Bayesian Vector Autoregressive Models

Introduction

- There are many popular time series models and all cannot be covered in a short course.
- In this course, will focus on models popular with empirical macroeconomists, characterized by:
- i) Multivariate in nature (macroeconomists interested in relationships between variables, not properties of a single variable).
- ii) Allow for parameters to change (e.g. over time, across business cycle, etc.).
- We will not cover univariate time series nor nonlinear time series models such as Markov switching, TAR, STAR, etc.
- See Bayesian Econometric Methods Chapters 17 and 18 for treatment of some of these models.
- We will discuss state space models (which can be used to model nonlinearities).

Time Series Modelling for Empirical Macroeconomics

- Vector Autoregressive (VAR) models popular way of summarizing inter-relationships between macroeconomic variables.
- Used for forecasting, impulse response analysis, etc.
- Economy is changing over time. Is model in 1970s same as now?
- Thus, time-varying parameter VARs (TVP-VARs) are of interest.
- Great Moderation of business cycle leads to interest in modelling error variances
- TVP-VARs with multivariate stochastic volatility is our end goal.
- Begin with Bayesian VARs
- A common theme: These models are over-parameterized so need shrinkage to get reasonable results (shrinkage = prior).

Bayesian VARs

- One way of writing VAR(p) model:

$$y_t = a_0 + \sum_{j=1}^{p} A_j y_{t-j} + \varepsilon_t$$

- y_t is $M \times 1$ vector
- ε_t is $M \times 1$ vector of errors
- a_0 is $M \times 1$ vector of intercepts
- A_j is an $M \times M$ matrix of coefficients.
- ε_t is i.i.d. $N(0, \Sigma)$.
- Exogenous variables or more deterministic terms can be added (but we don't to keep notation simple).

- Several alternative ways of writing the VAR (and we will use some alternatives below).
- One way: let y be $MT \times 1$ vector ($y = (y_1', .., y_T')$) and ε stacked conformably
- $x_t = \left(1, y_{t-1}', .., y_{t-p}'\right)$

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_T \end{bmatrix}$$

- $K = 1 + Mp$ is number of coefficients in each equation of VAR and X is a $T \times K$ matrix.
- The VAR can be written as:

$$y = \left(I_M \otimes X\right) \alpha + \varepsilon$$

- $\varepsilon \sim N\left(0, \Sigma \otimes I_M\right)$.

- Another way of writing VAR:
- Let Y and E be $T \times M$ matrices placing the T observations on each variable in columns next to one another.
- Then can write VAR as

$$Y = XA + E$$

- In first VAR, α is $KM \times 1$ vector of VAR coefficients, here A is $K \times M$
- Relationship between two: $\alpha = vec(A)$
- We will use both notations below (and later on, when working with restricted VAR need to introduce yet more notation).

Likelihood Function

- Likelihood function can be derived and shown to be of a form that breaks into two parts (see Bayesian Econometric Methods Exercise 17.6)
- First of these parts α given Σ and another for Σ
- $$\alpha|\Sigma, y \sim N\left(\widehat{\alpha}, \Sigma \otimes \left(X'X\right)^{-1}\right)$$
- Σ^{-1} has Wishart form

$$\Sigma^{-1}|y \sim W\left(S^{-1}, T-K-M-1\right)$$

- where $\widehat{A} = \left(X'X\right)^{-1}X'Y$ is OLS estimate of A, $\widehat{\alpha} = vec\left(\widehat{A}\right)$ and

$$S = \left(Y - X\widehat{A}\right)'\left(Y - X\widehat{A}\right)$$

Digression

- Remember regression models had parameters β and σ^2
- There proved convenient to work with $h = \frac{1}{\sigma^2}$
- In VAR proves convenient to work with Σ^{-1}
- In regression h typically had Gamma distribution
- With VAR Σ^{-1} will typically have Wishart distribution
- Wishart is matrix generalization of Gamma
- Details see appendix to textbook.
- If Σ^{-1} is $W(C, c)$ then "Mean" is cC and c is degrees of freedom.
- Note: easy to take random draws from Wishart.

Prior Issue 1

- VARs are not parsimonious models: α contains KM parameters
- For a VAR(4) involving 5 dependent variables: 105 parameters.
- Macro data sets: number of observations on each variable might be a few hundred.
- Without prior information, hard to obtain precise estimates.
- Features such as impulse responses and forecasts will tend to be imprecisely estimated.
- Desirable to "shrink" forecasts and prior information offers a sensible way of doing this shrinkage.
- Different priors do shrinkage in different ways.

Prior Issue 2

- Some priors lead to analytical results for the posterior and predictive densities.
- Other priors require MCMC methods (which raise computational burden).
- E.g. recursive forecasting exercise typically requires repeated calculation of posterior and predictive distributions
- In this case, MCMC methods can be very computationally demanding.
- May want to go with not-so-good prior which leads to analytical results, if ideal prior leads to slow computation.

Prior Issue 3

- Priors differ in how easily they can handle extensions of the VAR defined above.
- Restricted VARs: different equations have different explanatory variables.
- TVP-VARs: Allowing for VAR coefficients to change over time.
- Heteroskedasticity
- Such extensions typically require MCMC, so no need to restrict consideration to priors which lead to analytical results in basic VAR

The Minnesota Prior

- The classic shrinkage priors developed by researchers (Litterman, Sims, etc.) at the University of Minnesota and the Federal Reserve Bank of Minneapolis.
- They use an approximation which simplifies prior elicitation and computation: replace Σ with an estimate, $\widehat{\Sigma}$.
- Original Minnesota prior simplifies even further by assuming Σ to be a diagonal matrix with $\widehat{\sigma}_{ii} = s_i^2$
- s_i^2 is OLS estimate of the error variance in the i^{th} equation
- If Σ not diagonal, can use, e.g., $\widehat{\Sigma} = \frac{S}{T}$.

- Minnesota prior assumes

$$\alpha \sim N\left(\underline{\alpha}_{Min}, \underline{V}_{Min}\right)$$

- Minnesota prior is way of automatically choosing $\underline{\alpha}_{Min}$ and $\underline{V}_{Min}$
- Note: explanatory variables in any equation can be divided as:
- own lags of the dependent variable
- the lags of the other dependent variables
- exogenous or deterministic variables

- $\underline{\alpha}_{Min} = 0$ implies shrinkage towards zero (a nice way of avoiding over-fitting).
- When working with differenced data (e.g. GDP growth), Minnesota prior sets $\underline{\alpha}_{Min} = 0$
- When working with levels data (e.g. GDP growth) Minnesota prior sets element of $\underline{\alpha}_{Min}$ for first own lag of the dependent variable to 1.
- Idea: Centred over a random walk. Shrunk towards random walk (specification which often forecasts quite well)
- Other values of $\underline{\alpha}_{Min}$ also used, depending on application.

- Prior mean: "towards what should we shrink?"
- Prior variance: "by how much should we shrink?"
- Minnesota prior: $\underline{V}_{Min}$ is diagonal.
- Let $\underline{V}_i$ denote block of $\underline{V}_{Min}$ for coefficients in equation i
- $\underline{V}_{i,jj}$ are diagonal elements of $\underline{V}_i$
- A common implementation of Minnesota prior (for $r = 1, .., p$ lags):

$$\underline{V}_{i,jj} = \begin{cases} \frac{\underline{a}_1}{r^2} \text{ for coefficients on own lags} \\ \frac{\underline{a}_2 \sigma_{ii}}{r^2 \sigma_{jj}} \text{ for coefficients on lags of variable } j \neq i \\ \underline{a}_3 \sigma_{ii} \text{ for coefficients on exogenous variables} \end{cases}$$

- Typically, $\sigma_{ii} = s_i^2$.

- Problem of choosing $\frac{KM(KM+1)}{2}$ elements of $\underline{V}_{Min}$ reduced to simply choosing , $\underline{a}_1, \underline{a}_2, \underline{a}_3$.
- Property: as lag length increases, coefficients are increasingly shrunk towards zero
- Property: by setting $\underline{a}_1 > \underline{a}_2$ own lags are more likely to be important than lags of other variables.
- See Litterman (1986) for motivation and discussion of these choices (e.g. explanation for how $\frac{\sigma_{ii}}{\sigma_{jj}}$ adjusts for differences in the units that the variables are measured in).
- Minnesota prior seems to work well in practice.
- Recent paper by Giannone, Lenza and Primiceri (in ReStat) develops methods for estimating prior hyperparameters from the data

Posterior Inference with Minnesota Prior

- Simple analytical results involving only the Normal distribution.
- $$\alpha|y \sim N\left(\overline{\alpha}_{Min}, \overline{V}_{Min}\right)$$
- $$\overline{V}_{Min} = \left[\underline{V}_{Min}^{-1} + \left(\widehat{\Sigma}^{-1} \otimes \left(X'X\right)\right)\right]^{-1}$$
- $$\overline{\alpha}_{Min} = \overline{V}_{Min}\left[\underline{V}_{Min}^{-1}\underline{\alpha}_{Min} + \left(\widehat{\Sigma}^{-1} \otimes X\right)' y\right]$$

Natural conjugate prior

- A drawback of Minnesota prior is its treatment of Σ.
- Ideally want to treat Σ as unknown parameter
- Natural conjugate prior allows us to do this in a way that yields analytical results.
- But (as we shall sell) has some drawbacks.
- In practice, noninformative limiting version of natural conjugate prior sometimes used (but noninformative prior does not do shrinkage)

- An examination of likelihood function (see also similar derivations for Normal linear regression model where Normal-Gamma prior was natural conjugate) suggests VAR natural conjugate prior:

$$\alpha|\Sigma \sim N\left(\underline{\alpha}, \Sigma \otimes \underline{V}\right)$$

- $$\Sigma^{-1} \sim W\left(\underline{S}^{-1}, \underline{\nu}\right)$$

- $\underline{\alpha}, \underline{V}, \underline{\nu}$ and $\underline{S}$ are prior hyperparameters chosen by the researcher.
- Noninformative prior: $\underline{\nu} = 0$ and $\underline{S} = \underline{V}^{-1} = cI$ and let $c \longrightarrow 0$.

Posterior when using natural conjugate prior

- Posterior has analytical form:

$$\alpha|\Sigma, y \sim N\left(\overline{\alpha}, \Sigma \otimes \overline{V}\right)$$

- $$\Sigma^{-1}|y \sim W\left(\overline{S}^{-1}, \overline{\nu}\right)$$

- where

$$\overline{V} = \left[\underline{V}^{-1} + X'X\right]^{-1}$$

- $$\overline{A} = \overline{V}\left[\underline{V}^{-1}\underline{A} + X'X\widehat{A}\right]$$

- $$\overline{S} = S + \underline{S} + \widehat{A}'X'X\widehat{A} + \underline{A}'\underline{V}^{-1}\underline{A} - \overline{A}'\left(\underline{V}^{-1} + X'X\right)\overline{A}$$

- $$\overline{\nu} = T + \underline{\nu}$$

- Remember: in regression model joint posterior for (β, h) was Normal-Gamma, but marginal posterior for β had t-distribution
- Same thing happens with VAR coefficients.
- Marginal posterior for α is a multivariate t-distribution.
- Posterior mean is $\overline{\alpha}$
- Degrees of freedom parameter is $\overline{\nu}$
- Posterior covariance matrix:

$$var\left(\alpha|y\right) = \frac{1}{\overline{\nu} - M - 1}\overline{S} \otimes \overline{V}$$

- Posterior inference can be done using (analytical) properties of t-distribution.
- Predictive inference can also be done analytically (for one-step ahead forecasts)

Problems with Natural Conjugate Prior

- Natural conjugate prior has great advantage of analytical results, but has some problems which make it rarely used in practice.
- To make problems concrete consider a macro example:
- The VAR involves variables such as output growth and the growth in the money supply
- Researcher wants to impose the neutrality of money.
- Implies: coefficients on the lagged money growth variables in the output growth equation are zero (but coefficients of lagged money growth in other equations would not be zero).

- Problem 1: Cannot simply impose neutrality of money restriction.
- The $(I_M \otimes X)$ form of the explanatory variables in VAR means every equation must have same set of explanatory variables.
- But if we do not maintain $(I_M \otimes X)$ form, don't get analytical conjugate prior (see Kadiyala and Karlsson, JAE, 1997 for details).

- Problem 2: Cannot "almost impose" neutrality of money restriction through the prior.
- Cannot set prior mean over neutrality of money restriction and set prior variance to very small value.
- To see why, let individual elements of Σ be σ_{ij}.
- Prior covariance matrix has form $\Sigma \otimes \underline{V}$
- This implies prior covariance of coefficients in equation i is $\sigma_{ii}\underline{V}$.
- Thus prior covariance of the coefficients in any two equations must be proportional to one another.
- So can "almost impose" coefficients on lagged money growth to be zero in ALL equations, but cannot do it in a single equation.
- Note also that Minnesota prior form $\underline{V}_{Min}$ is not consistent with natural conjugate prior.

Some interesting approaches I will not discuss

- Choosing prior hyperparameters by using dummy observations (fictitious prior data set), see Sims and Zha (1998, IER).
- Using prior information from macro theory (e.g. DSGE models), see Ingram and Whiteman (1994, JME) and Del Negro and Schorfheide (2004, IER).
- Villani (2009, JAE): priors about means of dependent variables
- Useful since researchers often have prior information on these.
- Write VAR as:

$$\widetilde{A}(L)(y_t - \widetilde{a}_0) = \varepsilon_t$$

- $\widetilde{A}(L) = I - \widetilde{A}_1 L - .. - \widetilde{A}_p L^p$, L is the lag operator
- $\widetilde{a}_0$ are unconditional means of the dependent variables.
- Gibbs sampling required.

A Macroeconomic Example

- Hybrid New Keynesian Phillips Curve (NKPC) Model
- Inflation (π_t) and y_t is output gap or unemployment rate

$$\pi_t = \beta_b \pi_{t-1} + \beta_f E_{t-1}(\pi_{t+1}) + \gamma y_t + \varepsilon_t.$$

- $E_{t-1}(\pi_{t+1})$ is expectation at $t-1$ of inflation at $t+1$
- Note: Adding equation for y_t will give a multivariate model.
- No feedback from π_t to y_t

Relating the NKPC to a VAR

- To take NKPC to data need to find rational expectations solution to get rid of $E_{t-1}(\pi_{t+1})$ term in NKPC
- Since no feedback from π_t to y_t can show solution is:

$$\pi_t = a_1 \pi_{t-1} + a_2 y_{t-1} + u_t$$

- where $a_1 = f_1(\beta_b, \beta_f)$ and $a_2 = f_2(\beta_b, \beta_f, \gamma, \rho)$ for functions f_1 and f_2
- Case 1: suppose y_t is

$$y_t = \rho y_{t-1} + v_t$$

- These 2 equations form a restricted VAR (reduced form model)
- Rational expectations macro models often lead to restricted VARs

- Problem: VAR has 3 parameters, a_1, a_2, and ρ, but structural model has 4 (β_f, β_b, γ, and ρ)
- Identification issues in rational expectations/DSGE models can be important.
- Case 2: suppose y_t is

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + v_t$$

- Identification problem is now solved since reduced form VAR now has 4 parameters a_1, a_2, ρ_1 and ρ_2
- But is this solution a good one? Identification depends on lag length. What if ρ_2 is near zero?
- Summary: macro theory can often lead to restricted VARs, but identification can be a worry

The Independent Normal-Wishart Prior

- Natural conjugate prior had $\alpha|\Sigma$ being Normal and Σ^{-1} being Wishart and VAR had same explanatory variables in every equation.
- Want more general setup without these restrictive features.
- Can do this with a prior for VAR coefficients and Σ^{-1} being independent (hence name "independent Normal-Wishart prior")
- And using a more general formulation for the VAR

- To allow for different equations in the VAR to have different explanatory variables, modify notation.
- To avoid, use "β" notation for VAR coefficients now instead of α.
- Each equation (for $m = 1, .., M$) of the VAR is:

$$y_{mt} = z'_{mt}\beta_m + \varepsilon_{mt},$$

- If we set $z_{mt} = \left(1, y'_{t-1}, .., y'_{t-p}\right)'$ for $m = 1, .., M$ then exactly same VAR as before.
- However, here z_{mt} can contain different lags of dependent variables, exogenous variables or deterministic terms.

- Vector/matrix notation:
- $y_t = (y_{1t}, .., y_{Mt})'$, $\varepsilon_t = (\varepsilon_{1t}, .., \varepsilon_{Mt})'$
- $$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_M \end{pmatrix}$$
- $$Z_t = \begin{pmatrix} z_{1t}' & 0 & \cdots & 0 \\ 0 & z_{2t}' & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & z_{Mt}' \end{pmatrix}$$
- β is $k \times 1$ vector, Z_t is $M \times k$ where $k = \sum_{j=1}^{M} k_j$.
- ε_t is i.i.d. $N(0, \Sigma)$.
- Can write VAR as:

$$y_t = Z_t \beta + \varepsilon_t$$

- Stacking:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix}$$

- $$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{pmatrix}$$

- $$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_T \end{pmatrix}$$

- VAR can be written as:

$$y = Z\beta + \varepsilon$$

- ε is $N(0, I \otimes \Sigma)$.

- Thus, VAR can be written as a Normal linear regression model with error covariance matrix of a particular form (SUR form).
- Independent Normal-Wishart prior:

$$p\left(\beta, \Sigma^{-1}\right) = p\left(\beta\right) p\left(\Sigma^{-1}\right)$$

- where

$$\beta \sim N\left(\underline{\beta}, \underline{V}_{\beta}\right)$$

- and

$$\Sigma^{-1} \sim W\left(\underline{S}^{-1}, \underline{\nu}\right)$$

- $\underline{V}_{\beta}$ can be anything the researcher chooses (not restrictive $\Sigma \otimes \underline{V}$ form of the natural conjugate prior).
- $\underline{\beta}$ and $\underline{V}_{\beta}$ could be set as in the Minnesota prior.
- A noninformative prior obtained by setting $\underline{\nu} = \underline{S} = \underline{V}_{\beta}^{-1} = 0$.

Posterior inference in the VAR with independent Normal-Wishart prior

- $p\left(\beta, \Sigma^{-1}|y\right)$ does not have a convenient form allowing for analytical results.
- But Gibbs sampler can be set up.
- Conditional posterior distributions $p\left(\beta|y, \Sigma^{-1}\right)$ and $p\left(\Sigma^{-1}|y, \beta\right)$ do have convenient forms
-
$$\beta|y, \Sigma^{-1} \sim N\left(\overline{\beta}, \overline{V}_{\beta}\right)$$
- where
$$\overline{V}_{\beta} = \left(\underline{V}_{\beta}^{-1} + \sum_{t=1}^{T} Z_t' \Sigma^{-1} Z_t\right)^{-1}$$
- and
$$\overline{\beta} = \overline{V}_{\beta}\left(\underline{V}_{\beta}^{-1}\underline{\beta} + \sum_{i=1}^{T} Z_t' \Sigma^{-1} y_t\right)$$

- $$\Sigma^{-1}|y,\beta \sim W\left(\overline{S}^{-1},\overline{\nu},\right)$$
- where
$$\overline{\nu} = T + \underline{\nu}$$
- $$\overline{S} = \underline{S} + \sum_{t=1}^{T}\left(y_t - Z_t\beta\right)\left(y_t - Z_t\beta\right)'$$
- Remember: for any Gibbs sampler, the resulting draws can be used to calculate posterior properties of any function of the parameters (e.g. impulse responses), marginal likelihoods (for model comparison) and/or to do prediction.

Prediction in VARs

- I will use prediction and forecasting to mean the same thing
- Goal predict y_τ for some period τ using data available at time $\tau - 1$
- For the VAR, Z_τ contains information dated $\tau - 1$ or earlier.
- For predicting at time τ given information through $\tau - 1$, can use:

$$y_\tau | Z_\tau, \beta, \Sigma \sim N\left(Z_t \beta, \Sigma\right)$$

- This result and Gibbs draws $\beta^{(s)}, \Sigma^{(s)}$ for $s = 1, .., S$ allows for predictive inference.
- E.g. predictive mean (a popular point forecast) could be obtained as:

$$E\left(y_\tau | Z_\tau\right) = \frac{\sum_{s=1}^{S} Z_t \beta^{(s)}}{S}$$

- Other predictive moments can be calculated in a similar fashion

Prediction in VARs

- Or can do predictive simulation:
- For each Gibbs draw $\beta^{(s)}, \Sigma^{(s)}$ simulate one (or more) $y_\tau^{(s)}$
- Result will be $y_\tau^{(s)}$ for $s = 1, .., S$ draws
- Plot them to produce predictive density
- Average them to produce predictive mean
- Take their standard deviation to produce predictive standard deviation
- etc.

Prediction in VARs

- Preceding material was about predicting y_τ using data available at time $\tau - 1$
- This is one-period ahead forecasting
- But what about h-period ahead forecast
- h is the forecast horizon
- E.g. with quarterly data forecasting a year ahead $h = 4$
- Can do direct or iterated forecasting

Direct Forecasting in VARs

- Direct forecasting is straightforward: simply redefine Z_τ
- Above defined each equation using $z_{m\tau} = \left(1, y'_{\tau-1}, .., y'_{\tau-p}\right)'$
- Replace this by $z_{m\tau} = \left(1, y'_{\tau-h}, .., y'_{\tau-p-h+1}\right)'$
- Then your model is always predicting y_τ using data available at time $\tau - h$
- All posterior and predictive formulae are as above
- If forecasting (e.g.) for $h = 1, 2, 3, 4$ must re-estimate model for each h

Iterated Forecasting in VARs

- Estimate the model once using $z_{m\tau} = \left(1, y'_{\tau-1}, .., y'_{\tau-p}\right)'$
- Remember result that

$$y_\tau | Z_\tau, \beta, \Sigma \sim N\left(Z_t\beta, \Sigma\right) \qquad (**)$$

- When forecasting y_τ using information available at time $\tau - h$ for $h > 1$ you face a problem using (**)
- Use $h = 2$ and $p = 2$ to illustrate
- In the model, y_τ depends on $y_{\tau-1}$ and $y_{\tau-2}$
- But as a forecaster, you do not know $y_{\tau-1}$ yet
- E.g. suppose you have data through 2015Q4
- When forecasting 2016Q1 ($h = 1$) will have data for 2015Q4 and 2015Q3
- So Z_t is known for $h = 1$
- But when forecasting 2016Q2 ($h = 2$) will not have data for 2016Q1 and not know Z_t

Iterated Forecasting in VARs

- Solution to problem:
- Do predictive simulation beginning with $h = 1$
- Use draw of $y_{\tau-1}^{(s)}$ (along with y_{t-2}, $\beta^{(s)}, \Sigma^{(s)}$) to plug into (**)
- This is called iteration
- For $h > 2$ just keep on iterating
- Strategy above will provide you with draws $y_{\tau-1}^{(s)}$ and $y_{\tau-2}^{(s)}$
- For $h = 3$ can use these to define appropriate Z_t for use in (**)
- etc.
- Which of iterated or direct forecasting is better?
- This seems to depend on the data set being used

Recursive and Rolling Forecasts

- Data runs from $t = 1, .., T$
- E.g. annual data set from 1960 through 2015
- Sometimes researcher is interested in out-of-sample forecasting:
- Forecasting 2016 (or 2017, 2018, etc.)
- 2016 is not yet observed = out of sample
- Sometimes researcher wants to know how well model might have forecast in past
- E.g. given data I had in 1995 how well would I have forecast 1996?
- In general, given data available at time $\tau - h$, how well would I forecast τ?
- Pseudo out-of-sample forecast evaluation

Recursive and Rolling Forecasts

- For pseudo out–of-sample forecast evaluation proceed as follows:
- choose a forecast evaluation period: $\tau = \tau_0, .., T$
- E.g. 1970 to 2015
- Note $\tau_0 > 1$ since you need at least some data to sensibly estimate the VAR
- Recursive forecasting involves:
- use data for $t = 1, .., \tau - h$ to forecast y_τ
- Repeat for $\tau = \tau_0, .., T$
- Note: can be computationally demanding (esp. if MCMC and predictive simulation used)
- Repeatedly estimate model on "expanding window" of data

Recursive and Rolling Forecasts

- Recursive forecasting uses all data available at $\tau - h$ to forecast
- But what if parameter change has occurred (e.g. 1960s data irrelevant for 1990s forecasting)?
- E.g. Recursive forecasts in 1990s will be contaminated with 1960s data
- Best solution: build parameter/regime change into your model (more in future on this)
- Rolling forecasts: same as recursive forecasts but use data from $t = \tau - h - \tau_1, .., \tau - h$ to estimate VAR for forecasting y_τ
- Fixed window of data (always use most recent τ_1 observations)

Evaluating Forecasts

- Suppose you have produced forecasts somehow (direct or iterated/recursive or rolling) for $\tau = \tau_0, .., T$ and have
- Predictive densities $p\left(y_\tau | Z_t\right)$
- Predictive means (point forecasts): $E\left(y_\tau | Z_t\right)$
- Note: in past point forecasts popular, now huge interest in uncertainty about future (e.g. Bank of England inflation fan charts)
- Predictive densities (or density forecasts) hot topic
- Usually will have forecasts from several models (e.g. comparing VAR to other modelling approaches)
- How do you decide whether your forecasts are good?
- Large literature exists on forecast evaluation
- Necessary to distinguish between random variable and its realization
- E.g. if y_{it} is random variable and y_{it}^R is the observed value (e.g. observed inflation in 2015)
- Here I will define two common approaches

Mean Squared Forecast Error (MSFE)

- MSFE is the most common way of measuring performance of point forecasts for a variable in the VAR (e.g. y_{it} = inflation)
- $$MSFE = \frac{\sum_{\tau=\tau_0}^{T} \left(y_{it}^{R} - E\left(y_{i\tau}|Z_t\right)\right)^2}{T - \tau_0 + 1}$$
- Many related variants such as Mean Absolute Forecast Error (MAFE):

$$MAFE = \frac{\sum_{\tau=\tau_0}^{T} \left|y_{it}^{R} - E\left(y_{i\tau}|Z_t\right)\right|}{T - \tau_0 + 1}$$

Predictive Likelihoods

- Most common way of evaluating performance of entire predictive density is with predictive likelihood
- Predictive likelihood is predictive density evaluated at the actual realization
- Predictive likelihood for variable i at time τ: $p\left(y_{i\tau} = y_{i\tau}^{R} | Z_{\tau}\right)$
- Common to present cumulative sums of log predictive likelihoods as measure of forecast performance:

$$\sum_{\tau=\tau_0}^{T} \log\left[p\left(y_{i\tau} = y_{i\tau}^{R} | Z_{\tau}\right)\right]$$

- Can show if $\tau_0 = 1$, cumulative sums of predictive likelihoods equal to log marginal likelihood
- Have interpretation similar to marginal likelihoods over forecast evaluation period

Stochastic Search Variable Selection (SSVS) in VARs

- There are many approaches which seek parsimony/shrinkage in VARs, take SSVS as a representative example
- SSVS is usually done in VAR where every equation has same explanatory variables
- Hence, return to our initial notation for VARs where X contains lagged dependent variable, α are VAR coefficients, etc.
- SSVS can be interpreted as a prior shrinks some VAR coefficients to zero
- Or as a model selection device (select the model with explanatory variables with non-zero coefficients)
- Or as a model averaging device (which averages over models with different non-zero coefficients).
- Can be implemented in various ways, here we follow George, Sun and Ni (2008, JoE)

- Remember: of basic idea for a VAR coefficient, α_j
- SSVS is hierarchical prior, mixture of two Normal distributions:

$$\alpha_j|\gamma_j \sim (1-\gamma_j)\, N\left(0, \kappa_{0j}^2\right) + \gamma_j N\left(0, \kappa_{1j}^2\right)$$

- γ_j is dummy variable.
- $\gamma_j = 1$ then α_j has prior $N\left(0, \kappa_{1j}^2\right)$
- $\gamma_j = 0$ then α_j has prior $N\left(0, \kappa_{0j}^2\right)$
- Prior is hierarchical since γ_j is unknown parameter and estimated in a data-based fashion.
- κ_{0j}^2 is "small" (so coefficient is shrunk to be virtually zero)
- κ_{1j}^2 is "large" (implying a relatively noninformative prior for α_j).

- Below we describe a Gibbs sampler for this model which provides draws of γ and other parameters
- SSVS can select a single restricted model.
- Run Gibbs sampler and calculate Pr $\left(\gamma_j = 1|y\right)$ for $j = 1, .., KM$
- Set to zero all coefficients with Pr $\left(\gamma_j = 1|y\right) < a$ (e.g. $a = 0.5$).
- Then re-run Gibbs sampler using this restricted model
- Alternatively, if the Gibbs sampler for unrestricted VAR is used to produce posterior results for the VAR coefficients, result will be Bayesian model averaging (BMA).

Gibbs Sampling with the SSVS Prior

- SSVS prior for VAR coefficients, α, can be written as:

$$\alpha|\gamma \sim N(0, DD)$$

- γ is a vector with elements $\gamma_j \in \{0, 1\}$,
- D is diagonal matrix with $(j,j)^{th}$ element d_j:

$$d_j = \begin{cases} \kappa_{0j} \text{ if } \gamma_j = 0 \\ \kappa_{1j} \text{ if } \gamma_j = 1 \end{cases}$$

- "default semi-automatic approach" to selecting κ_{0j} and κ_{1j}
- Set $\kappa_{0j} = c_0 \sqrt{\widehat{var}(\alpha_j)}$ and $\kappa_{1j} = c_1 \sqrt{\widehat{var}(\alpha_j)}$
- $\widehat{var}(\alpha_j)$ is estimate from an unrestricted VAR
- E.g. OLS or a preliminary Bayesian estimate from a VAR with noninformative prior
- Constants c_0 and c_1 must have $c_0 \ll c_1$ (e.g. $c_0 = 0.1$ and $c_1 = 10$).

- We need prior for γ and a simple one is:

$$\Pr\left(\gamma_j = 1\right) = \underline{q}_j$$
$$\Pr\left(\gamma_j = 0\right) = 1 - \underline{q}_j$$

- $\underline{q}_j = \frac{1}{2}$ for all j implies each coefficient is *a priori* equally likely to be included as excluded.
- Can use same Wishart prior for Σ^{-1}
- Note: George, Sun and Ni also show how to do SSVS on off-diagonal elements of Σ

- Gibbs sampler sequentially draws from $p\left(\alpha|y,\gamma,\Sigma\right)$, $p\left(\gamma|y,\alpha,\Sigma\right)$ and $p\left(\Sigma^{-1}|y,\gamma,\alpha\right)$
- $$\alpha|y,\gamma,\Sigma \sim N(\overline{\alpha}_{\alpha},\overline{V}_{\alpha})$$
- where
$$\overline{V}_{\alpha} = [\Sigma^{-1} \otimes (X'X) + (DD)^{-1}]^{-1}$$
- $$\overline{\alpha}_{\alpha} = \overline{V}_{\alpha}[(\Psi\Psi') \otimes (X'X)\hat{\alpha}]$$
- $$\hat{A} = (X'X)^{-1}X'Y$$
- $$\hat{\alpha} = vec(\hat{A})$$

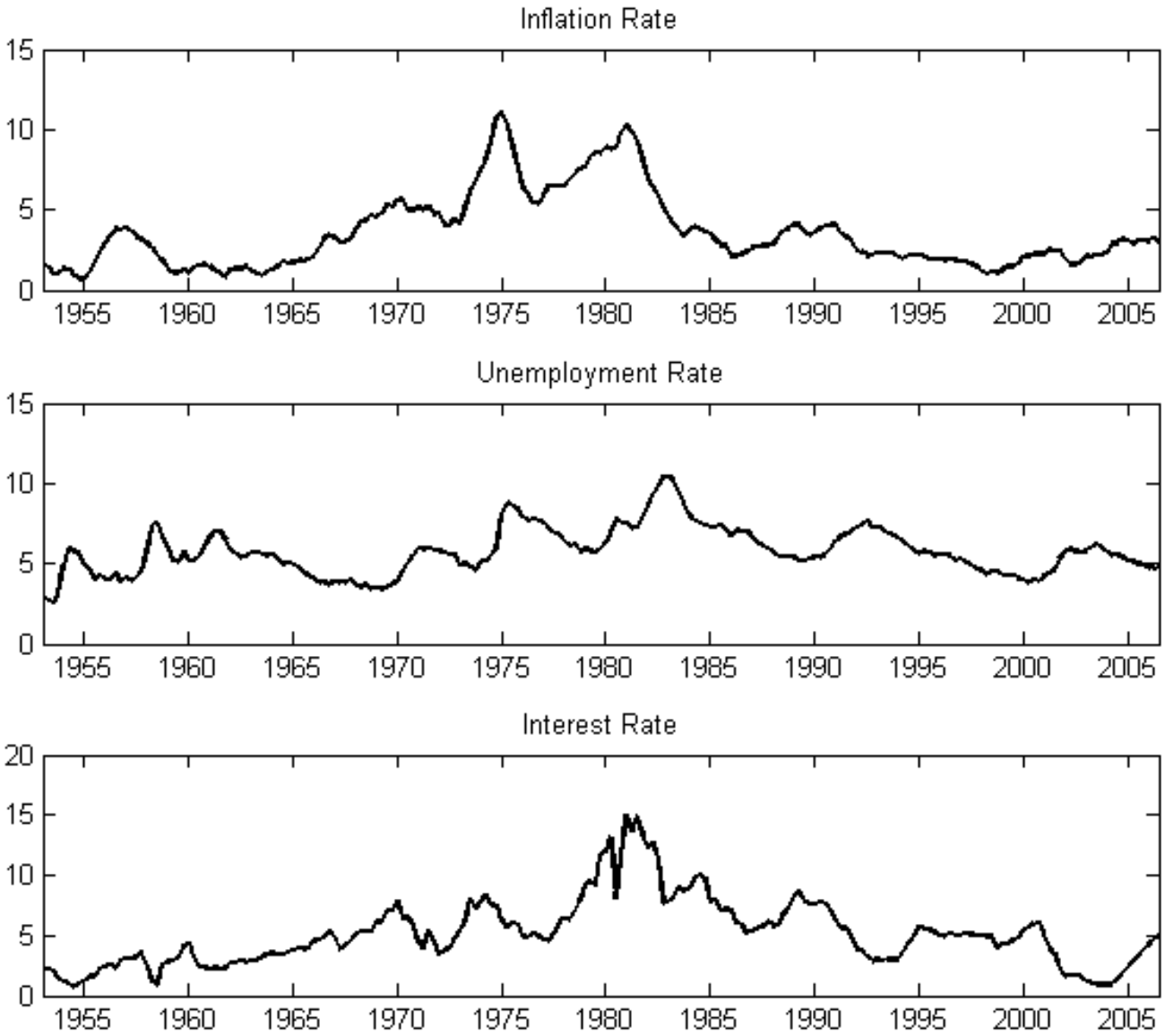
- $p\left(\gamma|y,\alpha,\Sigma\right)$ has γ_j being independent Bernoulli random variables:
- $$\begin{aligned}\Pr\left(\gamma_j=1|y,\alpha,\Sigma\right)&=\overline{q}_j\\ \Pr\left(\gamma_j=0|y,\alpha,\Sigma\right)&=1-\overline{q}_j\end{aligned}$$
- where

$$\overline{q}_j=\frac{\frac{1}{\kappa_{1j}}\exp\left(-\frac{\alpha_j^2}{2\kappa_{1j}^2}\right)\underline{q}_j}{\frac{1}{\kappa_{1j}}\exp\left(-\frac{\alpha_j^2}{2\kappa_{1j}^2}\right)\underline{q}_j+\frac{1}{\kappa_{0j}}\exp\left(-\frac{\alpha_j^2}{2\kappa_{0j}^2}\right)\left(1-\underline{q}_j\right)}$$

- $p\left(\Sigma^{-1}|y,\gamma,\alpha\right)$ has similar Wishart form as previously, so I will not repeat here.

Illustration of Bayesian VAR Methods in a Small VAR

- Data set: standard quarterly US data set from 1953Q1 to 2006Q3.
- Inflation rate $\Delta\pi_t$, the unemployment rate u_t and the interest rate r_t
- $y_t = (\Delta\pi_t, u_t, r_t)'$.
- These three variables are commonly used in New Keynesian VARs.
- The data are plotted in Figure 1.
- We use unrestricted VAR with intercept and 4 lags


Inflation Rate
15
10
5
0
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
Unemployment Rate
15
10
5
0
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
Interest Rate
20
15
10
5
0
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005

- We consider 6 priors:
- Noninformative: Noninformative version of natural conjugate prior
- Natural conjugate: Informative natural conjugate prior with subjectively chosen prior hyperparameters
- Minnesota: Minnesota prior
- Independent Normal-Wishart: Independent Normal-Wishart prior with subjectively chosen prior hyperparameters
- SSVS-VAR: SSVS prior for VAR coefficients and Wishart prior for Σ^{-1}
- SSVS: SSVS on both VAR coefficients and error covariance

- Point estimates for VAR coefficients often are not that interesting, but Table 1 presents them for 2 priors
- With SSVS priors, $\Pr\left(\gamma_j = 1|y\right)$ is the "posterior inclusion probability" for each coefficient, see Table 2
- Model selection using $\Pr\left(\gamma_j = 1|y\right) > \frac{1}{2}$ restricts 25 of 39 coefficients to zero.

Table 1. Posterior mean of VAR Coefficients for Two Priors						
	Noninformative			SSVS - VAR		
	$\Delta\pi_t$	u_t	r_t	$\Delta\pi_t$	u_t	r_t
Intercept	0.2920	0.3222	-0.0138	0.2053	0.3168	0.0143
$\Delta\pi_{t-1}$	1.5087	0.0040	0.5493	1.5041	0.0044	0.3950
u_{t-1}	-0.2664	1.2727	-0.7192	-0.142	1.2564	-0.5648
r_{t-1}	-0.0570	-0.0211	0.7746	-0.0009	-0.0092	0.7859
$\Delta\pi_{t-2}$	-0.4678	0.1005	-0.7745	-0.5051	0.0064	-0.226
u_{t-2}	0.1967	-0.3102	0.7883	0.0739	-0.3251	0.5368
r_{t-2}	0.0626	-0.0229	-0.0288	0.0017	-0.0075	-0.0004
$\Delta\pi_{t-3}$	-0.0774	-0.1879	0.8170	-0.0074	0.0047	0.0017
u_{t-3}	-0.0142	-0.1293	-0.3547	0.0229	-0.0443	-0.0076
r_{t-3}	-0.0073	0.0967	0.0996	-0.0002	0.0562	0.1119
$\Delta\pi_{t-4}$	0.0369	0.1150	-0.4851	-0.0005	0.0028	-0.0575
u_{t-4}	0.0372	0.0669	0.3108	0.0160	0.0140	0.0563
r_{t-4}	-0.0013	-0.0254	0.0591	-0.0011	-0.0030	0.0007

Table 2. Posterior Inclusion Probabilities for VAR Coefficients: SSVS-VAR Prior			
	$\Delta\pi_t$	u_t	r_t
Intercept	0.7262	0.9674	0.1029
$\Delta\pi_{t-1}$	1	0.0651	0.9532
u_{t-1}	0.7928	1	0.8746
r_{t-1}	0.0612	0.2392	1
$\Delta\pi_{t-2}$	0.9936	0.0344	0.5129
u_{t-2}	0.4288	0.9049	0.7808
r_{t-2}	0.0580	0.2061	0.1038
$\Delta\pi_{t-3}$	0.0806	0.0296	0.1284
u_{t-3}	0.2230	0.2159	0.1024
r_{t-3}	0.0416	0.8586	0.6619
$\Delta\pi_{t-4}$	0.0645	0.0507	0.2783
u_{t-4}	0.2125	0.1412	0.2370
r_{t-4}	0.0556	0.1724	0.1097

Impulse Response Analysis

- Impulse response analysis is commonly done with VARs
- Given my focus on the Bayesian econometrics, as opposed to macroeconomics, I will not explain in detail
- The VAR so far is a reduced form model:

$$y_t = a_0 + \sum_{j=1}^{p} A_j y_{t-j} + \varepsilon_t$$

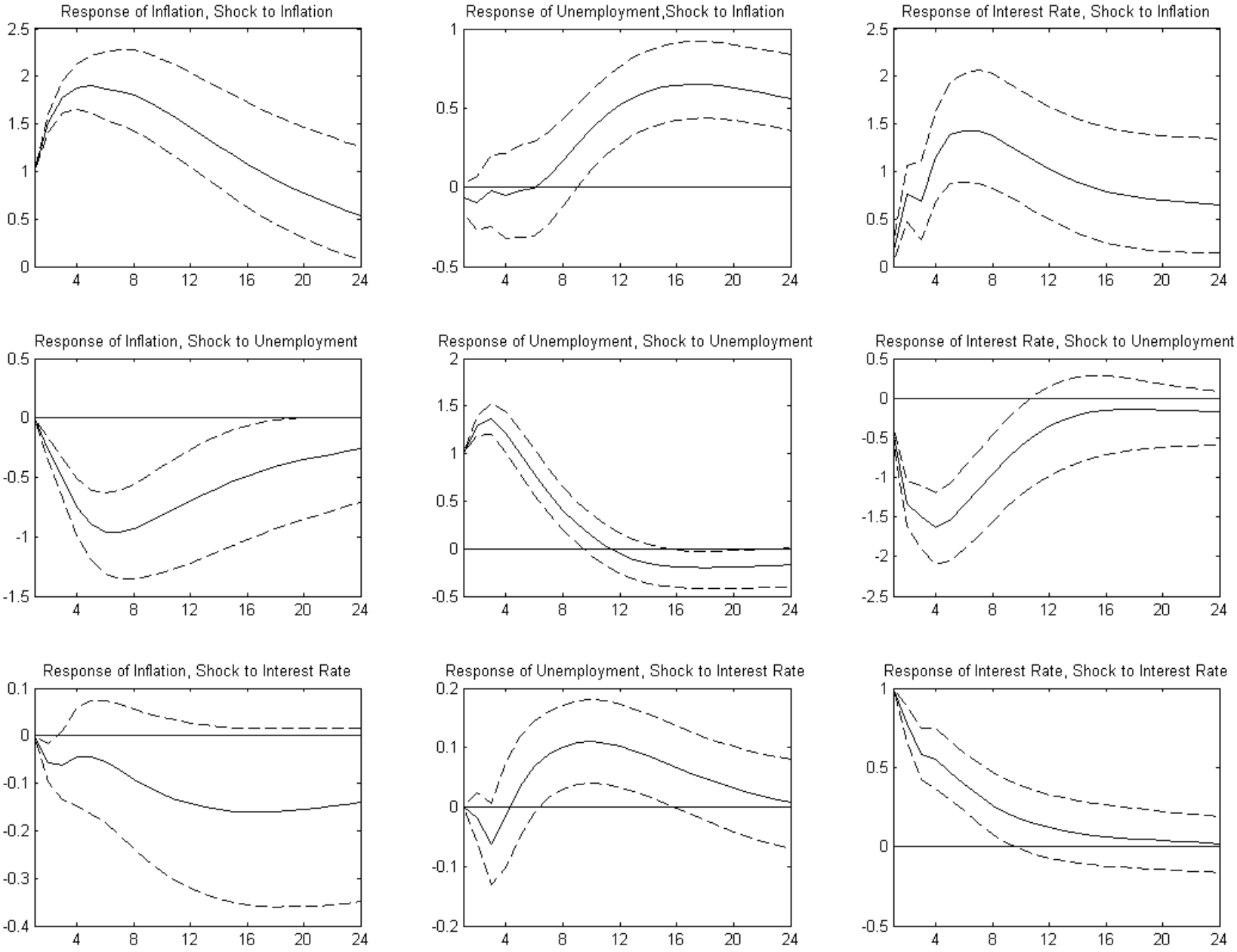
- where $var(\varepsilon_t) = \Sigma$
- Macroeconomists often work with structural VARs:

$$C_0 y_t = c_0 + \sum_{j=1}^{p} C_j y_{t-j} + u_t$$

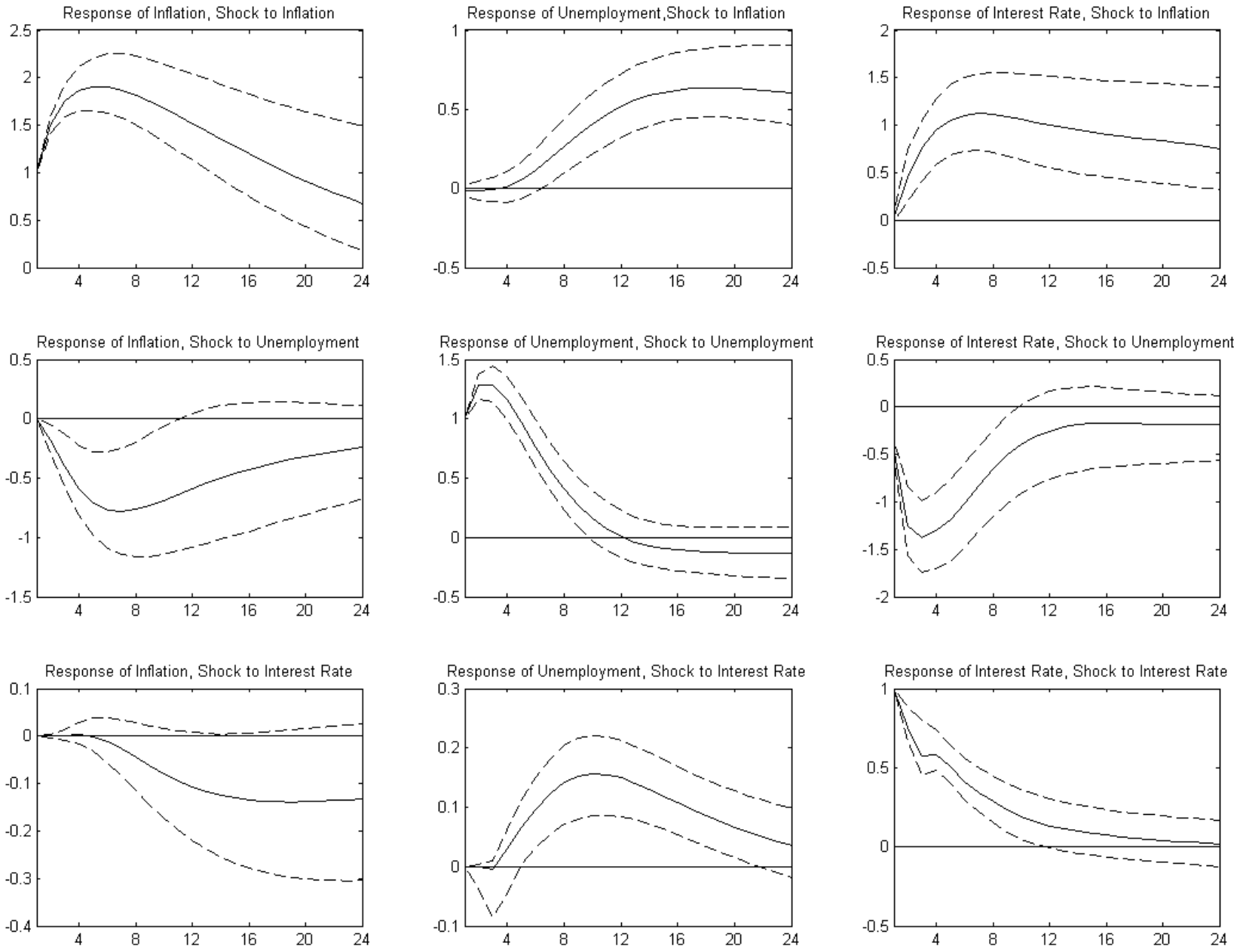
- where $var(u_t) = I$
- u_t are shocks which have an economic interpretation (e.g. monetary policy shock)

- Macroeconomist interested in effect of (e.g.) monetary policy shock now on all dependent variables in future = impulse response analysis
- Need to restrict C_0 to identify model.
- We assume C_0 lower triangular
- This is a standard identifying assumption used, among many others, by Bernanke and Mihov (1998), Christiano, Eichanbaum and Evans (1999) and Primiceri (2005).
- Allows for the interpretation of interest rate shock as monetary policy shock.
- Aside: sign-restricted impulse responses of Uhlig (2005) are increasingly popular

- Figures 2 and 3 present impulse responses of all variables to shocks
- Use two priors: the noninformative one and the SSVS prior.
- Posterior median is solid line and dotted lines are 10^{th} and 90^{th} percentiles.
- Priors give similar results, but a careful examination reveals SSVS leads to slightly more precise inferences (evidenced by a narrower band between the 10^{th} and 90^{th} percentiles) due to the shrinkage it provides.



Impulse Responses for Noninformative Prior



Impulse Responses for SSVS Prior

Illustration of Bayesian Methods in Larger VARs

- This illustration is based on my paper: "Forecasting with Medium and Large Bayesian VARs," Journal of Applied Econometrics, 2013.
- VARs have a long and successful tradition in the forecasting literature
- VARs are parameter-rich models and shrinkage can greatly improve forecast performance
- Bayesian methods popular since prior information offers a formal way of shrinking forecasts.
- Number of dependent variables is in VARs is usually small (e.g. 2 or 3)
- However, Banbura, Giannone and Reichlin (2010, JAE) started a literature working with larger Bayesian VARs.
- Other recent examples include Carriero, Clark and Marcellino (2011, Cleveland Fed), Carriero, Kapetanios and Marcellino (2009, JoF), Giannone, Lenza, Momferatou and Onorante (2010, ECB), Gefang (2013, JoF) and Korobilis (2013, JAE)

- BGR "medium" VAR has 20 dep vars and "large" VAR has 130
- Usually, when working with so many macroeconomic variables, factor methods are used
- We will discuss factor methods in later lecture
- However, BGR find that medium and large Bayesian VARs can forecast better than factor methods
- Perhaps Bayesian VARs should be used even when researcher has dozens or hundreds of variables?
- Dimensionality of α is key
- Large VAR with quarterly data might have $n = 100$ and $p = 4$ so α contains over 40000 coefficients.
- With monthly data it would have over 100000 coefficients.
- For a medium VAR, α might have about 1500 coefficients with quarterly data.
- Σ is parameter rich: $\frac{n(n+1)}{2}$ elements.

- Number of parameters may far exceed the number of observations.
- In theory, this is no problem for Bayesian methods.
- These combine likelihood function with prior.
- Even if parameters in likelihood function are not identified, combining with prior will (under weak conditions) lead to valid posterior density
- But how well do they work in practice?
- Role of prior information becomes more important as likelihood is less informative

Reminder: Natural conjugate priors for VARs

- BGR work with a natural conjugate prior which provides analytical results for predictive density.
- Natural conjugate prior has the form:

$$\alpha|\Sigma \sim N\left(\underline{\alpha}, \Sigma \otimes \underline{V}\right)$$

$$\Sigma^{-1} \sim W\left(\underline{S}^{-1}, \underline{\nu}\right)$$

- Traditional Minnesota prior replaces Σ with an estimate, $\widehat{\Sigma}$.
- Assumes Σ to be a diagonal matrix with elements s_i^2
- s_i^2 is OLS estimate of error variance in AR(p) model for i^{th} variable.
- Wishing to allow for correlations between the errors, BGR treats Σ as unknown matrix with prior inspired by Minnesota prior

- Posterior is:

$$\alpha|\Sigma, Y \sim N\left(\overline{\alpha}, \Sigma \otimes \overline{V}\right)$$

$$\Sigma^{-1}|Y \sim W\left(\overline{S}^{-1}, \overline{\nu}\right)$$

- Key thing: analytical formulae exist for posterior and predictive density
- Can choose prior hyperparameters to shrink forecasts towards a random walk or white noise
- λ is a single scalar used to control degree of shrinkage
- BGR use prior which coincides with traditional Minnesota prior except:
- Σ is treated as unknown
- A single λ is used for shrinkage instead of the two (traditionally: less shrinkage on own lags than other lags).

- Benefit 1 of natural conjugate prior: analytical results are available (no posterior simulation required)
- Benefit 2: For large Bayesian VARs: the $\Sigma \otimes \overline{V}$ form for posterior covariance matrix of α enormously simplifies computation.
- Posterior covariance involves inverting an $nK \times nK$ matrix
- With natural conjugate prior, this can be done by inverting Σ $(n \times n)$ and $\overline{V}$ $(K \times K)$ — much much easier for large VARs
- But there are disadvantages of natural conjugate prior relating to $\Sigma \otimes \underline{V}$ form.
- E.g. Prior variance of the coefficients on the same explanatory variable in any two equations must be proportional to one another
- So cannot shrink own lags differently than other lags as in Minnesota prior

- Main thing I investigate in this application is use of SSVS methods
- Another thing is the following:
- Use Minnesota prior with two prior hyperparameters controlling shrinkage (λ_1 and λ_2)
- Traditional Minnesota prior has Σ being diagonal
- We do this, but also modification of standard Minnesota prior where upper-left hand block of Σ (corresponding to a reduced set of important variables which are the ones being forecast) is not diagonal.
- Minnesota and BGR prior involve choosing prior hyperparameters in some way (i.e. prior means, shrinkage parameters, prior for Σ).
- Note: Prior for Σ depends on data (s_i^2) which may offend Bayesian purists

Non-conjugate SSVS Prior

- How well do SSVS methods work in large VARs?
- In this lecture I described George, Ni and Sun (2008) implementations
- This used a non-conjugate prior
- They used MCMC methods drawing from $p(\alpha|Y,\Sigma,\gamma)$, $p(\gamma|Y,\Sigma,\alpha)$ and $p(\Sigma|Y,\alpha,\gamma)$
- Easy forms for all, but a key stumbling block:

$$var(\alpha|Y,\Sigma,\gamma) = [\Sigma^{-1} \otimes (X'X) + D^{-1}]^{-1}$$

- Inversion of a $Kn \times Kn$ matrix must be done for each MCMC draw.
- For medium VARs slow but feasible
- For large VARs it is infeasible.

Conjugate SSVS Prior

- There also is a conjugate SSVS prior developed by Brown, Vannucci and Fearn (1998, JRSS, B)
- Big advantage: does not require inversion of $Kn \times Kn$ matrix at each MCMC draw.
- But has some other drawbacks:
- With non-conjugate γ has Kn elements (each individual coefficient is either included/excluded)
- With conjugate we have $\widetilde{\gamma}$ with K elements (coefficients on a single explanatory variables are either included/excluded in all equations)
- With conjugate we cannot do SSVS on off-diagonal elements of Σ
- Posterior simulation only involves drawing from $p\left(\widetilde{\gamma}|Y\right)$ (much easier than with non-conjugate, but still much more computationally demanding than BGR and Minnesota prior)

Combining Minnesota Prior with SSVS prior

- SSVS priors require choice of prior variances in Normal mixture prior
- George, Sun and Ni (2008) want κ_{0j}^2 very small and κ_{1j}^2 large and recommend a "semi-automatic" way of choosing them
- We use this approach
- However, the Minnesota prior is Normal and the second element in the SSVS mixture prior also Normal
- Why not simply use Minnesota prior for second Normal?
- SSVS can either choose to shrink any coefficient to zero or to use the Minnesota prior.
- We also use this Minnesota+SSVS approach

Data

- See paper for details, updated version of data set used in Stock and Watson (2008)
- 168 US variables from 1959Q1 through 2008Q4
- Variables all transformed to stationarity (usually by differencing or log differencing)
- Variables are divided into four groups:
- First group: three main variables we are interested in forecasting (output, prices and interest rates)
- Second group: 17 variables commonly used for forecasting first group (partly motivated by the monetary model of Christiano, Eichenbaum and Evans (1999) and partly includes variables found to be useful for forecasting in other studies)
- Third group: 20 variables sometimes used in forecasting exercises (this includes most of remaining aggregate variables)
- Fourth group: Remaining 128 variables (mostly components making up the aggregate variables already included in the other groups)

- Small VARs use first group ($n = 3$)
- Medium VARs use first two groups ($n = 20$)
- Medium-large VARs use first three groups ($n = 40$)
- Large VARs use all groups ($n = 168$).

Other Modelling Choices

- Lag length is four in all VARs
- For Minnesota priors (and combination Minnesota plus SSVS priors) have to choose prior shrinkage parameter(s) λ (or λ_1 and λ_2).
- We follow BGR:
- Estimate VARs on a training sample (data through 1969Q4)
- Do forecasting exercise within this training sample.
- For the small VAR no shrinkage is done ($\lambda \rightarrow \infty$).
- In medium, medium-large and large VARs, λ (or λ_1 and λ_2) chosen to yield sum of MSFEs for three main variables in training sample as close as possible to the small VAR.

Forecast Metrics

- Rolling and recursive forecast exercises
- Rolling forecasts: use a window of ten years
- We obtain predictive density for $y_{\tau+h}$ using data available through time τ for $h = 1$ and 4
- $\tau = \tau_0, .., T - h$ where τ_0 is 1969Q4
- Use MSFEs to assess performance of point forecasts
- Use cumulative sums of log predictive likelihoods to assess performance of predictive densities
- Forecast GDP growth, inflation and the interest rate
- We present MSFE as a proportion of the MSFE produced by random walk forecasts

Factor Methods

- Compare our Bayesian VAR approaches to each other and to factor methods (described in later lecture)
- Add lags of factors to the small VAR involving
- Factors constructed using principal components based on the remaining 165 variables.
- Use three factors and implement variants where we include one and four lags of these factors
- We use noninformative prior

Summary of Forecasting Approaches

1. Minnesota Prior as in BGR
2. Traditional Minnesota Prior (Σ diagonal)
3. Traditional Minnesota Prior (Σ not diagonal)
4. SSVS Conjugate prior (semi-automatic selection of prior hyperparameters)
5. SSVS Conjugate prior + Minnesota Prior
6. SSVS Non-conjugate prior (semi-automatic selection of prior hyperparameters)
7. SSVS Non-conjugate prior + Minnesota Prior
8. Factor methods with one lag of factors
9. Factor methods with four lags of factors

- Computational constraints: doing posterior simulation at each point in time in a recursive forecasting exercise very computationally demanding
- Remember, too, the need to invert huge-dimensional matrices when constructing posterior covariance matrix with non-conjugate SSVS
- These constraints mean:
- For Minnesota priors we present results for small, medium, medium-large and large VARs
- For conjugate SSVS we present results for small, medium and medium-large VARs
- For non-conjugate SSVS we present results for small and medium VARs
- For factor methods we use all 168 variables

Forecasting Results

- We are forecasting three variables (output=GDP, inflation=CPI and interest rates=FFR), at two horizons ($h = 1$ and 4) and are doing recursive and rolling forecasts.
- Thus 12 tables in paper
- Begin with table which summarizes which approach is best in each of the 12 cases

Method	GDP	CPI	FFR
Using MSFE to Measure Forecast Performance			
$h = 1$ rec.	Minn. Prior as BGR $n = 40$	Minn. Pri. Σ not diag., $n = 20$	Minn. Pri. as BGR, $n = 40$
$h = 4$ rec.	Minn. Prior as BGR, $n = 168$	SSVS No-con+ Min Prior $n = 3$	SSVS No-conj. sem-auto $n = 3$
$h = 1$ roll.	SSVS Conj. + Min Pri $n = 40$	Minn. Prior Σ diag, $n = 40$	SSVS Conj + Min Pri, $n = 20$
$h = 4$ roll.	Min Pr, Σ not diag, $n = 168$	Minn. Pri Σ not diag $n = 168$	SSVS Nonconj sem-auto $n = 3$
Using Pred. Likes. to Measure Forecast Performance			
$h = 1$ rec.	Minn. Pri as BGR, $n = 40$	Minn. Prior Σ diag., $n = 20$	SSVS Conj. + Minn. Pri $n = 20$
$h = 4$ rec.	SSVS Noncon sem-auto $n = 3$	SSVS Noncon+ Min Prior, $n = 3$	SSVS Noncon + Minn. Pri, $n = 3$
$h = 1$ roll.	Min Prior Σ not diag, $n = 168$	SSVS Non-conj semi-auto $n = 3$	SSVS Conj semiauto $n = 20$
$h = 4$ roll.	Min Prior Σ not diag, $n = 168$	SSVS Conj. semi-auto $n = 3$	SSVS Non-conj semiauto $n = 3$

General Discussion of Results

- No one single forecasting method predominates in all cases.
- In practice some approaches doing well in some cases, but not necessarily in others.
- Nevertheless, a few interesting stories emerge:
- Factor methods never lead to the best forecast performance.
- In all cases, most of our ways of implementing VARs lead to better (and often much better) forecast performance than factor models
- Recommendation: working with high-dimensional Bayesian VARs is an alternative worth considering.

- Who wins our 24 forecast "races"?
- 24 arises since 12 cases can be evaluated either using MSFEs or sums of log predictive likelihoods
- SSVS approaches have 13 "wins"
- Minnesota prior approaches win 11 times
- In terms of VAR dimensionality, large, medium-large and medium VARs each win five times and small VARs win nine times
- Fairly even split. Cannot recommend (e.g.) "you should always use SSVS" or "you should always work with large VARs"

More Detailed Discussion of Results

- Recursive and rolling results qualitatively similar so only present recursive here
- Factor model with four lags always does worse than one lag so omit
- Small VARs often forecast well, but in many cases, reading across rows in tables find improvements in forecasts
- But improvements tend to be small or non-existent when we move beyond $n = 20$
- This is consistent with BGR findings

- Exceptions to previous pattern usually with $h = 4$
- Here small VARs with SSVS priors often yield best forecasting performance.
- Pattern: SSVS methods forecast better than Minnesota priors in small VARs, but pattern is not always continued with medium and medium-large VARs.
- Minnesota priors: BGR's specification often works well.
- However, often an alternative forecasts slightly better than BGR.
- Traditional Minnesota prior had different degrees of shrinkage for coefficients on own lags than on other lags – this often improves forecast performance.
- Our version of Minnesota prior which allows upper left 3×3 block of Σ to be unrestricted often forecasts quite well.
- Combining SSVS with Minnesota often works well (may be a good conservative choice)

Table 1: GDP Forecasting for $h = 1$ MSFE's above pred. likes				
	$n = 3$	$n = 20$	$n = 40$	$n = 168$
Minn. Prior as in BGR	0.6504 −206.37	0.5552 −192.29	0.5084 −186.60	0.5225 −223.78
Minn. Prior Σ diagonal	0.7065 −211.85	0.5774 −204.84	0.6381 −205.52	0.5631 −202.39
Minn. Prior Σ not diagonal	0.7065 −205.97	0.5489 −195.40	0.5402 −193.49	0.5305 −192.81
SSVS Conjugate semi-automatic	0.6338 −200.66	0.6776 −199.90	0.6983 −197.66	n.a.
SSVS Conjugate plus Minn. Prior	0.6062 −198.77	0.5577 −192.53	0.5368 −192.4	n.a.
SSVS Non-conj. semi-automatic	0.6061 −198.40	0.6407 −205.12	n.a.	n.a.
SSVS Non-conj. plus Minn. Prior	0.6975 −204.71	0.6466 −203.92	n.a.	n.a.
Factor Model $p = 1$	n.a.	n.a.	n.a.	0.6441 −195.10

Table 2: CPI Forecasting for $h = 1$, MSFE's above pred. likes				
	$n = 3$	$n = 20$	$n = 40$	$n = 168$
Minn. Prior as in BGR	0.3471 −201.23	0.3029 −195.90	0.3172 −210.09	0.3309 −322.27
Minn. Prior Σ diagonal	0.3317 −190.85	0.2756 −182.18	0.3252 −200.55	0.2834 −184.78
Minn. Prior Σ not diagonal	0.3317 −203.95	0.2664 −184.06	0.2718 −188.27	0.3019 −197.45
SSVS Conjugate semi-automatic	0.3138 −187.82	0.2724 −191.15	0.3061 −197.66	n.a.
SSVS Conjugate plus Minn. Prior	0.3086 −186.70	0.3088 −197.64	0.3601 −222.30	n.a.
SSVS Non-conj. semi-automatic	0.3197 −193.92	0.3161 −196.47	n.a.	n.a.
SSVS Non-conj. plus Minn. Prior	0.3252 −191.45	0.2910 −187.58	n.a.	n.a.
Factor Model $p = 1$	n.a.	n.a.	n.a.	0.3133 −191.83

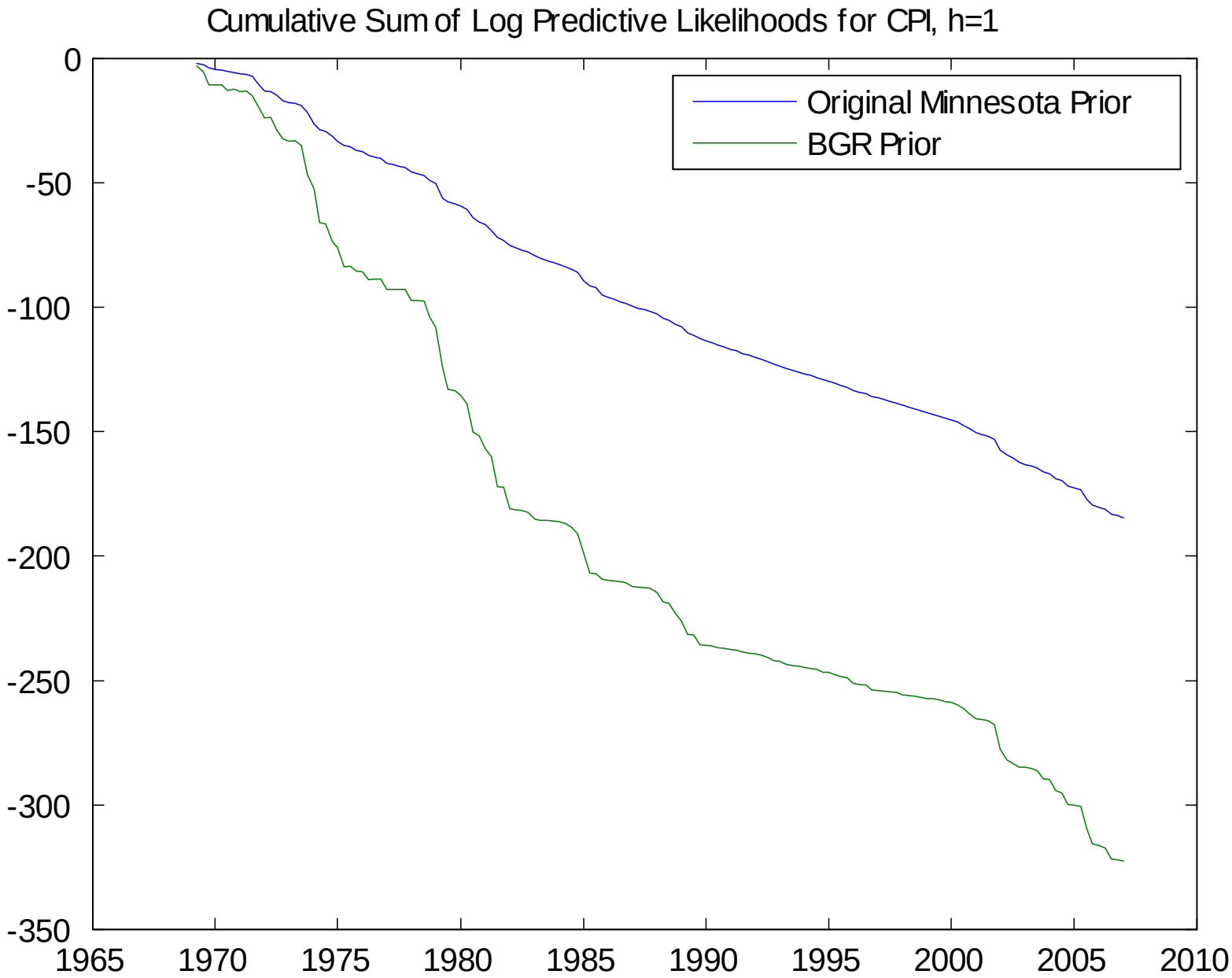
Table 3: FFR Forecasting for $h = 1$, MSFE's above pred. likes				
	$n = 3$	$n = 20$	$n = 40$	$n = 168$
Minn. Prior as in BGR	0.6192 −238.40	0.5136 −229.14	0.5084 −243.71	0.5224 −266.66
Minn. Prior Σ diagonal	0.8351 −247.02	0.5355 −238.79	0.6218 −263.05	0.5532 −239.84
Minn. Prior Σ not diagonal	0.8351 −267.29	0.5164 −249.09	0.5530 −249.49	0.5223 −258.28
SSVS Conjugate semi-automatic	0.7944 −247.25	0.6329 −245.25	0.5936 −256.02	n.a.
SSVS Conjugate plus Minn. Prior	0.7554 −243.16	0.5134 −228.54	0.5354 −251.98	n.a.
SSVS Non-conj. semi-automatic	0.8439 −252.43	0.5790 −237.16	n.a.	n.a.
SSVS Non-conj. plus Minn. Prior	0.7436 −252.68	0.5431 −228.86	n.a.	n.a.
Factor Model $p = 1$	n.a.	n.a.	n.a.	0.7360 −232.66

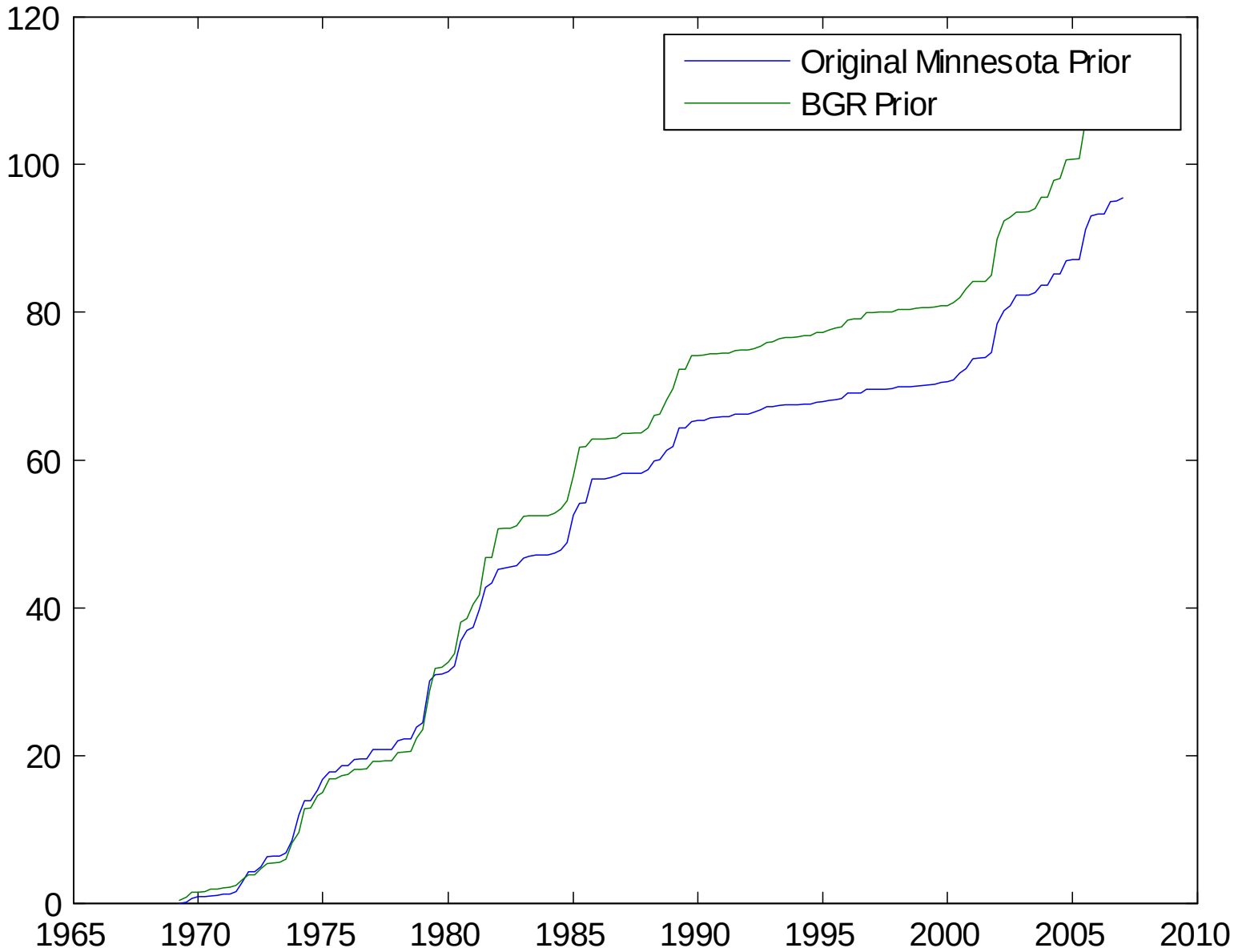
Table 4: GDP Forecasting for $h = 4$, MSFE's above pred. likes				
	$n = 3$	$n = 20$	$n = 40$	$n = 168$
Minn. Prior as in BGR	0.7437 −220.57	0.6094 −214.71	0.5717 −214.38	0.5420 −277.46
Minn. Prior Σ diagonal	0.7437 −219.25	0.6100 −214.02	0.5728 −210.99	0.5656 −210.06
Minn. Prior Σ not diagonal	0.7437 −220.58	0.6214 −213.28	0.5831 −209.50	0.5780 −209.37
SSVS Conjugate semi-automatic	0.6129 −211.36	0.6473 −212.35	0.8881 −239.87	n.a.
SSVS Conjugate plus Minn. Prior	0.8404 −222.91	0.8357 −219.64	0.6387 −222.50	n.a.
SSVS Non-conj. semi-automatic	0.6147 −207.80	0.7535 −293.21	n.a.	n.a.
SSVS Non-conj. plus Minn. Prior	0.8438 −221.57	0.6670 −219.01	n.a.	n.a.
Factor Model $p = 1$	n.a.	n.a.	n.a.	0.7662 −223.24

Table 5: CPI Forecasting for $h = 4$, MSFE's above pred. likes				
	$n = 3$	$n = 20$	$n = 40$	$n = 168$
Minn. Prior as in BGR	0.5254 −209.51	0.5217 −219.35	0.5246 −235.65	0.5044 −262.55
Minn. Prior Σ diagonal	0.5254 −216.43	0.5191 −217.60	0.5207 −218.45	0.5124 −216.64
Minn. Prior Σ not diagonal	0.5254 −214.64	0.5203 −216.07	0.5214 −217.57	0.5197 −217.14
SSVS Conjugate semi-automatic	0.4990 −211.36	0.6042 −225.02	0.6847 −253.84	n.a.
SSVS Conjugate plus Minn. Prior	0.4759 −199.86	0.7031 −246.64	0.4853 −220.44	n.a.
SSVS Non-conj. semi-automatic	0.5010 −208.26	0.7723 −226.36	n.a.	n.a.
SSVS Non-conj. plus Minn. Prior	0.4683 −194.39	0.4883 −201.62	n.a.	n.a.
Factor Model $p = 1$	n.a.	n.a.	n.a.	0.5608 −214.72

Table 6: FFR Forecasting for $h = 4$, MSFE's above pred. likes				
	$n = 3$	$n = 20$	$n = 40$	$n = 168$
Minn. Prior as in BGR	0.6679 −243.31	0.5868 −249.63	0.5670 −264.80	0.5717 −319.40
Minn. Prior Σ diagonal	0.6677 −281.95	0.6075 −278.11	0.5946 −273.70	0.6379 −281.92
Minn. Prior Σ not diagonal	0.6679 −246.90	0.5882 −244.77	0.5894 −240.50	0.6362 −245.34
SSVS Conjugate semi-automatic	0.5508 −236.00	0.5873 −249.46	0.7408 −273.60	n.a.
SSVS Conjugate plus Minn. Prior	0.6259 −235.57	0.6716 −258.47	0.5370 −255.89	n.a.
SSVS Non-conj. semi-automatic	0.5265 −231.16	0.8811 −268.06	n.a.	n.a.
SSVS Non-conj. plus Minn. Prior	0.6184 −228.80	0.5282 −233.67	n.a.	n.a.
Factor Model $p = 1$	n.a.	n.a.	n.a.	0.7027 −244.52

- In terms of MSFEs, the advantages of unrestricted Σ (as in BGR) are relatively small.
- In terms of predictive likelihoods, allowing for unrestricted Σ can occasionally lead to poor forecast performance.
- Example: forecasting CPI for $h = 1$.
- MSFEs say best method uses Minnesota prior with 20-variate VAR where MSFE is 0.2664.
- In large VARs, the MSFEs for Minnesota and BGR are only slightly higher (0.2834 and 0.3309).
- But in large VARs predictive likelihoods vastly different between Minnesota prior (-184.78) and BGR (-322.27).
- Figures plot cumulative sum of log pred likes and cumulative sum of squared forecast errors for these two priors for large VAR.
- Predictive likelihoods and MSFEs (point forecasts) can say very different things
- Dispersion/tail behaviour of predictives can matter.
- In this case, BGR has good point forecast but is yielding unnecessarily disperse predictive distribution due to its need to estimate many more parameters in Σ


Cumulative Sum of Log Predictive Likelihoods for CPI, h=1
Original Minnesota Prior
BGR Prior
0
-50
-100
-150
-200
-250
-300
-350
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010


Cumulative Sum of Squared Forecast Errors for CPI, h=1
Original Minnesota Prior
BGR Prior
120
100
80
60
40
20
0
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010

Summary of Large VAR application

- We consider various priors which differ in how they do shrinkage in Bayesian VARs
- Also differ in computational burden
- Investigate forecast performance in a substantive empirical example and find:
- Bayesian VARs consistently out-forecast factor models even in large VARs
- Generally, improvements in forecasting are small or non-existent beyond $n = 20$ or $n = 40$
- With such medium and medium-large VARs, SSVS methods are possible and often forecast better than Minnesota priors
- Different variants of Minnesota priors are also worth considering

Conclusion of VAR Lecture

- Lecture began with summary of basic methods and issues which arise with Bayesian VAR modelling and addressed questions such as:
- Why is shrinkage necessary?
- How should shrinkage be done?
- With recent explosion of interest in large VARs, need for answers for such questions is greatly increased
- Many researchers now developing models/methods to address them
- I have described one popular category focussing on SSVS methods
- But many more exist (e.g. variants on LASSO) and are coming out all the time
- Recent survey paper: Sune Karlsson: Forecasting with Bayesian Vector Autoregressions (to appear in Handbook of Economic Forecasting, volume 2)