using a dataset of 23 quarterly US variables

All models have Cholesky SV (Cogley and Sargent, 2005)

Estimation Results

	normal-gamma	Minnesota-type normal-gamma
κ_1	0.0007	0.041
	(0.0001)	(0.0171)
κ_2	0.0007	0.0006
	(0.0001)	(0.0001)
ν_ψ	0.13	0.15
	(0.004)	(0.012)

- under the new prior, κ_1 increases 58 times and κ_2 decreases
- find strong evidence of cross-variable shrinkage
- ν_ψ is very small — Lasso might be too restrictive

	Minnesota	Minnesota-type normal-gamma
κ_1	0.093	0.041
	(0.0152)	(0.0171)
κ_2	0.0028	0.0006
	(0.0003)	(0.0001)

- both κ_1 and κ_2 are substantially larger under Minnesota
- local component handles ‘large’ coefficients; global component shrinks the coefficients more aggressively

VARs with SV

A few recent SV specifications designed for large systems:

- Carriero, Clark and Marcellino (2016) consider a large VAR with a common SV
- Carriero, Clark and Marcellino (2019) estimate a 125-variable VAR with 125 SV processes
- Kastner (2019) considers a huge, sparse factor SV model

Interesting trade-off between parsimony, flexibility and speed of estimation

Lack of tools to select among these SV models

Comparing SV Specifications

(see Comparing Stochastic Volatility Specifications for Large Bayesian VARs)

Develop new methods to compute marginal likelihoods of large VARs

Key ingredients: conditional Monte Carlo and adaptive importance sampling

- analytically integrate out the VAR coefficients
- construct an adaptive importance sampling estimator to integrate out the SV/factors via Monte Carlo

Importance sampling density obtained by minimizing the Kullback-Leibler divergence to the ideal zero-variance density

Common Stochastic Volatility

Consider the following VAR with the common SV (VAR-CSV):

$$\mathbf{y}_t = \mathbf{a}_0 + \mathbf{A}_1\mathbf{y}_{t-1} + \cdots + \mathbf{A}_p\mathbf{y}_{t-p} + \varepsilon_t \quad \varepsilon_t \sim \mathcal{N}(\mathbf{0}, e^{h_t}\boldsymbol{\Sigma})$$

- $\mathbf{a}_0$ is an $n \times 1$ vector of intercepts; $\mathbf{A}_1, \ldots, \mathbf{A}_p$ are all $n \times n$ coefficient matrices
- the error covariance matrix is scaled by a common, time-varying factor that can be interpreted as the overall macroeconomic volatility
- h_t follows a zero-mean AR(1) process

Pros:

- if the natural conjugate prior is used, estimation is fast — minutes even for very large systems
- complexity: $\mathcal{O}(n^3)$ as opposed to $\mathcal{O}(n^4)$ for other SV specifications

Cons:

- seemingly restrictive — only one common SV
- the natural conjugate prior does not allow for cross-variable shrinkage

Cholesky Stochastic Volatility

Consider a VAR with SV in the structural form (VAR-SV):

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

- $\mathbf{A}_0$ is an $n \times n$ lower triangular matrix with ones on the diagonal
- $\mathbf{\Sigma}_t = \text{diag}(\exp(h_{1,t}), \ldots, \exp(h_{n,t}))$

Each of the log-volatility $h_{i,t}$ follows an AR(1) process

Recursive system; can estimate it equation by equation

Complexity is $\mathcal{O}(n^4)$

Much more flexible than VAR-CSV: n SV processes instead of one

Can also accommodate more flexible priors

Factor Stochastic Volatility

$$\mathbf{y}_t = \mathbf{a}_0 + \mathbf{A}_1\mathbf{y}_{t-1} + \cdots + \mathbf{A}_p\mathbf{y}_{t-p} + \varepsilon_t,$$
$$\varepsilon_t = \mathbf{L}\mathbf{f}_t + \mathbf{u}_t,$$
$$\begin{pmatrix} \mathbf{u}_t \\ \mathbf{f}_t \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_t \end{pmatrix}\right)$$

- $\mathbf{f}_t = (f_{1,t}, \ldots, f_{r,t})'$ is a $r \times 1$ vector of latent factors
- $\mathbf{L}$ is the associated $n \times r$ factor loading matrix
- $\mathbf{\Sigma}_t = \text{diag}(e^{h_{1,t}}, \ldots, e^{h_{n,t}})$ and $\mathbf{\Omega}_t = \text{diag}(e^{h_{n+1,t}}, \ldots, e^{h_{n+r,t}})$

Each of the log-volatility $h_{i,t}$ follows an AR(1) process

Given the factors, the n equations are unrelated; can estimate the system equation by equation

Complexity is $\mathcal{O}(n^4)$

Very flexible covariance structure: $n + r$ SV processes

Can also accommodate more flexible priors

Monte Carlo Experiments

Conduct a series of Monte Carlo experiments (100 datasets each)

Show that the new method can

- distinguish common SV, Cholesky SV and factor SV
- discriminate between homoskedastic vs heteroskedastic models
- identify the correct number of factors in FSV

Empirical Application

Compare VARs of different sizes along two dimensions

SV specifications: common SV, Cholesky SV, factor SV

Minnesota priors with and without 2 features:

- cross-variable shrinkage
- fixed vs estimated shrinkage hyperparameters

Using datasets of 7, 15 and 30 US quarterly macro and financial variables

Comparing SV Specifications

	VAR-CSV	VAR-SV	VAR-FSV
$n = 7$	$-2{,}410$	$-2{,}312$	$-2{,}318$
	(0.1)	(0.3)	(0.4)
$n = 15$	$-6{,}618$	$-6{,}442$	$-6{,}454$
	(0.1)	(0.4)	(0.8)
$n = 30$	$-12{,}024$	$-11{,}555$	$-11{,}567$
	(0.1)	(0.6)	(1.8)

- Cholesky SV and FSV perform much better than common SV for all n
- Cholesky SV is the best, FSV close second
- (all 3 SV models outperform the homoskedastic VAR)

Comparing Shrinkage Priors

Compare different types of Minnesota priors

Focus on 2 features: cross-variable shrinkage and fixed vs estimated shrinkage hyperparameters

Consider two benchmarks:

- Symmetric prior: set $\kappa_1 = \kappa_2$
- Subjective prior: set $\kappa_1 = 0.04$ and $\kappa_2 = 0.0016$

Focus on $n = 15$

Symmetric Vs Asymmetric Priors

	VAR-SV	VAR-FSV ($k = 4$)
Symmetric prior	−6,588	−6,658
	(0.4)	(1.1)
Asymmetric prior	−6,442	−6,454
	(0.5)	(0.8)

- for both models, the asymmetric prior significantly outperforms the symmetric version
- strong evidence for cross-variable shrinkage

Subjective Vs Asymmetric Priors

	VAR-CSV	VAR-SV	VAR-FSV ($k = 4$)
Subjective prior	$-6{,}702$	$-6{,}597$	$-6{,}491$
	(0.1)	(0.4)	(0.9)
Symmetric prior	$-6{,}618$	$-6{,}588$	$-6{,}658$
	(0.1)	(0.4)	(1.1)
Asymmetric prior	-	$-6{,}442$	$-6{,}454$
		(0.5)	(0.8)

- also beneficial to estimate the shrinkage hyperparameters rather than fixing them subjectively
- hard to have one set of hyperparameter values that work well for different variables and sample periods

Decomposing Gains in ML

	VAR-CSV	VAR-SV	VAR-FSV ($k = 4$)
Symmetric prior	−6,618	−6,588	−6,658
	(0.1)	(0.4)	(1.1)
Asymmetric prior	-	−6,442	−6,454
		(0.5)	(0.8)

- superior performance of Cholesky SV and FSV can mostly be attributed to the more flexible priors
- e..g, for VAR-SV $-6,442 + 6618 = 173$; 30 comes from more flexible likelihood, 146 comes from more flexible prior
- starker conclusion for VAR-FSV

A Few Tips for Estimating Large VARs

Equation-by-equation estimation (Carriero, Clark and Marcellino, 2019)

Reparameterize the system to get n unrelated regressions

Use precision sampler (instead of Kalman Filter) to draw SV/factors

Equation-by-Equation Estimation

Sample VAR coefficients equation by equation instead of drawing them in one step

Computational complexity reduces from $\mathcal{O}(n^6)$ to $\mathcal{O}(n^4)$

(There's an issue in the original algorithm, Carriero, Chan, Clark and Marcellino (2021) have a fix with the same order of complexity)

10 to 50 times faster for n up to 40

(Run out of memory for larger n)

Reparameterization

Under the structural-form VAR, we have n unrelated regressions

Can estimate the equations in parallel (embarrassingly parallel)

5 to 10 times faster compared to Carriero, Chan, Clark and Marcellino (2021) — even before parallelization

Precision Sampler

Draw SV or factors in one block using the precision sampler (Chan and Jeliazkov, 2009)

Works for any conditionally linear Gaussian and some nonlinear state space models

Key idea: the precision matrix of the states is banded

2 to 10 times faster compared to Kalman-filter based smoothers

Still Too Slow?

If everything fails, ditch MCMC (?!)

A promising alternative is variational Bayes

Approximate the posterior using a convenient parametric family by minimizing the Kullback-Leibler divergence

Not an exact method like MCMC, but substantially faster (minutes instead of hours/days)

For large VAR applications, see Gefang, Koop, and Poon (2019) and Chan and Yu (2020)

Main Takeaways

Useful features for shrinkage priors:

- cross-variable shrinkage
- data-dependent hyperparameters
- heavier tails than Gaussian/local variance component

SV is empirically important

Cholesky SV is the best, but FSV is also competitive

Choosing a flexible shrinkage prior is as important as selecting a flexible SV specification

Thank You for Your Attention!

This talk is based on

- Comparing Stochastic Volatility Specifications for Large Bayesian VARs
- Minnesota-Type Adaptive Hierarchical Priors for Large Bayesian VARs
- Asymmetric Conjugate Priors for Large Bayesian VARs
- Fast and Accurate Variational Inference for Large Bayesian VARs with Stochastic Volatility (joint with Xuewen Yu)

For working papers and codes, google

`joshua chan purdue`