Hierarchical Priors for Vector Autoregressive Models

Introduction

- Many new methods are gaining popularity in Bayesian econometrics which involve hierarchical priors
- Useful for Big Data contexts (VARs or regressions with $K > T$)
- Machine learning methods which automatically shrink many coefficients to zero so as to ensure parsimony
- First illustrate the basic ideas of two popular methods in regression:
- Stochastic search variable selection (SSVS) and Least Absolute Shrinkage and Selection Operator (LASSO)
- Then move on to VARs
- End with an extension of the LASSO: Dirichlet-Laplace hierarchical prior

Variable Selection and Shrinkage Using Hierarchical Priors

- Any sort of prior information can be used to overcome lack of data information with Big Data regression or VAR
- E.g. Minnesota prior is subjective prior suggested by empirical wisdom of earlier researchers
- But what if researcher does not have such prior information or does not wish to use it?
- Hierarchical priors are a common alternative
- I introduce two popular ones: LASSO and SSVS
- Many others (and not all Bayesian)
- Korobilis, D. (2013). Hierarchical shrinkage priors for dynamic regressions with many predictors. International Journal of Forecasting 29, 43-59.

SSVS: Overview

- To show main ideas assume (for now) β is a scalar
- Remember prior shrinkage can be done through prior variance: $\beta \sim N(0, \underline{V})$
- If $\underline{V}$ is small, then strong prior information β is near 0.
- E.g. $\underline{V} = 0.0001$ then $\Pr\left(-0.0196 \leq \beta \leq 0.0196\right) = 0.95$
- If $\underline{V}$ is big then prior becomes more non-informative
- If $\underline{V} = 100$ then $\Pr\left(-19.6 \leq \beta \leq 19.6\right) = 0.95$
- Note: exactly what "small" and "large" means depends on the empirical application and units of measurement of data

SSVS: Overview

- SSVS prior:

$$\beta|\gamma \sim (1-\gamma)\, N\left(0, \tau_0^2\right) + \gamma N\left(0, \tau_1^2\right)$$

- τ_0 is small and τ_1 is large
- $\gamma = 0$ or 1.
- If $\gamma = 0$, tight prior shrinking coefficient to be near zero
- If $\gamma = 1$, non-informative prior and β estimated in a data- based fashion.
- SSVS treats γ as unknown and estimates it
- Data choose whether to select a variable or omit it (in the sense of shrinking its coefficient to be very near zero).
- Can be implemented in various ways, here we follow George, Sun and Ni (2008, Journal of Econometrics)

SSVS: Overview

- prior for β is hierarchical: depends on γ which has its own prior.
- Gibbs sampler takes draw of γ and, conditional on these, results for independent Normal-Gamma prior used to draw β and h.
- If $\gamma = 1$ use $N\left(0, \tau_1^2\right)$ prior, else use $N\left(0, \tau_0^2\right)$
- Output from this GIbbs sampler can be used to:
- Do something similar to Bayesian model averaging (BMA): averages over restricted (when $\gamma = 0$ is drawn) and unrestricted ($\gamma = 1$) models
- Do Bayesian model selection (BMS)
- If $\Pr\left(\gamma = 1|y\right) > \frac{1}{2}$ choose unrestricted model, else choose restricted model
- Can use threshold other than $\frac{1}{2}$

SSVS in Multiple Regression

- Any Bayesian textbook gives posterior results for regression model with prior
$$N\left(\underline{\beta},\underline{V}\right)$$
- SSVS prior makes specific choices for $\underline{\beta}$ and $\underline{V}$
- $\underline{\beta}=0$ so as to shrink coefficients towards zero
-
$$\underline{V}=DD$$
- D is diagonal matrix with elements
$$d_i=\begin{cases}\tau_{0i} \text{ if } \gamma_i=0\\ \tau_{1i} \text{ if } \gamma_i=1\end{cases}$$
- We now have $i=1,..,K$
- $\gamma_i\in\{0,1\}$ indicating whether each variable is excluded
- Small/large prior variances, τ_{0i}^2 and τ_{1i}^2, for each variable

SSVS: Gibbs Sampler

- Conditional on draw of γ we are in familiar world
- Use independent Normal-Gamma posterior for β and h
- What about γ?
- Needs a prior
- A simple choice is:

$$\Pr(\gamma_i = 1) = \underline{q}_i$$
$$\Pr(\gamma_i = 0) = 1 - \underline{q}_i$$

- Non-informative choice is $\underline{q}_i = \frac{1}{2}$ (each coefficient is *a priori* equally likely to be included as excluded)

SSVS: Gibbs Sampler

- Can show conditional posterior distribution is Bernoulli:

$$\Pr\left(\gamma_i = 1 | y, \gamma\right) = \overline{q}_i,$$
$$\Pr\left(\gamma_i = 0 | y, \gamma\right) = 1 - \overline{q}_j,$$

- where

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

SSVS: Choosing Small and Large Prior Variances

- Researcher must choose τ_{0i}^2 and τ_{1i}^2
- Want τ_{0i}^2 to imply virtually all of prior probability is attached to region where β_i is so small as to be negligible
- Approximate rule of thumb: 95% of the probability of a distribution lies within two standard deviations from its mean.
- E.g. is $\tau_{0i} = 0.01$ small?
- Expresses a prior belief that β_i is less than 0.02 in absolute value.
- Is $\beta_i = 0.02$ a "small" value or not?
- Depends on empirical application at hand and units dependent and explanatory variables are measured in
- Sometimes researcher can subjectively make good choices for τ_{0i}
- But often not, want a method of choosing them that does not require (much) prior input from researcher

SSVS: Choosing Small and Large Prior Variances

- Common to use "default semi-automatic approach"
- Choose τ_{0i}^2 and τ_{1i}^2 based on initial estimation procedure.
- Use initial estimates (e.g. OLS) from regression with all exp vars:
- produce $\widehat{\sigma}_i$ – the standard error of β_i.
- Set $\tau_{0i} = \frac{1}{c} \times \widehat{\sigma}_i$ and $\tau_{1i} = c \times \widehat{\sigma}_i$ for large value for c (e.g. $c = 10$ or 100).
- Basic idea: $\widehat{\sigma}_i$ is estimate of the standard deviation of β_i
- Question: how do we choose small value for prior variance of β_i?
- Answer: choose one which is small relative to its standard deviation

LASSO: Theory

- LASSO = Least absolute shrinkage and selection operator
- Developed as a frequentist shrinkage and variable selection method for Big Data regression models
- Frequentist intuition: OLS estimates minimize sum of squared residuals

$$(y - X\beta)' (y - X\beta)$$

- LASSO minimizes

$$(y - X\beta)' (y - X\beta) + \lambda \sum_{j=1}^{k} \left|\beta_j\right|$$

- adds penalty term which depends on magnitude of the regression coefficients
- Bigger values for $\left|\beta_j\right|$ penalized (shrink towards zero)
- λ is shrinkage parameter.

LASSO: Theory

- LASSO estimate can be given a Bayesian interpretation:
- equivalent to Bayesian posterior modes if Laplace prior used for β
- I will not define Laplace distribution since will not work with it directly due to following:
- Laplace distribution can be written as scale mixture of Normals (i.e. a mixture of Normal distributions with different variances):

$$\begin{aligned} \beta_i &\sim N\left(0, h^{-1}\tau_i^2\right) \\ \tau_i^2 &\sim Exp\left(\frac{\lambda^2}{2}\right) \end{aligned}$$

- $Exp\left(.\right)$ is exponential distribution (special case of Gamma)
- Hierarchical prior: depends on τ_i^2 (parameters to be estimated) which have own prior
- Note: smaller $\tau_i^2 =$ stronger shrinkage of β_i
- Can show λ plays same role as frequentist λ above

LASSO: Theory

- Bayesian inference can be done using MCMC
- Main idea: conditional on τ_i^2, prior is Normal prior
- Can use standard results for Normal linear regression to obtain $p(\beta|y, h, \tau)$ and $p(h|y, \beta, \tau)$ where $\tau = (\tau_1, .., \tau_K)'$
- All we need is new blocks in MCMC algorithm for drawing τ and λ
- Details given in next slide, but note basic strategy same as for SSVS:
- Use hierarchical Normal prior for β
- Conditional on some other parameters (here τ, with SSVS it was γ) obtain Normal linear regression model
- So just need to work out conditional posterior for these other parameters
- Note: many variants on LASSO (elastic net LASSO) adopt similar strategy

LASSO: Theory

- Write LASSO prior covariance matrix of β as
$$\underline{V} = h^{-1}DD$$
- D is diagonal matrix with diagonal elements τ_i for $i = 1, .., K$
- Then $\beta | y, h, \tau$ is $N\left(\overline{\beta}, \overline{V}\right)$ where
$$\overline{\beta} = \left(X'X + (DD)^{-1}\right)^{-1} X'y$$
- $$\overline{V} = h^{-1}\left(X'X + (DD)^{-1}\right)^{-1}$$
- $h | y, \beta, \tau$ is $G(\overline{s}^{-2}, \overline{\nu})$ with
$$\overline{\nu} = N + K$$
- $$\overline{s}^2 = \frac{(y - X\beta)'(y - X\beta) + \beta'(DD)^{-1}\beta}{\overline{\nu}}$$

LASSO: Theory

- Easier to draw from $\frac{1}{\tau_i^2}$ for $i = 1, .., K$ as posterior conditionals are independent of one another and with inverse Gaussian distributions.
- Inverse Gaussian, $IG(.,.)$, is rarely used in econometrics.
- Standard ways for drawing from IG exist (all we need for MCMC)
- $p\left(\frac{1}{\tau_i^2}|y, \beta, h, \lambda\right)$ is $IG(\overline{c}_i, \overline{d}_i)$ with $\overline{d} = \lambda^2$

$$\overline{c}_i = \sqrt{\frac{\lambda^2}{h\beta_i^2}}$$

- Need prior for λ, convenient to use $\lambda^2 \sim G\left(\underline{\mu}_\lambda, \underline{\nu}_\lambda\right)$
- With this $p\left(\lambda^2|y, \tau\right)$ is $G(\overline{\mu}_\lambda, \overline{\nu}_\lambda)$ with

$$\overline{\nu}_\lambda = \underline{\nu}_\lambda + 2K$$

-

$$\overline{\lambda} = \frac{\underline{\nu}_\lambda + 2K}{2\sum_{i=1}^{K} \tau_i^2 + \frac{\underline{\nu}_\lambda}{\underline{\mu}_\lambda}}$$

Stochastic Search Variable Selection (SSVS) in VARs

- Now let us turn to the VAR
- Remember: over-parameterization concerns can be acute
- VAR has $K = 1 + M \times p$ explanatory variables in each of M equations
- There are many approaches which seek parsimony/shrinkage in VARs, take SSVS as a good 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.

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

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

- 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:

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

- $\underline{q}_j = \frac{1}{2}$ for all j implies each coefficient is *a priori* equally likely to be included as excluded.
- We will use standard Wishart prior for Σ^{-1}
- However, George, Sun and Ni also show how to do SSVS on off-diagonal elements of Σ (see Computer Tutorial 2)

- 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:
- $$\begin{array}{c} \Pr\left(\gamma_j = 1|y,\alpha,\Sigma\right) = \overline{q}_j \\ \Pr\left(\gamma_j = 0|y,\alpha,\Sigma\right) = 1 - \overline{q}_j \end{array}$$
- 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.
- We use unrestricted VAR with intercept and 4 lags and 2 priors:
- Noninformative: Noninformative version of natural conjugate prior
- SSVS: SSVS on both VAR coefficients and error covariance

- Compare SSVS to noninformative prior results to see role of shrinkage
- In small model such as this, role is small (will be bigger in VARs with more dependent variables)
- 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:

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

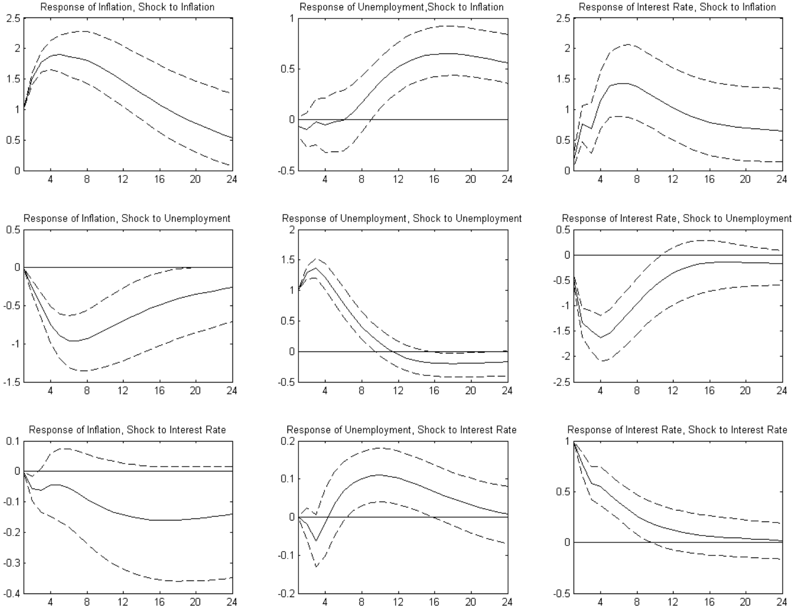
- where $var(\varepsilon_t) = \Sigma$
- Macroeconomists often work with structural VARs:

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

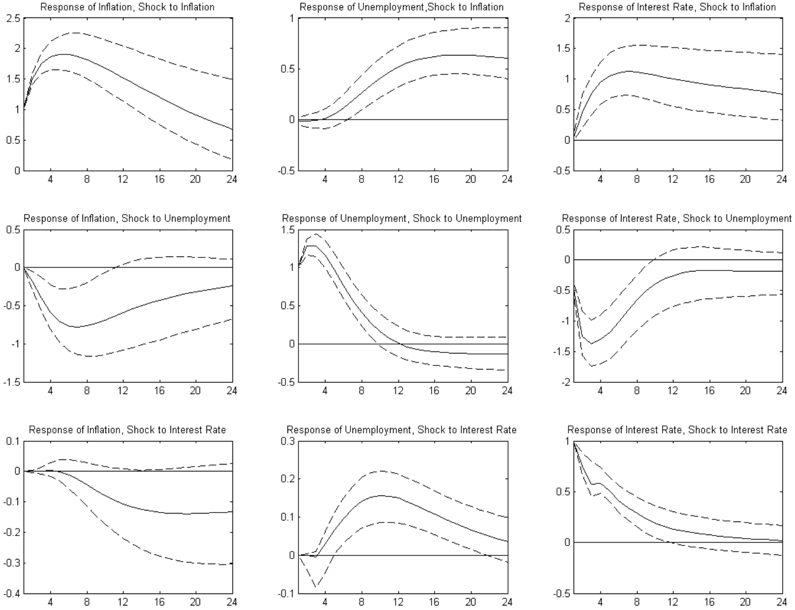
- where $var(u_t) = I$
- u_t are shocks which have an economic interpretation (e.g. monetary policy shock)

- 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

- Next figures 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

The Dirichlet-Laplace Prior

- The LASSO has been used in VARs
- E.g. Gefang (2014) Bayesian Doubly Adaptive Elastic-Net Lasso for VAR Shrinkage, International Journal of Forecasting.
- Bayesian LASSO is a Laplace hierarchical prior
- Recently an extension of LASSO involving the Dirichlet-Laplace hierarchical prior gaining in popularity
- Has good theoretical properties
- Bhattacharya et al (2015) Dirichlet–Laplace Priors for Optimal Shrinkage, Journal of American Statistical Association
- Less sensitive to choice of prior hyperparameters such as $\underline{\mu}_{\lambda}, \underline{\nu}_{\lambda}$ for LASSO
- Is being successfully used with VARs
- Kastner and Huber (2017) Sparse Bayesian vector autoregressions in huge dimensions

The Dirichlet-Laplace Prior

- Let α_j be a VAR coefficient
- $j = 1, .., k$ VAR coefficients
- The Dirichlet-Laplace prior:

$$\alpha_j \sim N(0, \psi_j \vartheta_j^2 \tau^2),$$

$$\psi_j \sim \text{Exp}(\frac{1}{2}),$$

$$\vartheta_j \sim \text{Dir}(a, \ldots, a),$$

$$\tau \sim \text{G}(\frac{ka}{2}, 2ka).$$

The Dirichlet-Laplace Prior

- Terminology: global-local shrinkage prior
- Remember: prior variance used to shrink
- τ^2 appears in prior variance for every coefficient (global)
- ψ_j appears only in prior variance for coefficient j (local)
- ϑ_j has Dirichlet distribution (see wikipedia)
- Dirichlet properties $0 \leq \vartheta_j \leq 1$ and

$$\Sigma_{j=1}^{k} \vartheta_j = 1$$

- Adding ϑ_j to LASSO prior leads to nicer theoretical properties
- Simplifies prior choice
- All you need to do is choose a
- Bhattacharya et al (2015) offers some simple suggestions (e.g. $a = \frac{1}{2}$)

Posterior Computation with Dirichlet-Laplace Prior

- Same idea as for SSVS
- Conditional on parameters of shrinkage prior (ψ_j, ϑ_j^2 and τ^2) we have Normal VAR
- Conditional posterior α takes standard form
- If Σ has Wishart prior, conditional posterior is Wishart
- Only new steps in Gibbs sampler for ψ_j, ϑ_j^2 and τ^2
- Formulae on next slide.
- GIG is generalized inverse Gaussian
- Exact details unimportant
- Key thing is that computer can easily take random draws from GIG and IG
- $\bullet$ just means the data and all the other parameters on next slide

Conditional Posteriors with Dirichlet-Laplace Prior

$$\psi_j|\bullet \sim IG(\tfrac{\vartheta_{j,}\tau}{|\alpha_j|}, 1), \quad \text{for } j = 1, \ldots, k$$

$$\tau|\bullet \sim GIG(k(\alpha - 1), 1, 2\sum_{j=1}^{k} \frac{|\alpha_j|}{\vartheta_{j,}}),$$

$$R_j|\bullet \sim GIG(\alpha - 1, 1, 2|\alpha_j|), \quad \text{for } j = 1, \ldots, k.$$

and

$$\vartheta_j = \frac{R_{j,}}{\sum_{j=1}^{k} R_{j,}}.$$

Dirichlet-Laplace Prior: Summary

- Code for the VAR Dirichlet-Laplace prior is provided as part of Computer Session 2
- You can experiment it in the session to learn more about it
- Has form where condition on some parameters, remainder of model is Normal VAR
- Makes computation simpler (add some new blocks to the Gibbs sampler)
- Other shrinkage priors are also gaining in popularity (e.g. Horseshoe prior)
- Folllett and Yu (2017) Achieving Parsimony in Bayesian VARs with the Horseshoe Prior
- Theoretical properties and simplicity of prior elicitation make Dirichlet-Laplace attractive

Conclusion

- VARs are parameter-rich models
- Hierarchical priors which automatically shrink unimportant coefficients to zero offer solution
- They are machine learning methods
- This lecture goes through SSVS, LASSO and Dirichlet-Laplace, but there are others
- MCMC methods used for posterior inference
- Too computationally demanding in truly large VARs
- E.g. they work with 25 variables, but not (yet) with 100
- Minnesota prior does work with 100+ variables
- Other machine learning methods are being developed for 100+ variables
- I will give an example in the last lecture