Bayesian Doubly Adaptive Elastic-Net Lasso For VAR Shrinkage

Deborah Gefang*
Department of Economics
University of Lancaster
email: d.gefang@lancaster.ac.uk

April 17, 2013

Abstract

We develop a novel Bayesian doubly adaptive elastic-net Lasso (DAELasso) approach for VAR shrinkage. DAELasso achieves variable selection and coefficients shrinkage in a data based manner. It constructively deals with the explanatory variables that tend to be highly collinear by encouraging grouping effect. In addition, it allows for different degree of shrinkages for different coefficients. Rewriting the multivariate Laplace distribution as a scale mixture, we establish closed-form conditional posteriors that can be drawn from a Gibbs sampler. Empirical analysis shows that forecast results produced by DAELasso and its variants are comparable to that of other popular Bayesian methods, which provides further evidence that the forecast performances of large and medium sized Bayesian VARs are relatively robust to prior choices, and in practice simple Minnesota types of priors can be more attractive relative to their complex and well designed alternatives.

*I would like to thank Gary Koop, Esther Ruiz and two anonymous referees for their constructive comments. I would also like to thank the conference participants of CFE11, ESEM2012, and RCEF2012 for helpful discussions. Any remaining errors are my own responsibility.

1 Introduction

Tibshirani's (1996) least absolute shrinkage and selection operator (Lasso) and its variants, such as elastic net Lasso (e-net Lasso) of Zou and Hastie (2005), and grouped Lasso of Yuan and Lin (2007) are widely used for variable selection and parameter shrinkage for large data set. Recently, Bayesian Lasso has gained popularity as it can be easily implemented through MCMC or a Gibbs sampler (e.g. Park and Casella, 2008; Kyung et al, 2010), and it can automatically achieve adaptive shrinkage to allow for different degree of shrinkage (e.g. Griffin and Brown, 2010; Leng et al, 2010). Despite being successful, the Lasso literature is mainly concentrated on single equation models. To our best knowledge, only a few studies in the frequentist framework (e.g., Hsu, Hung and Chang, 2008; Song and Bickel, 2011) explore Lasso shrinkage for vector autoregressive (VAR) models. And these available methods can be too restrictive as they either assume the covariance matrix of the VAR errors to be diagonal or assume its off-diagonal elements are much smaller than the diagonal ones.

This paper develops a novel Bayesian Lasso method for VAR shrinkage. Considering large VAR models usually have highly correlated explanatory variables, we propose using doubly adaptive e-net Lasso (DAELasso) for macroeconomic research. DAELasso extends the adaptive e-net Lasso of Zou and Zhang (2009) for single equation models into VAR context. Unlike adaptive e-net Lasso that only adapts the tuning parameters of the L_1 norm, DAELasso allows for tuning parameters of both the L_1 and L_2 norms to be

adapted.[1] While Lasso generally only picks up one variable among a group of highly correlated variables, DAELasso has the potential of selecting all the important variables by encouraging grouping effects through e-net and adaptive shrinkage. Our scale mixture prior leads to closed-form conditional posteriors that can be directly drawn from a Gibbs sampler. Compared to its frequentist counterparts, DAELasso is more flexible as it does not need to impose unrealistic restrictions on the covariance matrix of the VAR errors. Hence it can better capture interdependencies between the variables. Considering that DAELasso can be too complicated for some data, in this paper, we also introduce four alternative Lasso types of VAR shrinkage methods: Lasso, adaptive Lasso, e-net Lasso, and adaptive e-net Lasso, each of them is a nested version of DAELasso.

Large Bayesian VARs are widely used for forecasting macroeconomic variables (e.g. Sims, 1972, 1980; Banbura et al, 2010). In empirical work, we evaluate the forecasting performance of DAELasso approach along with its variants. In addition, we compare the forecasting performance of these Lasso types of methods with that of the popular Baysian VAR shrinkage methods reviewed in Koop (2011). Those priors include the traditional Minnesota prior of Doan et al (1984) and Litterman (1986) and its natural variants (e.g. Kadiyala and Karlsson, 1997, Banbura et al, 2010), the stochastic search variable selection (SSVS) prior of George et al (2008), and the family of SSVS plus Minnesota priors of Koop (2011). We employ Koop's (2011) data

[1]Note that if we have formal and informal economic theory at hand to group the data, it can be more desirable to have other type of Lasso, such as the grouped Lasso, instead of e-net Lasso for VAR shrinkage. Yet in general we do not have such information, and e-net Lasso turns out to be the most appealing choice for it encourages the grouping effect supported by the data itself (Zou and Hastie, 2005).

set that contains 20 US macroeconomic series, which is originally compiled by James H. Stock and Mark W. Watson. The data runs from 1959Q1 to 2008Q4. In line with Koop (2011), we conduct rolling and recursive forecast exercises and calculate both the mean squared forecast error (MSFE) and predictive likelihood measures. Using relatively uninformative priors, we find DAELasso approach leads to forecasting results that compares favorably or equally well to other Bayesian VAR shrinkage methods. This suggests that DAELasso approach is an appropriate complement to the available Baysian VAR toolkit.

This paper is closely related to the growing literature on forecasting with many macroeconomic variables.[2] Popular methods in the existing literature include factor models (e.g. Stock and Watson, 2002a,b; Forni el al, 2000) and Bayesian shrinkage with a wide range of prior choices (e.g. De Mol et al, 2008; Banbura et al, 2010). De Mol et al (2008) and Koop (2011) find that these forecasts tend to be highly correlated, and Bayesian VARs do tend to forecast better than factor models. Our empirical results provide further evidence for the robustness of Bayesian shrinkage methods for forecasting with a large number of predictors. In addition, our findings suggest that, as stress in Koop (2011), so far the simple and less computationally costly Minnesota priors can be more attractive to practitioners as they forecast equally well or even better than their much complicated counterparts.

The remainder of the paper is organized as following. Section 2 develops the Bayesian DAELasso methods. Section 3 presents four alternative Lasso types of VAR shrinkage methods that nested in DAELasso. Section 4 present

[2] I thank a referee for pointing this out.

the empirical findings. Section 5 concludes. The data list and results for prior sensitivity analysis are provided in the appendix.

2 The Estimator and Bayesian Methods

Without loss of generality, we assume that all the variables are centered. Let Y be a $T \times N$ dependent variables, X be a $T \times Nk$ matrix contains the k lags of each dependent variable, and B be the coefficient matrix of dimension $Nk \times N$. In matrix form, an unrestricted VAR model of N variables takes the following form:

$$Y = XB + E \tag{1}$$

where E is a $T \times N$ matrix for *i.i.d.* error terms with its t^{th} row distributed as $\mathrm{N}(0, \Sigma)$.

Given the assumptions of the error term, the likelihood function of model (1) can be expressed as

$$L(b, \Sigma) \propto |\Sigma|^{-\frac{T}{2}} exp\{-\frac{1}{2} tr(Y - XB)^{'}(Y - XB)\Sigma^{-1}\} \tag{2}$$

Note that when $X^{'}X$ is not of full rank, which is often the case when we have more parameters than the number of observations, the least squares estimator $B_{LS} = (X^{'}X)^{+}X^{'}Y$ is noisy, where $(X^{'}X)^{+}$ is the Moore-Penrose generalized inverse of $X^{'}X$.

Vectorizing the matrices, we can transform model (1) into

$$y = (I_n \otimes X)\beta + e \tag{3}$$

where $y = vec(Y)$, $\beta = vec(B)$, $e = vec(E)$ and $e \sim N(0, \Sigma \otimes I_T)$. The dimension of β is $N^2k \times 1$.

We define the DAELasso estimator for a VAR as following:

$$\hat{\beta}_{dL} = arg\, min_{\beta}\{[y - (I_n \otimes X)\beta]'[y - (I_n \otimes X)\beta] + \sum_{j=1}^{N^2k} \lambda_{1,j}|\beta_j| + \sum_{j=1}^{N^2k} \lambda_{2,j}\beta_j^2\} \tag{4}$$

where $\lambda_{1,j}$ and $\lambda_{2,j}$, for $j = 1, 2, ..., N^2k$, are positive tuning parameters associated with the L_1 and L_2 penalties, respectively. We allow for different tuning parameters for different β_j to allow for different degree of shrinkages. For notational convenience, we define $\Lambda_1 = diag(\lambda_{1,1}, \lambda_{1,2}, ..., \lambda_{1,N^2k})$ and $\Lambda_2 = diag(\lambda_{2,1}, \lambda_{2,2}, ..., \lambda_{2,N^2k})$. Note that DAELasso defined in equation (4) is closely related to Zou and Zhang's (2009) adaptive e-net Lasso for single equations.

2.1 Priors

In the Bayesian framework, univariate Bayesian Laplace prior, which can be expressed as a scale mixtures of Normals with an exponential density (Andrews and Mallows, 1974), is widely used to enforce sparsity induced by the L_1 penalty in Lasso (e.g, Park and Casella, 2008; Leng et al, 2010; Korobilis, 2011). It is natural to consider extending the univariate Bayesian Laplace prior into multivariate analysis. However, this is not so straightforward. As noted by van Gerven et al (2009, 2010), the commonly used multivariate Laplace distributions (e.g, Kotz et al, 2001; Eltoft el al, 2006) generally do not factorize into a product of univariate Laplace distributions that can be associated with the individual coefficients.

Our approach is directly motivated by van Gerven et al's (2009, 2010) multivariate Laplace prior for single equation models. van Gerven et al (2009, 2010) use a scale mixture of Normals to reflect their prior knowledge of the interactions between the coefficients. Our scale mixture prior is similar to theirs, however, our prior is more about ensuring the priors associated with the L_1 norm are conditional on the unrestricted covariance matrix of the VAR errors. Conditioning on the covariance matrix of the VAR errors is important because otherwise the posterior may not be unimodal (Park and Casella, 2008). The posterior of van Gerven et al (2009, 2010) is not in a tractable form, and they use approximate inference methods for posterior computations. By contrast, our prior can lead to tractable conditional posteriors that can be directly drawn from Gibbs sampler.

We consider a conditional multivariate mixture prior of the following form:

$$\begin{aligned}\pi(\beta|\Sigma,\Gamma,\Lambda_1,\Lambda_2) \propto \prod_{j=1}^{N^2k} \{ & \frac{\sqrt{\lambda_{2,j}}}{\sqrt{2\pi}} exp(-\frac{\lambda_{2,j}}{2}\beta_j^2) \\ & \times \int_0^\infty \frac{1}{\sqrt{2\pi f_j(\Gamma))}} exp[-\frac{1}{2f_j(\Gamma)}\beta_j^2] d(f_j(\Gamma))\} \\ & \times \{|M|^{-\frac{1}{2}} exp(-\frac{1}{2}\Gamma' M^{-1}\Gamma)\}^2\end{aligned} \tag{5}$$

where $\Gamma = [\gamma_1, \gamma_2, ..., \gamma_{N^2k}]'$, $M = \Sigma \otimes I_{Nk}$, and $f_j(\Gamma)$ is a function of Γ and Λ_1 to be defined later. In this mixture prior, the terms associated with the L_1 penalty are conditional on Σ through $f_j(\Gamma)$. This is important as otherwise the posterior will be not unimodal due to the 'sharp corners' of the

L_1 penalty (Park and and Casella, 2008). In equation (5), the variances of β_a and β_b for $a \neq b$ are related through M. However, β_a and β_b themselves are independent of each other.

We need to find an appropriate $f_j(\Gamma)$ which provides us tractable posteriors. The last term in equation (5) takes the form of a multivariate Normal distribution $\Gamma \sim N(0, M)$. For ease of exposition, we first write the $N^2k \times N^2k$ covariance matrix M as following:

$$M = \begin{pmatrix} M_{1,1} & ... & M_{1,j} & M_{1,j+1} & ... & M_{1,N^2k} \\ ... & ... & ... & ... & ... & ... \\ M_{j,1} & ... & M_{j,j} & M_{j,j+1} & ... & M_{j,N^2k} \\ M_{j+1,1} & ... & M_{j+1,j} & M_{j+1,j+1} & ... & M_{j+1,N^2k} \\ ... & ... & ... & ... & ... & ... \\ M_{N^2k,1} & ... & M_{N^2k,j} & M_{N^2k,j+1} & ... & M_{N^2k,N^2k} \end{pmatrix} \tag{6}$$

Let $H_j = (M_{j,j+1}, ..., M_{j,N^2k}) \begin{pmatrix} M_{j+1,j+1} & ... & M_{j+1,N^2k} \\ ... & ... & ... \\ M_{N^2k,j+1} & ... & M_{N^2k,N^2k} \end{pmatrix}^{-1}$.

We next construct independent variables τ_j for $j = 1, 2, ..., N^2k$ using standard textbook techniques (e.g. Anderson, 2003; Muirhead 1982).

$$\tau_1 = \gamma_1 + H_1(\gamma_2, \gamma_3, ..., \gamma_{N^2k})' \tag{7}$$

$$\tau_2 = \gamma_2 + H_2(\gamma_3, \gamma_4, ..., \gamma_{N^2k})' \tag{8}$$

...

$$\tau_{N^2K-1} = \gamma_{N^2k-1} + H_{N^2k-1}\gamma_{N^2k} \tag{9}$$

$$\tau_{N^2K} = \gamma_{N^2k} \tag{10}$$

The joint density of $\tau_1, \tau_2, ..., \tau_{N^2k}$ is

$$N(\tau_1|0, \sigma^2_{\gamma_1})N(\tau_2|0, \sigma^2_{\gamma_2})...N(\tau_{N^2k}|0, \sigma^2_{\gamma_{N^2k}}) \tag{11}$$

where $\sigma^2_{\gamma_j} = M_{j,j} - H_j(M_{j,j+1}, ..., M_{j,N^2k})'$, with $\sigma^2_{\gamma_{N^2k}} = M_{N^2k,N^2k}$. Note that it is computationally feasible to derive $\sigma^2_{\gamma_j}$ when M is sparse.

The Jacobian of transforming $\Gamma \sim N(0, M)$ to (11) is 1. Defining $\eta_j = \tau_j/\lambda_{1,j}$, we can write (11) as

$$N(\eta_1|0, \sigma^2_{\gamma_1}\lambda^{-2}_{1,1})N(\eta_2|0, \sigma^2_{\gamma_2}\lambda^{-2}_{1,2})...N(\eta_{N^2k}|0, \sigma^2_{\gamma_{N^2k}}\lambda^{-2}_{1,N^2k}) \tag{12}$$

Let $f_j(\Gamma) = 2(\eta^2_j)$. Our scale mixture prior in (5) can be rewritten as:

$$\begin{aligned}\pi(\beta|\Sigma, \Gamma, \Lambda_1, \Lambda_2) \propto \prod_{j=1}^{N^2k} \{ & \frac{\sqrt{\lambda_{2,j}}}{\sqrt{2\pi}} exp(-\frac{\lambda_{2,j}}{2}\beta^2_j) \\ & \times \int_0^\infty \frac{1}{\sqrt{2\pi(2\eta^2_j))}} exp[-\frac{\beta^2_j}{2(2\eta^2_j)}] d(2\eta^2_j) \\ & \times \frac{\lambda^2_{1,j}}{2\sigma^2_{\gamma_j}} exp[-\frac{1}{2}\frac{2\eta^2_j}{(\sigma^2_{\gamma_j})/\lambda^2_{1,j}}]\}\end{aligned} \tag{13}$$

The last two terms in (13) constitute a scale mixture of Normals (with an exponential mixing density), which can be expressed as the univari-

ate Laplace distribution $\frac{\lambda_{1,j}}{2\sqrt{\sigma^2_{\gamma_j}}} exp(-\frac{\lambda_{1,j}}{\sqrt{\sigma^2_{\gamma_j}}}|\beta_j|)$. Readers familiar with the Bayesian Lasso literature can easily find that this is an adaptive version of the Laplace prior of Park and Casella (2008) that enforces the $L1$ penalty. Overall, the priors in equation (13) resembles the conditional prior for the e-net Lasso used in Kyung et al (2010). However, our prior is more complicated than that of Kyung et al (2010). First, we allow different levels of shrinkages for all the coefficients.[3] Second, the prior distribution of the coefficients are conditional on the covariance matrix of the multivariate model.

Equation (13) shows that the conditional prior for β_j is $N(0, \frac{2\eta_j^2}{2\lambda_{2,j}\eta_j^2+1})$, and the conditional prior for β is

$$\beta|\Gamma, \Sigma, \Lambda_1, \Lambda_2 \sim N(0, D^*_\Gamma) \tag{14}$$

where $D^*_\Gamma = diag([\frac{2\eta_1^2}{2\lambda_{2,1}\eta_1^2+1}, \frac{2\eta_2^2}{2\lambda_{2,2}\eta_2^2+1}, ..., \frac{2\eta^2_{N^2k}}{2\lambda_{2,N^2k}\eta^2_{N^2k}+1}])$. The tightness of the prior for each β_j depends on $\frac{2\eta_j^2}{2\lambda_{2,j}\eta_j^2+1}$. If $\frac{2\eta_j^2}{2\lambda_{2,j}\eta_j^2+1}$ is small, β_j will be shrunk towards zero. If $\frac{2\eta_j^2}{2\lambda_{2,j}\eta_j^2+1}$ is large, the prior for β_j can become quite uninformative.

Priors for Σ, $\lambda^2_{1,j}$ and $\lambda_{2,j}$ can be elicited following standard practice in VAR and Lasso literature. In this paper, we set Wishart prior for Σ^{-1} and Gamma priors for $\lambda^2_{1,j}$ and $\lambda_{2,j}$: $\Sigma^{-1} \sim W(\underline{S}^{-1}, \underline{\nu})$, $\lambda^2_{1,j} \sim G(\underline{\mu}_{\lambda^2_{1,j}}, \underline{\nu}_{\lambda^2_{1,j}})$, $\lambda_{2,j} \sim G(\underline{\mu}_{\lambda_{2,j}}, \underline{\nu}_{\lambda_{2,j}})$.[4]

[3] In the literature, Griffin and Brown (2010) explore using different priors to achieve adaptive Lasso which automatically adapt. Leng et al (2010) propose using varying degree of shrinkage for the tuning parameter. Our adaptive approach is in spirit of Leng et al (2010).

[4] Please refer to Koop (2003), p326, for Gamma distribution, and Zellner (1971), p389, for Wishart distribution.

2.2 Posteriors and Gibbs Sampler

Combining the priors and likelihood, the following full conditional posteriors can be easily derived.

The full conditional posterior for β is $\beta \sim N(\overline{\beta}, \overline{V}_\beta)$, where $\overline{V}_\beta = [(I_N \otimes X)^{'})(\Sigma^{-1} \otimes I_{Nk})(I_N \otimes X) + (D_\Gamma^*)^{-1}]^{-1}$, and $\overline{\beta} = \overline{V}_\beta[(I_N \otimes X)^{'}(\Sigma^{-1} \otimes I_{Nk})y]$. The Full conditional posterior for Σ^{-1} is $W(\overline{S}^{-1}, \overline{\nu})$, with $\overline{S}^{-1} = (Y - XB)^{'}(Y - XB) + 2Q^{'}Q + \underline{S}^{-1}$ and $\overline{\nu} = T + 2Nk + \underline{\nu}$, with $vec(Q) = \Gamma$. The Full conditional posterior for $\lambda_{1,j}^2$ is $G(\bar{\mu}_{\lambda_{1,j}}, \bar{\nu}_{\lambda_{1,j}})$, where $\bar{\nu}_{\lambda_{1,j}} = \underline{\nu}_{\lambda_{1,j}} + 2$ and $\bar{\mu}_{\lambda_{1,j}} = \frac{\bar{\nu}_{\lambda_{1,j}} \sigma_j^2 \underline{\mu}_{\lambda_{1,j}}}{2\tau_j^2 \underline{\mu}_{\lambda_{1,j}} + \underline{\nu}_{\lambda_{1,j}} \sigma_{\gamma_j}^2}$. The Full conditional posterior for $\lambda_{2,j}$ is $G(\bar{\mu}_{\lambda_{2,j}}, \bar{\nu}_{\lambda_{2,j}})$, where $\bar{\nu}_{\lambda_{2,j}} = \underline{\nu}_{\lambda_{2,j}} + 1$ and $\bar{\mu}_{\lambda_{2,j}} = \frac{\underline{\mu}_{\lambda_{2,j}} \bar{\nu}_{\lambda_{2,j}}}{\underline{\nu}_{\lambda_{2,j}} + \underline{\mu}_{\lambda_{2,j}} \beta_j^2}$. Finally the full conditional posterior of $\frac{1}{2\eta_j^2}$ is Inverse Gaussian: $IG(\sqrt{\frac{\lambda_{1,j}^2}{\beta_j^2 \sigma_{\gamma_j}^2}}, \frac{\lambda_{1,j}^2}{\sigma_{\gamma_j}^2})$.[5] Γ can not be directly drawn from the posteriors. But it can be recovered in each Gibbs iteration using the draws of $\frac{1}{2\eta_j^2}$ and Σ .

Conditional on arbitrary starting values, the Gibbs sampler contains the following six steps:

1. draw $\beta|\Sigma, \Lambda_1, \Lambda_2, \Gamma$ from $N(\overline{\beta}, \overline{V}_\beta)$;
2. draw $\Sigma^{-1}|\beta, \Lambda_1, \Lambda_2, \Gamma$ from $W(\overline{S}^{-1}, \overline{\nu})$
3. draw $\lambda_{1,j}^2|\beta, \Sigma, \Lambda_{1,-j}, \Lambda_2, \Gamma$ from $G(\bar{\mu}_{\lambda_{1,j}}, \bar{\nu}_{\lambda_{1,j}})$ for $j = 1, 2, ...N^2k$
4. draw $\lambda_{2,j}|\beta, \Sigma, \Lambda_1, \Lambda_{2,-j}, \Gamma$ from $G(\bar{\mu}_{\lambda_{2,j}}, \bar{\nu}_{\lambda_{2,j}})$ for $j = 1, 2, ...N^2k$
5. draw $\frac{1}{2\eta_j^2}|\beta, \Sigma, \Lambda_1, \Lambda_2$ from $IG(\sqrt{\frac{\lambda_{1,j}^2}{\beta_j^2 \sigma_{\gamma_j}^2}}, \frac{\lambda_{1,j}^2}{\sigma_{\gamma_j}^2})$ for $j = 1, 2, ...N^2k$.

[5] We adopt the same form of the inverse-Gaussian density used in Park and Casella (2008).

6. calculate Γ based on draws of Σ and $\frac{1}{2\eta_j^2}$ in the current iteration.

3 Related Lasso Types of VAR Shrinkage

DAELasso provides a general method to shrink both the variable and parameter space of a VAR. However, with the number of tuning parameters two times the number of coefficients, DAELasso might be subject to the criticism of demanding too much from the data. In this section, we introduce four alternative scaled mixture priors for β that respectively associated with Lasso, adaptive Lasso, e-net Lasso, and adaptive e-net Lasso. Note that these four Lassos are all nested in DAELasso. Thus their posteriors can be easily worked out using the procedures presented for DAELasso shrinkage. For brevity, we relegate the technical details to the online appendix.

3.1 Lasso VAR Shrinkage

Following Song and Bickel (2011), we define Lasso estimator for a VAR as:

$$\hat{\beta}_L = arg\, min_\beta \{ [y - (I_n \otimes X)\beta]' [y - (I_n \otimes X)\beta] + \lambda_1 \sum_{j=1}^{N^2 k} |\beta_j| \} \tag{15}$$

Correspondingly, the conditional multivariate mixture prior for β takes the following form:

$$\begin{aligned} \pi(\beta|\Sigma, \Gamma, \lambda_1) \propto \prod_{j=1}^{N^2 k} \{ \int_0^\infty \frac{1}{\sqrt{2\pi f_j(\Gamma))}} exp[-\frac{1}{2 f_j(\Gamma)} \beta_j^2] d(f_j(\Gamma)) \} \\ \times \{ |M|^{-\frac{1}{2}} exp(-\frac{1}{2} \Gamma' M^{-1} \Gamma) \}^2 \end{aligned} \tag{16}$$

Let $f_j(\Gamma) = 2(\eta_j^2)$, the scale mixture prior is:

$$\begin{aligned}\pi(\beta|\Sigma,\Gamma,\lambda_1) \propto \prod_{j=1}^{N^2k} \{ \int_0^\infty \frac{1}{\sqrt{2\pi(2\eta_j^2))}} exp[-\frac{\beta_j^2}{2(2\eta_j^2)}] d(2\eta_j^2) \\ \times \frac{\lambda_1^2}{2\sigma_{\gamma_j}^2} exp[-\frac{1}{2}\frac{2\eta_j^2}{(\sigma_{\gamma_j}^2)/\lambda_1^2}]\}\end{aligned} \tag{17}$$

where $\eta_j = \tau_j/\lambda_1$.

3.2 Adaptive Lasso VAR Shrinkage

We define the adaptive Lasso estimator for a VAR as:

$$\hat{\beta}_{AL} = arg\, min_\beta \{ [y - (I_n \otimes X)\beta]' [y - (I_n \otimes X)\beta] + \sum_{j=1}^{N^2k} \lambda_{1,j} |\beta_j| \} \tag{18}$$

Correspondingly, the conditional multivariate mixture prior for β takes the following form:

$$\begin{aligned}\pi(\beta|\Sigma,\Gamma,\Lambda_1) \propto \prod_{j=1}^{N^2k} \{ \int_0^\infty \frac{1}{\sqrt{2\pi f_j(\Gamma))}} exp[-\frac{1}{2f_j(\Gamma)}\beta_j^2] d(f_j(\Gamma)) \} \\ \times \{ |M|^{-\frac{1}{2}} exp(-\frac{1}{2}\Gamma' M^{-1}\Gamma) \}^2\end{aligned} \tag{19}$$

Let $f_j(\Gamma) = 2(\eta_j^2)$, the scale mixture prior is:

$$\begin{aligned}\pi(\beta|\Sigma,\Gamma,\Lambda_1) \propto \prod_{j=1}^{N^2k} \{ \int_0^\infty \frac{1}{\sqrt{2\pi(2\eta_j^2))}} exp[-\frac{\beta_j^2}{2(2\eta_j^2)}] d(2\eta_j^2) \\ \times \frac{\lambda_{1,j}^2}{2\sigma_{\gamma_j}^2} exp[-\frac{1}{2}\frac{2\eta_j^2}{(\sigma_{\gamma_j}^2)/\lambda_{1,j}^2}]\}\end{aligned} \tag{20}$$

where $\eta_j = \tau_j/\lambda_{1,j}$.

3.3 E-net Lasso VAR Shrinkage

We define the e-net Lasso estimator for a VAR as:

$$\hat{\beta}_{EL} = arg\, min_{\beta}\{[y-(I_n \otimes X)\beta]'[y-(I_n \otimes X)\beta] + \lambda_1 \sum_{j=1}^{N^2k} |\beta_j| + \lambda_2 \sum_{j=1}^{N^2k} \beta_j^2\} \tag{21}$$

Correspondingly, the conditional multivariate mixture prior for β takes the following form:

$$\begin{aligned} \pi(\beta|\Sigma,\Gamma,\lambda_1,\lambda_2) \propto \prod_{j=1}^{N^2k} \{ & \frac{\sqrt{\lambda_2}}{\sqrt{2\pi}} exp(-\frac{\lambda_2}{2}\beta_j^2) \\ & \times \int_0^{\infty} \frac{1}{\sqrt{2\pi f_j(\Gamma))}} exp[-\frac{1}{2f_j(\Gamma)}\beta_j^2] d(f_j(\Gamma))\} \\ & \times \{|M|^{-\frac{1}{2}} exp(-\frac{1}{2}\Gamma' M^{-1}\Gamma)\}^2 \end{aligned} \tag{22}$$

Let $f_j(\Gamma) = 2(\eta_j^2)$, the scale mixture prior is:

$$\begin{aligned} \pi(\beta|\Sigma,\Gamma,\lambda_1,\lambda_2) \propto \prod_{j=1}^{N^2k} \{ & \frac{\sqrt{\lambda_2}}{\sqrt{2\pi}} exp(-\frac{\lambda_2}{2}\beta_j^2) \\ & \times \int_0^{\infty} \frac{1}{\sqrt{2\pi(2\eta_j^2))}} exp[-\frac{\beta_j^2}{2(2\eta_j^2)}] d(2\eta_j^2) \\ & \times \frac{\lambda_1^2}{2\sigma_{\gamma_j}^2} exp[-\frac{1}{2}\frac{2\eta_j^2}{(\sigma_{\gamma_j}^2)/\lambda_1^2}]\} \end{aligned} \tag{23}$$

where $\eta_j = \tau_j/\lambda_1$.

3.4 Adaptive E-net Lasso VAR Shrinkage

In line with Zou and Zhang (2009), we define the adaptive e-net Lasso estimator for a VAR as following:

$$\hat{\beta}_{AEL} = arg\, min_{\beta}\{[y-(I_n \otimes X)\beta]'[y-(I_n \otimes X)\beta] + \sum_{j=1}^{N^2k} \lambda_{1,j}|\beta_j| + \lambda_2 \sum_{j=1}^{N^2k} \beta_j^2\} \tag{24}$$

Correspondingly, the conditional multivariate mixture prior for β takes the following form:

$$\begin{aligned} \pi(\beta|\Sigma,\Gamma,\Lambda_1,\lambda_2) \propto \prod_{j=1}^{N^2k} \{ & \frac{\sqrt{\lambda_2}}{\sqrt{2\pi}} exp(-\frac{\lambda_2}{2}\beta_j^2) \\ & \times \int_0^{\infty} \frac{1}{\sqrt{2\pi f_j(\Gamma))}} exp[-\frac{1}{2f_j(\Gamma)}\beta_j^2] d(f_j(\Gamma))\} \\ & \times \{|M|^{-\frac{1}{2}} exp(-\frac{1}{2}\Gamma' M^{-1}\Gamma)\}^2 \end{aligned} \tag{25}$$

Let $f_j(\Gamma) = 2(\eta_j^2)$. The scale mixture prior in (25) can be rewritten as:

$$\begin{aligned} \pi(\beta|\Sigma,\Gamma,\Lambda_1,\lambda_2) \propto \prod_{j=1}^{N^2k} \{ & \frac{\sqrt{\lambda_2}}{\sqrt{2\pi}} exp(-\frac{\lambda_2}{2}\beta_j^2) \\ & \times \int_0^{\infty} \frac{1}{\sqrt{2\pi(2\eta_j^2))}} exp[-\frac{\beta_j^2}{2(2\eta_j^2)}] d(2\eta_j^2) \\ & \times \frac{\lambda_{1,j}^2}{2\sigma_{\gamma_j}^2} exp[-\frac{1}{2}\frac{2\eta_j^2}{(\sigma_{\gamma_j}^2)/\lambda_{1,j}^2}]\} \end{aligned} \tag{26}$$

where $\eta_j = \tau_j/\lambda_{1,j}$.

4 Empirical Illustration

4.1 Data

In macroeconomics, it is a standard practice to assess models by their forecasting performance (e.g. Litterman, 1986; Giannone et al, 2010). Koop (2011) provides an extensive forecasts evaluation for seven popular Bayesian VAR priors. We employ the data set of Koop (2011), an updated version of that used in Stock and Watson (2008), for the out-of sample forecasting analysis. The data set contains twenty quarterly macroeconomic series including a measure of economic activity (GDP, real GDP), prices (CPI, the consumer price index), an interest rate (FFR, the Fed funds rate), and other seventeen variables.[6] Four of the seventeen variables are those used in the monetary model of Christiano el al (1999). The rest of the thirteen variables contain important aggregated information of the economy. The time series span from 1959Q1 to 2008Q4. A full list of the variables is provided in Appendix A. Detailed data descriptions please refer to Koop (2011) and Stock and Watson (2008). Data are transformed to stationarity and standardized same as Koop (2011).[7]

[6]These 20 variables are used for medium-size VAR in Koop (2011). Koop (2011) also examines the VAR forecasts using medium-large VAR, which contains 40 variables, and large VAR, which contains 168 variables. We only focus on Koop's (2011) medium-size VAR in this paper due to two considerations. First, it is computationally costly to use DAELasso priors to estimate the medium-large and large VARs. Second, it is shown in the literature (e.g. Banbura et al, 2010; Koop, 2011) that most of the gains in forecasting performance are achieved by using medium VARs of about 20 key variables.

[7]I am grateful to Mark Watson for providing the data. In addition, I am grateful to Gary Koop for sharing the Matlab code for data transformation.

4.2 Forecast Evaluation

Same as Koop (2011), we conduct rolling and recursive forecast exercises and calculate both the mean squared forecast error (MSFE) and predictive likelihood measures using reduced form VAR of order four. The window length for the rolling estimation is set to be ten years. Recursive and rolling forecasts are conducted for t_0+h, t_0+1+h,...T, where t_0 is 1969Q4. Let y^f_{t+h} be the h^{th} period forecast of y using data available at time t, and y_{t+h} be the real value for y observed at $t+h$. The MSFE measure for the variable y_i is calculated as an average of the mean squared errors of the point estimates:

$$MSFE = \frac{\sum_{t=t_0}^{T-h}[y_{i,t+h} - E(y^f_{i,t+h}|Data_t)]^2}{T-h-t_0+1} \tag{27}$$

The predictive likelihood is used to evaluate the entire predictive distribution. In particular, the following sum of the log predictive likelihood is used:

$$\sum_{t=t_0}^{T-h} log[p(y^f_{i,t+h} = y_{i,t+h}|Data_t)] \tag{28}$$

For DAELasso, we need to elicit priors for $\lambda^2_{1,j}$, $\lambda_{2,j}$, and Σ. It is practically impossible to set informative priors for each $\lambda^2_{1,j}$ and $\lambda_{2,j}$, thus we set relatively uninformative priors for $\lambda^2_{1,j}$ (or λ^2_1) and $\lambda_{2,j}$ (or λ_2) to be $G(1, 0.001)$ and $G(1, 0.01)$, respectively. The prior for Σ^{-1} is set to be $W((N-1)I_N, 1)$, which is also relatively uninformative. There is room for improving the forecasting performance of DAELasso, such as by eliciting more informative priors. We do not explore this possibility in the current paper because the goal of our exercise is to find out whether a DAELasso

with relatively uninformative priors can provide acceptable forecasting results. For comparison, the priors for Lasso, adaptive Lasso, e-net Lasso, and adaptive e-net Lasso are set in the same manner. Following Koop (2011) and Geweke and Amisano (2011), we calculate log predictive likelihood in each replication of the Gibbs sampler using posterior draws of the parameters, then take the average when the Gibbs ends. For comparison, we also use the mean posterior parameters to calculate the log predictive likelihood. It turns out that the results are similar.

To save space, we relegate a comprehensive comparison between the forecasts of DAELasso and its variants and that of the priors used in Koop (2011) into the on-line appendix. In this section, we only report the forecasts results for the five most important priors: DAELasso, adaptive e-net Lasso, adaptive Lasso, the natural conjugate prior used in Banbura et al (2010), which is labelled 'Minn. Prior as in BGR', and a combination of the conjugate SSVS prior and Minnesota prior, which is labelled 'SSVS Conjugate plus Minn. Prior' proposed by Koop (2011). The last two priors are investigated in Koop (2011). We refer to Koop (2011) for a lucid description of these priors.

Results presented in Tables 1-4 show that the forecasting performance of DAELasso approach is comparable to that of the popular Bayesian shrinkage methods explored in Koop (2011). Overall, the forecasts of all methods are highly correlated with each other. Compared to the priors investigated in Koop (2011), DAELasso and its variants tend to forecast better for GDP and CPI in terms of point forecasts, but not as well for FFR. The results for predictive loglikelihood yielded by DAELasso approach are more mixed. Yet,

Table 1: Rolling Forecasting for $h = 1$

	GDP	CPI	FFR
DAELasso	0.58 (-198.9)	0.32 (-192.7)	0.57 (-211.7)
adaptive e-net Lasso	0.67 (-195.8)	0.40 (-199.4)	0.63 (-215.0)
adaptive Lasso	0.77 (-225.6)	0.31 (-209.2)	0.62 (-228.3)
Minn. Prior as in BGR	0.58 (-190.5)	0.34 (-209.2)	0.51 (-177.4)
SSVS Non-conj. plus Minn. Prior	0.68 (-197.9)	0.34 (-195.2)	0.52 (-177.2)

Notes:

MSFEs as proportion of random walk MSFEs.

Sum of log predictive likelihoods in parentheses.

Table 2: Rolling Forecasting for $h = 4$

	GDP	CPI	FFR
DAELasso	0.55 (-206.9)	0.48 (-205.9)	0.65 (-230.9)
adaptive e-net Lasso	0.53 (-195.7)	0.47 (-204.4)	0.55 (-219.9)
adaptive Lasso	0.74 (-233.9)	0.54 (-223.0)	0.78 (-247.7)
Minn. Prior as in BGR	0.59 (-217.1)	0.55 (-227.7)	0.59 (-213.4)
SSVS Non-conj. plus Minn. Prior	0.63 (-209.9)	0.51 (-201.3)	0.58 (-198.1)

Notes:

MSFEs as proportion of random walk MSFEs.

Sum of log predictive likelihoods in parentheses.

Table 3: Recursive Forecasting for $h = 1$

	GDP	CPI	FFR
DAELasso	0.55	0.29	0.56
	(-210.4)	(-190.9)	(-224.2)
adaptive e-net Lasso	0.67	0.40	0.63
	(-242.0)	(-201.6)	(-239.8)
adaptive Lasso	0.62	0.28	0.60
	(-219.2)	(-196.4)	(-226.8)
Minn. Prior as in BGR	0.56	0.30	0.51
	(-192.3)	(-195.9)	(-229.1)
SSVS Non-conj. plus Minn. Prior	0.65	0.29	0.54
	(-203.9)	(-187.6)	(-228.9)

Notes:

MSFEs as proportion of random walk MSFEs.

Sum of log predictive likelihoods in parentheses.

Table 4: Recursive Forecasting for $h = 4$

	GDP	CPI	FFR
DAELasso	0.54	0.48	0.61
	(-218.3)	(-206.6)	(-239.6)
adaptive e-net Lasso	0.53	0.47	0.55
	(-215.6)	(-207.0)	(-247.3)
adaptive Lasso	0.63	0.52	0.66
	(-228.0)	(-214.7)	(-242.2)
Minn. Prior as in BGR	0.61	0.52	0.59
	(-214.7)	(-219.4)	(-249.6)
SSVS Non-conj. plus Minn. Prior	0.67	0.49	0.53
	(-219.0)	(-201.6)	(-233.7)

Notes:

MSFEs as proportion of random walk MSFEs.

Sum of log predictive likelihoods in parentheses.

DAELasso's performances are qualitatively similar to that of the Bayesian VAR methods studied in Koop (2011).

When we focus on the three Lasso types of forecasts, we find for $h = 1$, DAELasso tends to provide the best GDP and FFR point forecasts while adaptive Lasso gives the best CPI point forecasts. This results seem to suggest that while GDP and FFR can be better forecasted by using many variables that might be highly collinear, CPI can be better forecasted using a smaller number of important variables that are not highly correlated with each other. Turning to the point forecasts for $h = 4$, we find adaptive e-net Lasso are performing better in most of the cases. Forecasts results for predictive loglikelihood varies for the three Lasso types of methods. But overall, DAELasso and adaptive e-net Lasso tend to outperform adaptive Lasso in most of the cases. This result suggests the importance of using e-net to capture the possible grouping effect.

We conduct prior sensitivity analysis using relatively tighter priors. The results are reported in Appendix B. We find that the tighter priors tend to give better forecasts. Yet, the more important information from the exercise is that the general pattern found in Tables 1-4 is robust to the prior choices.

4.3 Sparsity or Stability?

As noted in De Mol et al (2008), in presence of collinearity, if a shrinkage method only enforces sparsity, like Lasso, it will be very sensitive to minor perturbations of the data. Traditional wisdom is to use e-net Lasso to enforce both $L1$ and $L2$ penalties (e.g. Zou and Hastie, 2005). Yet, it is less clear if using e-net Lasso is the most effective way to improve the stability of

variable selections. In this paper, we look into this question by examining the posterior mean estimates for DAELasso and its variants. We find that using similar priors, e-net Lasso tends to enforce slightly less sparsity than Lasso. Yet, DAELasso, adaptive e-net Lasso and adaptive Lasso tend to hugely improve the stability of variable selection by pulling the parameters for all the highly collinear variables towards zero. Our findings suggest that, in terms of enforcing stability, adaptive shrinkage is more effective than e-net.[8]

Tables 5-7 present the correlations between the posterior mean estimates for the coefficients in the GDP, CPI, and FFR equations for DAELAsso and its variants. Here coefficients for all the recursive estimations are pooled together.[9] Interestingly, we find that in all cases, the coefficients for the adaptive types of models are highly correlated, and the coefficients for e-net Lasso and Lasso are highly correlated. This implies that the degree of sparsity enforced are more affected by whether adaptive shrinkage is used.

Table 5: Correlations Between the Coefficients in the GDP Equation

	DAELasso	adaptive e-net Lasso	adaptive Lasso	e-net Lasso	Lasso
DAELasso	1.00				
adaptive e-net Lasso	0.97	1.00			
adaptive Lasso	0.97	0.97	1.00		
e-net Lasso	0.70	0.70	0.69	1.00	
Lasso	0.66	0.65	0.66	0.97	1.00

Table 8 presents the means and standard deviations for the posterior es-

[8]Our finding that adaptive types of Lasso models tend to give more stable variable selection results, however, is based on the same or similar priors we elicited for the shrinkage parameters across models. This scenario can be changed if the priors for the shrinkage parameters of different models are very different.

[9]Results for rolling estimates are very similar.

Table 6: Correlations Between the Coefficients in the CPI Equation

	DAELasso	adaptive e-net Lasso	adaptive Lasso	e-net Lasso	Lasso
DAELasso	1.00				
adaptive e-net Lasso	0.97	1.00			
adaptive Lasso	0.97	0.97	1.00		
e-net Lasso	0.73	0.74	0.73	1.00	
Lasso	0.69	0.69	0.69	0.96	1.00

Table 7: Correlations Between the Coefficients in the FFR Equation

	DAELasso	adaptive e-net Lasso	adaptive Lasso	e-net Lasso	Lasso
DAELasso	1.00				
adaptive e-net Lasso	0.97	1.00			
adaptive Lasso	0.98	0.97	1.00		
e-net Lasso	0.74	0.76	0.74	1.00	
Lasso	0.71	0.71	0.71	0.97	1.00

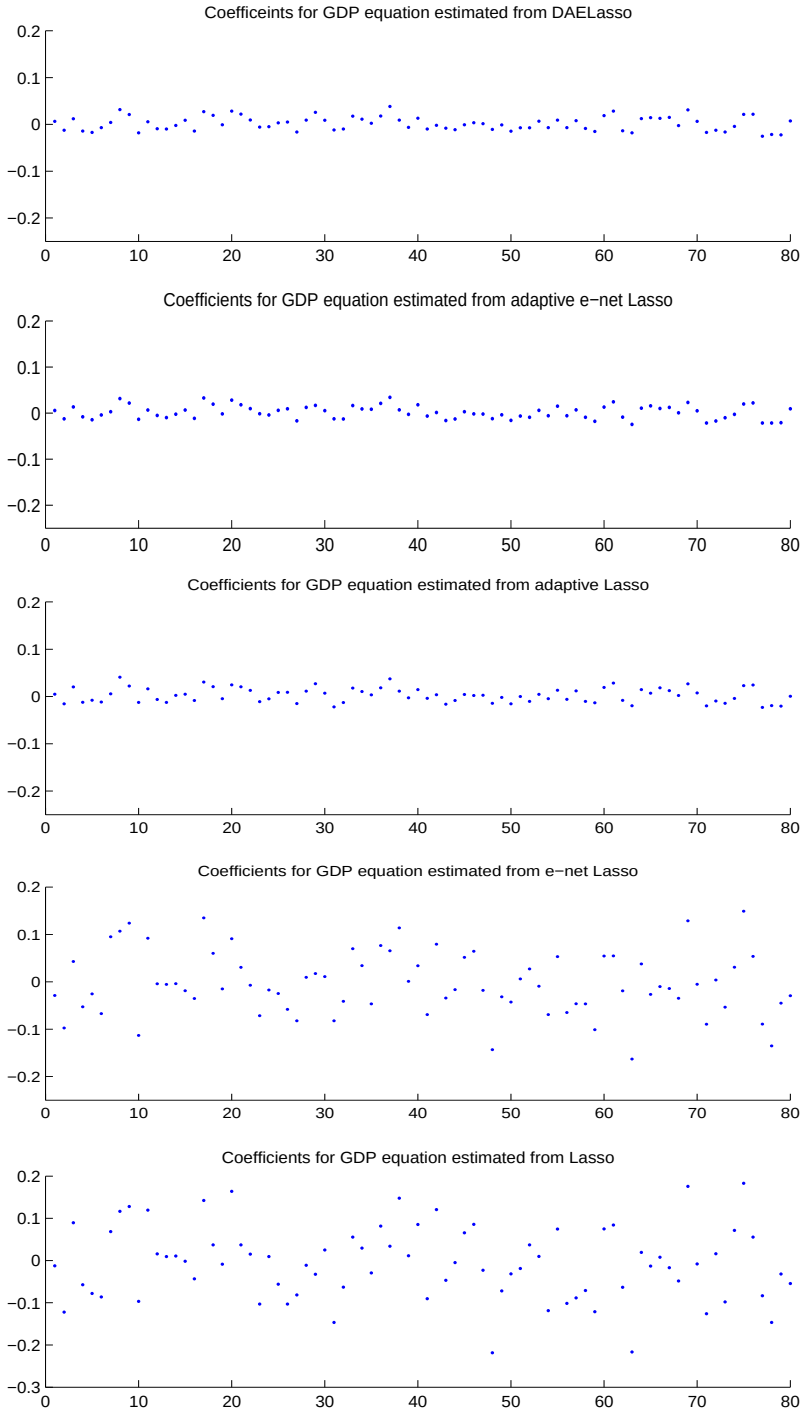
timates for the coefficients in the GDP, CPI, and FFR equations. In general, results for the three adaptive types of Lassos are very similar, results for e-net Lasso and Lasso are very similar. A closer look reveals that compared to Lasso, e-net Lasso yields posterior estimates that are slightly more centered at zero. However, the most prominent pattern that leaps out of Table 8 is that the standard deviations for the coefficients in e-net Lasso and Lasso are almost 10 times lager than their counterparts in the three adaptive types of models.

Table 8: Means and Standard Deviations for the Coefficients

		DAELasso	adaptive e-net Lasso	adaptive Lasso	e-net Lasso	Lasso
GDP	mean	0.0021	0.0020	0.0021	-0.0013	-0.0029
	std.	(0.0191)	(0.0180)	(0.0191)	(0.1032)	(0.1267)
CPI	mean	0.0013	0.0012	0.0013	0.0000	-0.0013
	std.	(0.0193)	(0.0184)	(0.0193)	(0.0799)	(0.0943)
FFR	mean	0.0053	0.0053	0.0054	0.0130	0.0150
	std.	(0.0222)	(0.0207)	(0.0222)	(0.0939)	(0.1062)

Figure 1 plots the posterior estimates of the coefficients in the GDP equation for the whole sample as an example. The coefficients are plotted against the sequence number of the variables. The similarities between the coefficients plots for DAELasso, adaptive e-net Lasso, and adaptive Lasso are striking. Same as the similarity between the coefficients plots for e-net Lasso and Lasso. As the coefficients estimated using the adaptable types of Lassos are much closer to zero than their counterparts estimated using e-net Lasso and Lasso, it is understandable that a minor perturbation in data tends to have more impact on the e-net Lasso and Lasso forecasts. For brevity, we do not present comparable figures for other variables and other sample periods, but they all support our general findings.

Figure 1


Coefficeints for GDP equation estimated from DAELasso
Coefficients for GDP equation estimated from adaptive e−net Lasso
Coefficients for GDP equation estimated from adaptive Lasso
Coefficients for GDP equation estimated from e−net Lasso
Coefficients for GDP equation estimated from Lasso

5 Conclusion

This paper proposes a Bayesian DAELasso approach for VAR shrinkage. We elicit a scale mixture prior which leads to closed-form conditional posteriors that can be directly drawn from a Gibbs sampler. The method is appealing as it can simultaneously achieve variable selection and coefficient shrinkage in a data based fashion. DAELasso constructively deals with multicollinearity problem by encouraging the grouping effect through both the e-net and adaptive shrinkages. Hence, its forecasts results can be more stable and less subject to the influence of minor data perturbations. Using relatively uninformative prior, we find that the forecasting performance of DAELasso is comparable to that of other popular Bayesian VAR shrinkage methods. This shows that DAELasso approach can be used as an appropriate addition to the available Bayesian VAR toolkits. The implementation of DAELasso approach is simple and straightforward. It can be easily extended into nonlinear framework to shed new light on macro economic analysis and forecasting.

It is interesting to find that the sophisticated and well designed priors we used for DAELasso and its variants produce forecasts that are correlated and equally accurate of those obtained with simple Minnesota prior and conjugate priors traditionally used to conduct inference in large and medium sized Bayesian VARs. This result has important implications in a broader context. First, it provides further evidence that good forecasting performances of Bayesian VARs are a general feature of the data and do not depend on specific features of the prior. Second, it confirms Koop's

(2011) finding that simple Minnesota priors such as the Normal-conjugate prior used by Banbura et al (2010) work well in medium and large VARs, which makes these simple priors attractive relative to the computationally more demanding alternatives.

References

[1] Anderson, T. W. (2003), An introduction to multivariate statistical analysis, *3rd Ed.* John Wiley and Sons: New Jersey.

[2] Andrews, D. F. and C. L. Mallows (1974), Scale mixtures of Normal distributions, *Journal of the Royal Statistical Society*, B 36, 99-102.

[3] Banbura, M, D. Giannone and L. Reichlin (2010), Large Bayesian vector auto regressions, *Journal of Applied Econometrics*, 25, 71-92.

[4] Christiano, L. J., M. Eichenbaum and C. L. Evans (1999), Monetary policy shocks: what have we learned and to what end? in Taylor J. B. and M. Woodford (*eds*) Handbook of Macroeconomics 1, 65-148, Elsevier: Amsterdam.

[5] De Mol, C., Giannone, D. and Reichlin, L. (2008), Forecasting using a large number of predictors: is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146, 318-28.

[6] Doan, T., R. Litterman and C. Sims (1984), Forecasting and conditional projections using realistic prior distributions, *Econometric Reviews*, 3, 1-100.

[7] Eltoft, T., T. Kim, and T. Lee (2006), On the multivariate Laplace distribution, *IEEE Signal Proc. Let.* 13 (5), 300-3.

[8] Forni, M., M. Hallin, M. Lippi and L. Reichlin (2000), The generalized dynamic factor model: Identification and estimation, *Review of Economics and Statistics*, 82, 540-54.

[9] Geweke J and J. Amisano (2011), Hierarchical Markov normal mixture models with applications to financial asset returns. Journal of Applied Econometrics 26: 129

[10] George, E., D. Sun and S. Ni (2008), Bayesian stochastic search for VAR model restrictions, *Journal of Econometrics*, 142, 553-80.

[11] Giannone, M., M. Lenza and G. Primiceri (2010), Prior selection for vector autoregressions, Working paper, Université Libre de Bruxelles.

[12] Griffin, J. E. and P. J. Brown (2010), Bayesian adaptive lassos with non-convex penalization, *Australian and New Zealand Journal of Statistics, forthcoming.*

[13] Hsu, N-J, H-L. Hung and Y-M. Chang (2008), Subset selection for vector autoregressive processes using Lasso, *Computational Statistics & Data Analysis*, 52, 3645-57.

[14] Kadiyala, K.R. and S. Karlsson (1997), Numerical methods for estimation and inference in Bayesian VAR-models, *Journal of Applied Econometrics*, 12(2), 99-132.

[15] Koop, G. (2003), Bayesian econometrics, John Wiley and Sons: Chichester.

[16] Koop, G. (2011), Forecasting with medium and large Bayesian VARs, *Journal of Applied Econometrics*, doi: 10.1002/jae.1270.

[17] Korobilis, D. (2011), Hierarchical shrinkage priors for dynamic regressions with many predictors, *MPRA* Paper No. 30380.

[18] Kotz, S., T. J. Kozubowski and K. Podgórski (2001), The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance, Birkhäuser Boston.

[19] Kyung, M., J. Gill, M. Ghosh, and G. Casella (2010), Penalized Regression, Standard Errors, and Bayesian Lassos, *Bayesian Analysis*, 5, 369-412.

[20] Leng, C, M. N. Tran and D. Nott (2010), Bayesian adaptive Lasso, *arXiv:1009.2300v1*

[21] Litterman, R.(1986), Forecasting with Bayesian vector autoregressions five years of experience, *Journal of Business and Economics Statistics*, 4, 25-38.

[22] Muirhead R. J. (1982), Aspects of Multivariate Statistical Theory, Wiley: New York.

[23] Park, T. and G. Casella (2008), The Bayesian Lasso, *Journal of the American Statistical Association*, 103, 681-86.

[24] Sims, C. (1972), Money, Income and Causality, *American Economic Review*, 62, 540-52.

[25] Sims, C. (1980), Macroeconomics and Reality, *Econometrica*, 48, 1-48.

[26] Song, S. and P. Bickel (2010), Large vector auto regressions, *arXiv:1106.3915v1.*

[27] Stock, J. H. and M. W. Watson (2002a), Forecasting using principal components from a large number of predictors, *Journal of the American Statistical Association*, 97, 1167-79.

[28] Stock, J. H. and M. W. Watson (2002b), Macroeconomic forecasting using diffusion indexes, *Journal of Business and Economic Statistics*, 20, 147-62.

[29] Stock, J. H. and M. W. Watson (2008), Forecasting in dynamic factor models subject to structural instability, in Castle, J. and N. Shephard (*eds*) The Methodology and Practice of Econometrics: A Festschrift in Honour of Professor David F. Hendry, 173205. Oxford University Press: Oxford.

[30] Tibshirani, R. (1996), Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society*, B 58, 267-88.

[31] van Gerven, M., B. Cseke, R. Oostenveld and T. Heskes (2009), Bayesian Source Localization with the Multivariate Laplace Prior, in Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta *(eds.)* Advances in Neural Information Processing Systems 22, 1901-09.

[32] van Gerven, M., B. Cseke, F. P. de Lange and T. Heskes (2010), Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior, *NeuroImage*, 50, 150-61.

[33] Yuan, M. and Y. Lin (2007), Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society*, B 68(1), 49-67.

[34] Zellner, A. (1971). An introduction to Bayesian inference in econometrics, John Wiley and Sons: New York.

[35] Zou, H. and T. Hastie (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society*, B 67, 301-20.

[36] Zou, H. and Zhang, H. H. (2009), On the adaptive elastic-Net with a diverging number of parameters, *Ann. Statist.*, 37, 1733-51.

Appendix

A Data List

We use 20 US macro series used in Koop (2011), which is an updated version of Stock and Watson (2008). The variables are transformed to stationarity and standardized same as Koop (2011). Below we cite the essential data description and transformation code from Koop (2011). Please refer to Koop (2011) for detailed data information.

The data transformation codes are as following: 1: no transformation; 2: first difference; 3: second difference; 4: log; 5: first difference of logged variables; 6: second difference of logged variables.

Table 9: The List of Variables

Short name	Trans. Code	Data Description
RGDP	5	Real GDP, quantity index (2000 = 100)
CPI	6	CPI all items
FFR	2	Interest rate: federal funds (e-ective) (% per annum)
Com: spot price (real)	5	Real spot market price index: all commodities
Reserves nonbor	3	Depository inst reserves: nonborrowed (mil$)
Reserves tot	6	Depository inst reserves: total (mil$)
M2	6	Money stock: M2 (bil$)
Cons	5	Real Personal Cons. Exp., Quantity Index
IP: total	5	Industrial production index: total
Capacity Util	1	Capacity utilization: manufacturing (SIC)
U: all	2	Unemp. rate: All workers, 16 and over (%)
HStarts: Total	4	Housing starts: Total (thousands)
PPI: fin gds	6	Producer price index: finished goods
PCED	6	Personal Consumption Exp.: price index
Real AHE: goods	5	Real avg hrly earnings, non-farm prod. worker
M1	6	Money stock: M1 (bil$)
S&P: indust	5	S&Ps common stock price index: industrials
10 yr T-bond	2	Interest rate: US treasury const. mat., 10-yr
Ex rate: avg	5	US effective exchange rate: index number
Emp: total	5	Employees, nonfarm: total private

B Prior Sensitivity Analysis

Tables 10-13 report the forecast results for DAELasso, adaptive e-net Lasso, and adaptive Lasso using tighter priors. The priors for $\lambda_{1,j}^2$ (or λ_1^2) and $\lambda_{2,j}$ (or λ_2) are set to be $G(1, 0.0001)$ and $G(1, 0.001)$, respectively. The prior for Σ^{-1} is set to be $W(100I_N, 1)$,

Table 10: Rolling Forecasting for $h = 1$

	GDP	CPI	FFR
DAELasso	0.58	0.33	0.55
	(-189.9)	(-185.1)	(-211.2)
adaptive e-net Lasso	0.58	0.33	0.55
	(-189.9)	(-186.2)	(-212.6)
adaptive Lasso	0.58	0.33	0.55
	(-189.7)	(-185.4)	(-211.5)

Notes:
MSFEs as proportion of random walk MSFEs.
Sum of log predictive likelihoods in parentheses.

Table 11: Rolling Forecasting for $h = 4$

	GDP	CPI	FFR
DAELasso	0.53	0.47	0.57
	(-201.9)	(-204.8)	(-233.2)
adaptive e-net Lasso	0.53	0.47	0.57
	(-201.8)	(-204.6)	(-233.2)
adaptive Lasso	0.53	0.47	0.57
	(-201.8)	(-204.2)	(-232.8)

Notes:
MSFEs as proportion of random walk MSFEs.
Sum of log predictive likelihoods in parentheses.

Table 12: Recursive Forecasting for $h = 1$

	GDP	CPI	FFR
DAELasso	0.58	0.31	0.54
	(-193.9)	(-182.8)	(-226.1)
adaptive e-net Lasso	0.59	0.31	0.55
	(-194.5)	(-183.3)	(-226.6)
adaptive Lasso	0.58	0.31	0.55
	(-194.0)	(-182.6)	(-226.4)

Notes:

MSFEs as proportion of random walk MSFEs.

Sum of log predictive likelihoods in parentheses.

Table 13: Recursive Forecasting for $h = 4$

	GDP	CPI	FFR
DAELasso	0.53	0.47	0.56
	(-204.6)	(-208.6)	(-248.9)
adaptive e-net Lasso	0.53	0.47	0.56
	(-204.3)	(-208.2)	(-248.2)
adaptive Lasso	0.53	0.47	0.56
	(-204.4)	(-208.2)	(-248.1)

Notes:

MSFEs as proportion of random walk MSFEs.

Sum of log predictive likelihoods in parentheses.