Composite Likelihood Methods for Large Bayesian VARs with Stochastic Volatility

Joshua Chan[1] Eric Eisenstat[2] Chenghan Hou[3] Gary Koop[4]

[1]Purdue University [2]University of Queensland [3]Hunan University [4]University of Strathclyde

Background: History of Large VARs

- Large VARs, involving 100 or more dependent variables, are increasingly used in a variety of macroeconomic applications.
- Pioneering paper: Banbura, Giannone and Reichlin (2010, JAE) "Large Bayesian Vector Autoregressions"
- Previous VARs: a few variables perhaps 10 at most
- Sample of recent large VAR papers:
- Carriero, Kapetanios and Marcellino (2009, IJF): exchange rates for many countries
- Carriero, Kapetanios and Marcellino (2012, JBF): US government bond yields of different maturities
- Giannone, Lenza, Momferatou and Onorante (2010): euro area inflation forecasting (components of inflation)
- Koop and Korobilis (2016, EER) eurozone sovereign debt crisis
- Kastner and Huber (2018): US macro forecasting
- Jarociński and Maćkowiak (2016, ReStat): Granger causality
- Banbura, Giannone and Lenza (2014, ECB): conditional forecasts/scenario analysis in euro area

Background: Why large VARs?

- Availability of more data
- More data means more information, makes sense to include it
- Concerns about missing out important information (omitted variables bias, fundamentalness, etc.)
- The main alternatives are factor models
- Principal components squeeze information in large number of variables to small number of factors
- But this squeezing is done without reference to explanatory power (i.e. squeeze first then put in regression model or VAR): "unsupervised"
- Large VAR methods are supervised and can easily see role of individual variables
- And they work: often beating factor methods in forecasting competitions

Background: Computation in large VARs

- E.g. large VAR with $N = 100$ variables and a lag length of $p = 13$:
- $100,000+$ VAR coefficients
- $5,050$ free parameters in error covariance.
- Bayesian prior shrinkage surmounts over-parameterization
- Standard choices exist: e.g. Minnesota prior
- Key point 1: Standard approaches are conjugate: analytical results exist (estimation and forecasting – no MCMC needed)
- Key point 2: Huge posterior covariance of VAR coefficients ($N^2p \times N^2p$ matrix): tough computation
- Key point 3: Conjugacy greatly simplifies: separately manipulate $N \times N$ and $Np \times Np$ matrices
- Key point 4: Using more realistic priors or extending model (e.g. to relax homoskedasticity assumption) loses conjugacy and, thus, computational feasibility
- Bottom line: Great tools exist for large homoskedastic Bayesian VARs with a particular prior, but cannot extend

Background: Multivariate Stochastic Volatility in VARs

- Allowing for error variances to change in macroeconomic VARs important
- E.g. Primiceri (2005, ReStud), Sims and Zha (2006, AER), Clark (2011, JBES), etc.
- Research question: How to add multivariate stochastic volatility in large VARs?
- Existing Bayesian literature is either:
- Homoskedastic
- Restrictive forms (e.g. Clark, Carriero and Marcellino, 2016, JBES + 2 working papers, Chan, 2016, working paper)
- Approximations (Koop and Korobilis, JOE, JOE and Koop, Korobilis and Pettenuzzo, 2016, JOE)
- Present paper: new approach using composite likelihoods

Vector Autoregressions with Stochastic Volatility (VAR-SV)

- y_t is N-vector of dependent variables (N large)
- VAR-SV is:

$$A_{0t} y_t = c + A_1 y_{t-1} + \cdots A_p y_{t-p} + \epsilon_t, \quad \epsilon_t \sim N(0, \Sigma_t),$$

- $\Sigma_t = \text{diag}\left(e^{h_{1,t}}, \ldots, e^{h_{n,t}}\right)$
-

$$A_{0t} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ a_{21,t} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1,t} & a_{n2,t} & \cdots & 1 \end{pmatrix}$$

- Rewrite as

$$y_t = X_t \beta + W_t a_t + \epsilon_t$$

- $X_t = I_n \otimes (1, y'_{t-1}, \ldots, y'_{t-p})$
- a_t is vector of free elements of A_{0t}

Vector Autoregressions with Stochastic Volatility

- $$W_t = \begin{pmatrix} 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ -y_{1,t} & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & -y_{1,t} & -y_{2,t} & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & -y_{1,t} & -y_{2,t} & \cdots & -y_{N-1,t} \end{pmatrix}$$
- $$\begin{aligned} h_t &= h_{t-1} + \epsilon_t^h, \quad \epsilon_t^h \sim N(0, \Sigma_h) \\ a_t &= a_{t-1} + \epsilon_t^a, \quad \epsilon_t^a \sim N(0, \Sigma_a) \end{aligned}$$
- $\Sigma_h = \mathrm{diag}(\sigma_{h,1}^2, \ldots, \sigma_{h,N}^2)$ and $\Sigma_a = \mathrm{diag}(\sigma_{a,1}^2, \ldots, \sigma_{a,\frac{N(N-1)}{2}}^2)$.
- Standard MCMC methods used for estimation and forecasting
- But these will not work with large VARs

Composite Bayesian Methods

- Likelihood function (assuming independent errors):

$$L(y;\theta) = \prod_{t=1}^{T} p(y_t|\theta) = \prod_{t=1}^{T} L(y_t;\theta)$$

- Composite likelihood

$$L^C(y;\theta) = \prod_{t=1}^{T}\prod_{i=1}^{M} L^C(y_{i,t};\theta)^{w_i}$$

- $y_{i,t}$ for $i = 1, .., M$ are sub-vectors of y_t
- $L^C(y_{i,t};\theta) = p(y_{i,t}|\theta)$
- w_i weight attached to sub-model i
- $\sum_{i=1}^{M} w_i = 1$
- Bayesian composite posterior

$$p^C(\theta|y) \propto L^C(y;\theta)\, p(\theta)$$

How do we use composite Bayesian methods?

- Instead of forecasting with large VAR-SV, forecast with many small VAR-SVs
- Let $y_t = \begin{pmatrix} y_t^* \\ z_t \end{pmatrix}$
- y_t^* contains N_* variables of interest
- z_t (with elements denoted by $z_{i,t}$) remaining variables.
- Sub-model i is VAR-SV using $y_{i,t} = \begin{pmatrix} y_t^* \\ z_{i,t} \end{pmatrix}$
- Our application uses 193 variables with $N_* = 3$
- Thus, 190 sub-models, each is a 4-variate VAR-SV

Theory of Composite Likelihood Methods

- Some asymptotic theory exists (e.g. Canova and Matthes, 2017)
- Require strong assumptions
- Overview: Varin, Reid and Firth (2011, Stat Sin)
- Pakel, Shephard, Sheppard and Engle (2014, working paper)
- Need asymptotic mixing assumptions about dependence over time, over variables and between different variables at different points in time
- In general, strong assumptions often not achieved in practice
- Hence, our justification is mostly empirical
- For our choice of sub-models, have proved convergence to a large restricted VAR-SV

Theory of Composite Likelihood Methods

- Kullback-Liebler divergence between composite likelihood VAR-SV and a restricted large VAR-SV is $D_{\text{KL}}(L \| \tilde{L}^C)$
- Theorem:
- If $\max\{w_i\}$ is decreasing in M and $\frac{\sqrt{M}}{g(M)} < \infty$ then

$$\lim_{M \to \infty} D_{\text{KL}}(L \| \tilde{L}^C) = 0$$

- Which restricted large VAR-SV? See paper
- What is $g(M)$?
- Shrinkage on some coefficients of same form as commonly done with Minnesota prior
- E.g. if $g(M) = \sqrt{M}$ then "other lags" shrunk more than "own lags"
- If $g(M) = 1$ (no such shrinkage) then $D_{\text{KL}}(L \| \tilde{L}^C)$ is at least bounded

Theory of Composite Likelihoods as Opinion Pools

- Bayesian theory uses idea of opinion pool
- Each sub-model is "agent" with "opinion" about a feature (e.g. a forecast) expressed through a probability distribution.
- Theory addresses "How do we combine these opinions?"
- Generalized logarithmic opinion pool equivalent to composite likelihood
- Nice properties (e.g. external Bayesianity)
- Linear opinion pools lead to other combinations of sub-models
- E.g. Hall and Mitchell (2007) or Geweke and Amisano (2011, JOE) optimal prediction pools
- In empirical work consider both composite likelihood and linear opinion pools

Choosing the Weights

- Various approaches considered
- Equal weights $w_i = \frac{1}{M}$
- Weights proportional to marginal likelihood of each sub-model
- Weights proportional to (exponential of) BIC of each sub-model
- Weights proportional to (exponential of) DIC of each sub-model
- In all above use likelihood/marginal likelihood for core variables only (y_t^*)

Computation

- Target: Draws from Bayesian composite posterior

$$p^C(\theta|y) \propto L^C(y;\theta)\, p(\theta)$$

- We have:
- 1. MCMC draws from M sub-models (4-variate VAR-SVs)
- 2. Weights, w_i for $i = 1, .., M$
- We develop accept-reject algorithm
- Parallelization benefits: Each sub-model run separately
- See paper for details

Macroeconomic Forecasting Using a Large Dataset

- FRED-QD data set from1959Q1- 2015Q3
- 193 quarterly US variables (transformed to stationarity)
- Three core variables: CPI inflation, GDP growth and the Federal Funds rate.
- Small data set: 7 variables
- Core variables + unemployment, industrial production, money (M2) and stock prices (S&P)
- Large data set: All 193 variables
- Lag length of 4

Organization

- With small data set use variety of models
- Computation is feasible (and over-parameterization concerns smaller)
- Large data set:
- Compare composite likelihoods methods to homoskedastic, conjugate prior, large VAR

Priors

- For composite likelihood approach prior elicitation less of an issue (small models)
- With large VARs prior elicitation is crucial (may or may not be disadvantage)
- For all models use comparable priors
- Hyperparameter choices inspired by Minnesota prior
- See paper for details

Models

- Variety of different weights in composite likelihood approaches
- Standard VAR-SV (Primiceri, 2005, ReStud)
- Homoskedastic VARs of different dimensions
- Carriero, Clark and Marcellino (CCM, 2016a,b)
- CCM1: common drifting volatility model
- VAR-SV with $a_t = 0$ and $\Sigma_t = e^{h_t}\Sigma$
- h_t is scalar stochastic volatility process
- CCM2: more flexible SV model
- VAR-SV with a_t constant
- Each equation error has own volatility, but restrictions on correlations

Table 1: Models used in Forecasting Exercise

VAR-HM	7-variable Homoskedastic VAR
VAR-SV	7-variable VAR with stochastic volatility
VAR-CCM1	7-variable model of CCM (2016a)
VAR-CCM2	7-variable model of CCM (2016b)
Large VAR	Large Homoskedastic VAR
VAR-CL-BIC	VAR-CL-SV with BIC based weights
VAR-CL-DIC	VAR-CL-SV with DIC based weights
VAR-CL-EQ	VAR-CL-SV with equal weights
VAR-CL-ML	VAR–SV with ML weights
VAR-CL-LIN	VAR–CL-SV with linear pool weights
VAR-LIN	VAR-SV with linear pool weights

Estimating Variances and Covariances

- Key variables of interest (common to all models) are $\sigma_{ij,t}$ for $i, j = 1, 2, 3$
- Small data set: VAR-SV will probably be closest to "true" specification (most flexible)
- Evaluate performance relative to VAR-SV
- In paper we show VAR-CL and VAR-LIN methods and VAR-CCM2 track VAR-SV fairly well
- VAR-CCM1 and VAR-HM

Forecasting

- Estimation results are encouraging, what about forecasting?
- Results for $h = 1$ and 4
- Two forecast evaluation periods:
- Beginning 1970Q1
- Beginning 2008Q1 (financial crisis and subsequent recession)
- I will (mostly) focus on results which begin in 2008Q1

Forecast Evaluation Metrics

- For 3 core variables individually:
- RMSFE
- MAFE
- ALPL $=$ average of log predictive likelihoods (higher value better)
- ACRPS $=$ average of conditional rank probability score (lower values better)
- Also joint ALPL based on joint predictive for core variables

Joint ALPL for Core Variables

Table 2: Forecasting Evaluation Using Joint ALPL for 3 Core Variables				
Horizon	$h = 1$		$h = 4$	
Evaluation begins:	1970Q1	2008Q1	1970Q1	2008Q1
VAR-HM	0.33	−0.58	−1.04	−1.60
VAR-SV	0.65	0.44	−1.04	−1.61
VAR-CCM1	0.06	−0.51	−0.98	−1.85
VAR-CCM2	0.90	0.52	−0.84	−1.58
Large VAR	−0.47	−1.69	−1.41	−2.02
VAR-CL-ML	0.90	**1.27**	−0.99	−1.49
VAR-CL-DIC	0.85	0.67	−0.72	−0.92
VAR-CL-BIC	0.90	1.15	−0.88	−1.51
VAR-CL-EQ	0.88	0.89	−0.71	−0.84
VAR-CL-LIN	0.89	0.92	**−0.71**	**−0.79**
VAR-LIN	**0.91**	1.01	−0.75	−0.83

Joint ALPL for Core Variables

- Best overall summary
- Composite likelihoods + linear opinion pool forecast best
- Weights: Marginal likelihood or BIC weights best (but only slightly)
- Homoskedastic large VAR does poorly
- CCM2 better than CCM1

Forecasting the Core Forecasts Individually

- See paper for detailed results
- General themes:
- Methods pooling many small sub-models forecast well
- Especially for 2008-2016 period
- Especially for inflation and interest rate
- Less so for GDP growth (VAR-SV is best)
- Large homoskedastic VARs forecast poorly
- CCM2 better than CCM1
- In general, CCM2 similar but a bit worse than composite likelihoods

Conclusion

- Composite likelihood methods allows VAR-SV with huge data sets
- Computationally and conceptually simple: average over many small models
- Other VAR-SV models have attractive features but are computationally infeasible with huge data sets
- In small data set, composite likelihood methods approximate other methods
- In large data set, composite likelihoods forecast better than large VAR