ECORE
INTERNATIONAL ASSOCIATION FOR RESEARCH AND TEACHING
Economics, Finance, Operations Research, Econometrics and Statistics

++ research ++ teaching ++

ECORE DISCUSSION PAPER

2011/46

VAR forecasting using Bayesian variable selection

Dimitris KOROBILIS

CORE
Université catholique de Louvain

ECARES
Université Libre de Bruxelles

CORE DISCUSSION PAPER
2011/22

VAR forecasting using Bayesian variable selection

Dimitris KOROBILIS[1]

May 2011

Abstract

This paper develops methods for automatic selection of variables in Bayesian vector autoregressions (VARs) using the Gibbs sampler. In particular, I provide computationally efficient algorithms for stochastic variable selection in generic linear and nonlinear models, as well as models of large dimensions. The performance of the proposed variable selection method is assessed in forecasting three major macroeconomic time series of the UK economy. Databased restrictions of VAR coefficients can help improve upon their unrestricted counterparts in forecasting, and in many cases they compare favorably to shrinkage estimators.

Keywords: forecasting, variable selection, time-varying parameters.

JEL Classification: C11, C32, C52, C53, E37

[1] Université catholique de Louvain, CORE, B-1348 Louvain-la-Neuve, Belgium.
E-mail: dimitrios.korompilis@uclouvain.be.

This paper presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The scientific responsibility is assumed by the author.

1 Introduction

Since the pioneering work of Sims (1980), a large part of empirical macroeconomic modeling is based on vector autoregressions (VARs). Despite their popularity, the flexibility of VAR models entails the danger of over-parameterization, which can lead to poor forecasts. This pitfall of VAR modelling was recognized early, and in response shrinkage methods have been proposed; see for example the so-called Minnesota prior (Doan, Litterman and Sims, 1984). Nowadays the applied econometricians' toolbox includes numerous efficient modelling tools to prevent the proliferation of parameters and eliminate parameter and model uncertainty: variable selection priors (George, Sun and Ni, 2008), steady-state priors (Villani, 2009), Bayesian model averaging (Garratt, Koop, Mise and Vahey, 2009) and factor models (Stock and Watson, 2006), to name but a few.

This paper develops a stochastic search algorithm for variable selection in linear and nonlinear vector autoregressions (VARs) using Markov Chain Monte Carlo (MCMC) methods. The term "stochastic search" simply means that if the model space is too large to assess in a deterministic manner (that is, enumerate and estimate all possible models, and decide on the best one using some goodness-of-fit measure), the algorithm will visit only the most probable models in a stochastic manner. In this paper, the general model form that I am studying is the reduced-form VAR model, which can be written using the following linear regression specification

$$y_t = c + B_1 y_{t-1} + B_2 y_{t-2} + ... + B_p y_{t-p} + \varepsilon_t \tag{1}$$

where y_t is an $m \times 1$ vector of $t = 1, ..., T$ time series observations on the dependent variables and the errors ε_t are assumed to be $N(0, \Sigma)$, where Σ is an $m \times m$ covariance matrix. The idea behind Bayesian variable selection is to introduce indicators γ_{ij} such that

$$\begin{aligned} B_{ij} &= 0 \text{ if } \gamma_{ij} = 0 \\ B_{ij} &\neq 0 \text{ if } \gamma_{ij} = 1 \end{aligned} \tag{2}$$

where B_{ij} is an element of the $m \times k$ coefficient matrix $B = \left[c, B_1', ..., B_p'\right]$, for

$i = 1, .., m$, $j = 1, ..., k$ and $k = p + 1$.

There are various benefits of using this approach over some of the shrinkage methods mentioned previously, such as the Minnesota prior or factor models. First, variable selection is automatic, meaning that along with estimates of the parameters we get associated probabilities of inclusion of each parameter in the "best" model. In that respect, the variables γ_{ij} indicate which elements of B should be included or excluded from the final optimal model. Selection of the optimal model is implemented among all possible 2^n, $n = mk$, VAR model combinations, without the need to estimate each and every one of these models. Second, this form of Bayesian variable selection is independent of the prior assumptions about the coefficients B. That is, if the researcher has defined any desirable prior for the parameters of the unrestricted model (1), adopting the variable selection restriction (2) needs no other modification than adding one extra block in the posterior sampler that draws from the conditional posterior of the γ_{ij}'s. An indirect implication of this approach is that, unlike other proposed stochastic search variable selection algorithms for VAR models (George et al. 2008; Korobilis, 2008), variable selection of this form may be adopted in VAR models which are nonlinear in the mean coefficients B.

In fact, in this paper I show that variable selection is very easy to adopt in the non-linear and richly parameterized, time-varying parameters vector autoregression (TVP-VAR). These models are currently very popular for measuring monetary policy and have been used extensively in academic research (Canova and Gambetti, 2009; Cogley and Sargent, 2002; Cogley, Morozov and Sargent, 2005; Koop, Leon-Gonzalez and Strachan, 2009; and Primiceri, 2005). Common feature of these papers is that they all fix the number of autoregressive lags to 2 for parsimony. This simplification is so popular because marginal likelihoods are difficult to obtain, especially in the presence of stochastic volatility where one has to rely on computationally expensive particle filtering methods (Koop and Korobilis, 2009a). Even if we assume that marginal likelihoods are readily available, these would allow only pairwise comparisons and hence all 2^n TVP-VAR models need to be estimated. Therefore, automatic variable selection is a convenient and fast way to overcome the computational and practical problems associated with (computationally) demanding nonlinear VAR models as well as simple linear models.

Apart from the TVP-VAR I examine closely the performance of Bayesian variable

selection on several VAR formulations with various prior specifications. In particular I begin with the simple linear VAR model with ridge regression, Minnesota, and adaptive shrinkage priors. Following this, variable selection for nonlinear models is introduced, where in addition to the TVP-VAR I consider a multivariate extension of the Koop and Potter (2007) structural breaks autoregressive model which allows to forecast breaks out-of-sample. Finally, given the recent interest in forecasting with large models (Bańbura, Giannone and Reichlin, 2010) as an alternative to dimension reduction using principal components (Stock and Watson, 2006), a modification of the stochastic restriction search useful for VARs of medium and large dimensions is established.

Although the methods described in this paper can be used for structural analysis (by providing data-based restrictions on the coefficients which could enhance identifying monetary policy for instance), the aim is to show how more parsimonious models can be selected to have a positive impact on macroeconomic forecasting.

The next section describes the mechanics behind variable selection in a general VAR setting. In Section 3, variable selection is established for specific cases of linear VAR models of small and larger dimensions, and nonlinear models. The paper concludes by evaluating the out-of-sample forecasting performance of VAR models using variable selection, for three key UK macroeconomic variables observed over the period 1971:Q1 - 2008:Q4.

2 Variable selection in vector autoregressions

To allow for different equations in the VAR to have different explanatory variables, rewrite equation (1) as a system of seemingly unrelated regressions (SUR)

$$y_t = z_t\beta + \varepsilon_t \tag{3}$$

where $z_t = I_m \otimes x_t = I_m \otimes (1, y_{t-1}, ..., y_{t-p})$ is a matrix of dimensions $m \times n$, $\beta = vec(B')$ is $n \times 1$, and $\varepsilon_t \sim N(0, \Sigma)$. When no parameter restrictions are present in equation (3), this model will be referred to as the unrestricted model. Bayesian variable selection is incorporated by defining and embedding in model (3) indicator

variables $\gamma = (\gamma_1, ..., \gamma_n)'$, such that $\beta_j = 0$ if $\gamma_j = 0$, and $\beta_j \neq 0$ if $\gamma_j = 1$. These indicators γ are treated as random variables by assigning a prior on them, and allowing the data likelihood to determine their posterior values. We can explicitly insert these indicator variables multiplicatively in the model[1] using the following form

$$y_t = z_t\theta + \varepsilon_t \tag{4}$$

where $\theta = \Gamma\beta$. Here Γ is an $n \times n$ diagonal matrix with elements $\Gamma_{jj} = \gamma_j$ on its main diagonal, for $j = 1, ..., n$. It is easy to verify that when $\gamma_j = \Gamma_{jj} = 0$ then θ_j is restricted and is equal to $\Gamma_{jj}\beta_j = 0$, while for $\gamma_j = \Gamma_{jj} = 1$ it holds that $\theta_j = \Gamma_{jj}\beta_j = \beta_j$, so that all possible 2^n VAR specifications can be explored and variable selection in this case is equivalent to model selection.

2.1 A generic VAR case

The restricted VAR specification (4) may serve as a generic formulation for the rest of the models. All we have to do is make sure that we can write the linear/nonlinear VAR models in SUR form. For instance, in the next section I show that when using nonlinear models we can arrive in a SUR form similar to equation (4), but in this case it will hold that $\theta = \Gamma g(\beta)$. Here $g(\beta)$ is any class of nonlinear functions of the VAR parameters β, with a prior density $F(\cdot)$, that is

$$p(g(\beta)) \sim F(a, G_0) \tag{5}$$

In this paper I focus on specifications of interest to macroeconomists who usually assume that $g(\beta)$ is a piecewise linear function (as it is the case with the class of structural breaks, Markov Switching and threshold autoregressive specifications, among others) but generalizations to other nonlinear or nonparametric functions is almost as straightforward.

Derivations are simplified if the indicators γ_j are a priori independent of each other for $j = 1, ..., n$, i.e. $p(\gamma) = \prod_{j=1}^{n} p(\gamma_j) = \prod_{j=1}^{n} p(\gamma_j|\gamma_{\setminus -j})$, where $_{\setminus -j}$ indexes all the elements of a vector but the $j - th$. Additionally, we can remove the effect of the covariance matrix by integrating this parameter using an a scale invariant

[1] See for example the formulation of variable selection in Kuo and Mallick (1997).

improper Jeffrey's prior. Hence we have

$$\gamma_j|\gamma_{\setminus -j} \sim Bernoulli\left(\pi_{0j}\right) \tag{6}$$
$$\Sigma \propto |\Sigma|^{-(m+1)/2} \tag{7}$$

where π_{0j} is the prior probability of the Bernoulli density, implying prior belief that coefficient j is restricted.

The following pseudo-algorithm demonstrates that the algorithm for the restricted model (4) actually adds only one block (which samples the restriction indicators γ) over the standard algorithm of the unrestricted VAR model (3). In the rest of the paper I define $y = (y_1, ..., y_T)'$ and $z = (z_1, ..., z_T)'$.

Bayesian Variable Selection Pseudo-Algorithm

1. Sample $g(\beta)$ from the conditional posterior (assuming it exists)[2] of the form

$$g(\beta)|\Sigma, y, z, \Gamma \sim L(y, z^*; g(\beta)|\Sigma, \Gamma) \times F(a, G_0)$$

where $L(y, z; g(\beta)|\Sigma, \Gamma)$ is the conditional likelihood (i.e. conditional on Σ, Γ being known). Here z_t^* is the restricted data matrix with $z_t^* = z_t\Gamma$

2. Sample each γ_j conditional on $\gamma_{\setminus -j}$, $g(\beta)$, Σ and the data from

$$\gamma_j|\gamma_{\setminus -j}, g(\beta), \Sigma, y, z \sim Bernoulli\left(\pi_{0j}\right) \tag{8}$$

preferably in random order j, $j = 1, ..., n$, where $\widetilde{\pi}_j = \frac{l_{0j}}{l_{0j}+l_{1j}}$, with

$$l_{0j} = p\left(y|\theta_j, \Sigma, \gamma_{\setminus -j}, \gamma_j = 1\right)\pi_{0j} \tag{9}$$
$$l_{1j} = p\left(y|\theta_j, \Sigma, \gamma_{\setminus -j}, \gamma_j = 0\right)(1 - \pi_{0j}) \tag{10}$$

3. Sample Σ as in the unrestricted VAR in (3), where now the mean equation

[2]For all the popular nonlinear models I consider, the posterior conditionals exist, so that a Metropolis step within the Gibbs sampler is not needed to sample from $g(\beta)$.

parameters are $\theta = \Gamma g(\beta)$.

$$\Sigma^{-1}|\beta, \gamma, y, z \sim Wishart\left(\widetilde{\alpha}, \widetilde{S}^{-1}\right) \tag{11}$$

where $\widetilde{\alpha} = T$ and $\widetilde{S} = \left(\sum_{t=1}^{T}(y_{t+h} - z_t\theta)'(y_{t+h} - z_t\theta)\right)$.

In this type of model selection, what we care about is which of the parameters θ are equal to zero, so that identifiability of $g(\beta)$ and γ plays no role. In a Bayesian setting identifiability is still possible, since if the likelihood does not provide information about a parameter, its prior does. When for a specific $j = 1, .., n$ we sample a $g\left(\beta_j\right) = 0$ then γ_j is identified by drawing from its prior: notice that in this case in equations (9) - (10) it holds that $p\left(y|\theta_j, \gamma_{\setminus -j}, \gamma_j = 1\right) = p\left(y|\theta_j, \gamma_{\setminus -j}, \gamma_j = 0\right)$, so that the posterior probability of the Bernoulli density, $\widetilde{\pi}_j$, will be equal to the prior probability π_{0j}. Similarly, when $\gamma_j = 0$ then $g\left(\beta_j\right)$ is identified from its prior: the j-th column of $z_t^* = z_t\Gamma$ will be zero, i.e. the likelihood provides no information about $g\left(\beta_j\right)$, and sampling from the posterior of $g\left(\beta_j\right)$ collapses to getting a draw from its prior. Nevertheless, in both of the above cases the result of interest is that the j-th parameter should be restricted since $\theta_j = 0$.

Posterior computation is based on Gibbs sampler with complete blocking. If the support of β is finite (see also the discussion of priors on β in the next section), then we can use the argument of Tierney (1991) to show that the Markov Chain is geometrically ergodic and that a Central Limit Theorem on this Markov Chain is available. Thus, convergence of the Gibbs sampler is expected to be quite rapid, and selection of the correct restrictions quite accurate. A simulation study in the working paper version of this article confirms that this is the case for both linear and nonlinear VAR models in small samples.

3 VAR formulations and priors

This section describes in detail some popular VAR specifications and various prior distributions on them that are considered in the empirical application of this paper. The main idea is to compare all linear and nonlinear VAR formulations using some popular priors routinely used in business and academia, with and without variable

selection. First, I show how each of these popular VAR models admit a SUR form. Then the model with variable selection is the one where the γ_j's are sampled from (8), and the corresponding unrestricted model is the one where we simply impose $\gamma_j = 1 \; \forall j$ without sampling from the posterior (as it will be clear in Section 4, this model is also equivalent to imposing the tight prior $\pi_{0j} = 1 \; \forall j$ on the restricted model). Some of the priors described here already provide some shrinkage (i.e. they provide data-based rules to restrict irrelevant VAR coefficients). This fact implies that we can examine how variable selection competes with traditional shrinkage (for instance the Minnesota prior), but also if combining variable selection and shrinkage priors in the same VAR model could help improve forecasting even further.

In order to do such a comparison, the intercepts are left unrestricted ($\gamma_j = 1$ if β_j is an intercept) and flat priors are placed on them in all instances. Similarly the covariance matrix is integrated out with the improper scale invariant (Jeffrey's) prior in equation (7). Finally, the hyperparameters π_{0j} found in equation (6) are set to $\pi_{0j} = 0.8$ implying that 80% of the predictors should be included in the final model. This assumption is reasonable for small trivariate VARs, since the "noninformative" choice $\pi_{0j} = 0.5$ implies that probably too many (i.e. 50%) VAR coefficients should be restricted. In subsection 3.4 I introduce variable selection specifically for large VARs. There I relax this assumption and propose setting the values of π_{0j} in the spirit of the Minnesota prior (i.e. penalize heavily more distant lags using the variable selection algorithm) which can assist in solving the curse of dimensionality problem in these models. Full Bayes and Empirical Bayes priors can also be used on π_{0j} and the reader can seek more information in Chipman, George and McCulloch (2001).

3.1 Linear VAR

The traditional VAR process with variable selection is fully described by equation (4), where β (and hence $\theta = \Gamma\beta$) enters the model linearly. Typical prior distributions for linear VAR models are based on the Normal density, i.e.

$$\beta \sim N_n\left(\underline{b}, \underline{V}\right)$$

In this paper I examine three types of eliciting prior hyperparameters based on the Normal distribution, all of which provide some form of shrinkage in the VAR coefficients (but no exact zero restrictions like variable selection does).

Ridge regression prior This is probably the most widely used prior in autoregressive models. The assumption is that $\underline{b} = 0_{n\times 1}$ and $\underline{V} = \lambda I_n$. The posterior mean/mode of the Bayes estimator is equal to the penalized least squares estimator which writes

$$\widetilde{\beta} = \left(z'z + \lambda^{-1} I_n\right)^{-1} z'y$$

which is equivalent to unrestricted LS for $\lambda \to \infty$. The reader should also note that for the case $\lambda \to \infty$ (in practical situations this translates to $\lambda = 100$ and above) variable selection cannot be performed. An intuitive explanation for this effect is that marginal likelihoods for model selection cannot be calculated with uninformative priors. Kuo and Mallick (1997) give a more detailed explanation about this issue and propose to use values of $\lambda \in [0.25, 25]$. Consequently, in the absence of prior information about the model coefficients, one can use a locally uninformative prior by setting $\lambda = 100$ (diffuse prior) on the intercepts and $\lambda = 9$ for autoregressive coefficients. In near-covariance stationary VAR processes the autoregressive coefficients are expected to be roughly less than one in absolute value, so a higher value of λ for these parameters is basically redundant.

Minnesota (Litterman) prior The Minnesota prior is very popular and is as old as the VAR literature in economics. This prior is due to the works of Bob Litterman and colleagues at Minnesota University and the Minneapolis Fed; see for instance Litterman (1986) and Doan, Litterman and Sims (1984). This Empirical Bayes formulation assumes the prior mean vector $\underline{b}$ is set equal to 1 for parameters on the first own lag of each variable (random walk prior) and zero otherwise, and $\underline{V}$ is a diagonal matrix with diagonal element the variance on lag r of variable j in equation i of the form

$$\underline{V}_{ij}^{r} = \begin{cases} 100 s_i^2 & \text{if intercept} \\ 1/r^2 & \text{if } i = j \\ \lambda \frac{s_i^2}{r^2 s_l^2} & \text{if } i \neq j \end{cases} \tag{12}$$

for $r = 1, ..., p$, $i = 1, ..., m$,and $j = 1, ..., k$ with $k = p + 1$. Here s_i^2 is the residual variance from the unrestricted p-lag univariate autoregression for variable i. The degree of shrinkage depends on a single hyperparameter λ[3], where again if $\lambda \to \infty$ we end up with unrestricted estimates similar to LS. Litterman (1986) originally introduced a hyperparameter for own lags as well, i.e. he used $\underline{V}_{ij}^r = \kappa/r^2$ if $i = j$ in equation (12). For small and medium VAR models it is the choice of λ that matters. I set $\kappa = 1$ which provides a "realistic" prior variance for own lag coefficients. In covariance-stationary VARs we do not expect these coefficients to be much larger than 1 especially for higher order lags, so $1/r^2$ should (and does) work fine. Selection of λ in contrast is dependent on the specific dataset and application considered. Selection of the shrinkage factor λ of the Minnesota prior is discussed in subsection 4.1.

Hierarchical Bayes Shrinkage prior Shrinkage priors based on Empirical Bayes methods, like the Minnesota prior, suffer from the fact that they are subjective constructs and might not appeal to the objective researcher. The formal Bayesian way to shrinkage in regressions is to use hierarchical priors on the regression coefficients so that the shrinkage parameter λ is chosen objectively by the data. In Korobilis (2011) I show that using hierarchical Normal-Gamma priors, we can recover many popular shrinkage estimators for sparse signals, like the least absolute shrinkage and selection operator (LASSO) of Tibshirani (1996) and its variants (Fused LASSO, Group LASSO, Elastic Net). Here I use a special case of adaptive shrinkage Normal-Gamma priors which is the hierarchical Normal-Jeffrey's prior of Hobert and Casela (1993) of the form

$$\begin{aligned} \beta &\sim N_n\left(0, \underline{V}\right), \quad \underline{V}_{jj} = \lambda_j,\ j = 1, ..., n \\ \lambda_j &\sim \begin{cases} 100 & \text{if } \beta_j \text{ is an intercept coefficient} \\ 1/\lambda_j & \text{otherwise} \end{cases} \end{aligned} \tag{13}$$

[3]Litterman (1986) originally introduced a hyperparameter for own lags as well, i.e. he used $\underline{V}_{ij}^r = \kappa/r^2$ if $i = j$ in equation (12). For small and medium VAR models it is the choice of λ that matters. I set $\kappa = 1$ which provides a "realistic" variance for own lag coefficients (we do not expect these coefficients to be much larger than 1).

In simple words, by placing a scale invariant Jeffreys' distribution on λ_j, its posterior value is determined solely by the data (hence λ is not a prior choice for the researcher). This is the simplest form of adaptive shrinkage, and can easily be used in VAR models. In Korobilis (2011) I show that LASSO-based Bayesian shrinkage (specifically the hierarchical version of the Elastic Net algorithm of Zou and Hastie, 2005) perform even better in forecasting than simple Normal-Jeffreys priors. However as explained in Park and Casela (2008) for LASSO-type priors we need to condition β_j on the model error variance, something not straightforward to do in a VAR model, unless we make simplifying assumptions like setting Σ to be diagonal.

3.2 Time-varying parameters VAR

Modern macroeconomic applications increasingly involve the use of VARs with mean regression coefficients and covariance matrices which drift every month/quarter. Nonetheless, forecasting with time-varying parameters VARs is not a new topic in economics. During the "Minnesota revolution" efficient approximation methods of forecasting with TVP-VARs were developed, with most notable contributions the ones by Doan, Litterman and Sims (1984) and Sims (1989); for a large-scale application in an 11-variable VAR see also Canova (1993). Using modern posterior simulator methods (Markov Chain Monte Carlo), TVP-VARs have been used recently very extensively for structural analysis (Primiceri, 2005; Cogley and Sargent, 2002) and forecasting (D'Agostino et al., 2009; Cogley et al., 2005), while Groen, Paap and Ravazzolo (2009) and Koop and Korobilis (2009b) are focusing on univariate predictions with the use of a large set of exogenous variables.

As mentioned in the Introduction, marginal likelihood calculations in this model are hard to implement. When specifically stochastic volatility is present, computationally expensive particle filtering methods are needed only to obtain a measure of fit for a single model. Estimation using Bayesian variable selection is not affected by specific modelling assumptions (like the inclusion or not of stochastic volatility) and can accommodate all possible model combinations efficiently in a single run of the Gibbs sampler.

A time-varying parameters VAR with constant covariance matrix (Homoskedas-

tic TVP-VAR) takes the form

$$y_t = c_t + B_{1,t}y_{t-1} + ... + B_{p,t}y_{t-p} + \varepsilon_t \tag{14}$$

where as before $\varepsilon_t \sim N(0, \Sigma)$ with Σ an $m \times m$ covariance matrix. This model can easily be written in the variable selection SUR form (4), by defining β_t to be the $n \times 1$ vector $\left[c_t', vec\left(B_{1,t}'\right), ..., vec\left(B_{p,t}'\right)\right]'$ of parameters and $z_t = I_m \otimes (1, y_{t-1}, ..., y_{t-p})'$ is an $m \times n$ matrix. In that case we have

$$y_t = z_t\theta_t + \varepsilon_t \tag{15}$$

$$\beta_t = \beta_{t-1} + \eta_t \tag{16}$$

where $\theta_t = \Gamma\beta_t$ and Γ is the $n \times n$ matrix defined in (4). Equation (16) defines a random walk evolution of the nonlinear VAR coefficients[4], for which it holds that $\eta_t \sim N(0, Q)$ with Q an $n \times n$ covariance matrix.

Note that variable selection in this case implies that a VAR coefficient either enters or exits the "true" model in all time periods $t = 1, ..., T$. In contrast, today there are methods in univariate regressions which allow different coefficients to be selected at different points in time. Most notably, Chan, Koop, Leon-Gonzalez and Strachan (2010) use such a flexible specification, however estimation relies on computationally intensive MCMC procedures which only allow them to consider a handful of variables. The efficient approximations we describe in Koop and Korobilis (2009b) allow dynamic model averaging (DMA) and selection (DMS) with up to around 20 predictors (i.e. to average or select among 2^{20} models at each period t). Nonetheless, the smallest typical VAR used in macroeconomics has three quarterly variables and four lags and an intercept (39 mean coefficients), which makes application of DMA computationally intensive.

While the priors for (Σ, Γ) are the same as in the previous cases (Jeffrey's-Bernoulli), it can be shown that conjugate priors for the remaining parameters of

[4] An autoregressive model of order one could be defined, but early empirical experience with these models (see Sims, 1989) suggests that the AR(1) coefficient is practically very close to 1.

the TVP-VAR model (Cogley and Sargent, 2002) are of the form

$$\begin{aligned} \beta_0 &\sim N_n\left(\underline{b}, \underline{V}\right) \\ Q^{-1} &\sim Wishart\left(\xi, R^{-1}\right) \end{aligned}$$

with β_0 being practically the initial condition of β_t. Note that a prior on each β_t, $t = 1, ..., T$, need not be specified since this is implicitly defined recursively as $\beta_t \sim N_n\left(\beta_{t-1}, Q\right)$. An important thing to underline is that the model allows the VAR coefficients β_t to evolve as random walks for T periods, so that shrinkage/tight priors must be used especially for Q (a detailed explanation why is given in Primiceri, 2005, Section 4.4). Cogley and Sargent (2005) and Primiceri (2005) use the OLS estimates of a simple VAR estimated on a training sample to inform their prior hyperparameters, and set their shrinkage coefficient (what was denoted as λ in the linear VAR priors) at a very small value. This approach is standard in Bayesian analysis, especially when marginal likelihoods are not readily available, but it results in discarding valuable information in the training sample.

In contrast the standard Minnesota prior can be used to inform the initial condition β_0 of the TVP-VAR coefficients, combined with a tight prior on Q. Subsequently, we can set $\underline{b}$ and $\underline{V}$ as in equation (12), while setting $\xi = 2\left(n+1\right)$ and $R = k_R {I_n}^5$, where n is the number of coefficients in β_t and k_R is a scaling factor which we have to choose. Following Cogley and Sargent's (2002) "business as usual" prior, i.e. the belief that the TVP-VAR coefficients should vary smoothly and not change abruptly each time period, I set $k_R = 0.0001$. This is the standard value used by Primiceri after implementing a sensitivity analysis, see Primiceri (2005, Section 4.4.1). Consequently, as in the linear VAR models, we only need to worry about the value of the shrinkage coefficient λ, a choice which is discussed in the empirical section.

[5] To replicate Primiceri's (2005) training sample prior, we can use $R = k_R \underline{V}$ where as before $\underline{V}$ is the Minnesota prior covariance matrix. However this assumption does not alter any of the forecasting results for the UK dataset used in the empirical section.

3.3 Structural breaks VAR

In theory and in practice, a VAR with structural breaks lies between the linear VARs (zero breaks) and the TVP-VAR (breaks in every period, i.e. T breaks) and should have been presented earlier. However one of all the possible formulations of structural breaks in the VAR coefficients, which is due to Koop and Potter (2007), is to write the model as a special case of the TVP-(V)AR presented above. Subsequently, following equations (15) and (16) we can write the structural breaks VAR using the form

$$y_t = z_t \theta_{s_t} + \varepsilon_t \tag{17}$$

$$\beta_{s_t} = \beta_{s_{t-1}} + \eta_{s_t}. \tag{18}$$

Here $\theta_{s_t} = \Gamma \beta_{s_t}$, $\eta_t \sim N(0, Q)$, and $s_t \in [1, ..., K+1]$ is a first order Markov process with block-diagonal transition matrix of the form

$$P = \begin{bmatrix} p_{11} & p_{12} & 0 & \cdots & 0 \\ 0 & p_{22} & p_{23} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ & & 0 & p_{KK} & p_{K,K+1} \\ 0 & \cdots & 0 & 0 & p_{K+1} \end{bmatrix}$$

which makes the structural breaks model a restricted form of a Markov switching VAR, since we can only move from one regime to the next, and never return to a previous regime. In this case we have a breaks between time period t and $t+1$ iff $s_t \neq s_{t+1}$. Uncertainty about the number of regimes is easily incorporated in a Bayesian context by setting a maximum number of breaks, say $K_{\max}$, and allowing the data to determine the "true" number of estimated breaks K, where $1 \leq K \leq K_{\max}$. In Bauwens, Koop, Korobilis and Rombouts (2011) we give exact implementation details on forecasting with a univariate version of this model, which I follow closely in this multivariate extension. Estimation details are provided in the Appendix.

The hyperparameters on the initial condition, $\beta_0 \sim N_n(\underline{b}, \underline{V})$, and the state

covariance matrix, $Q^{-1} \sim Wishart\left(\xi, R^{-1}\right)$, are based on Sims's version of the Minnesota prior explained in the previous subsection. The additional parameters on this model are the transition probabilities $p_{ij} = \Pr\left[s_t = i | s_{t-1} = j\right]$, for which I use the typical Beta prior for the diagonal elements $p_{ii} \sim Beta\left(\rho_1, \rho_2\right)$, $i = 1, ..., K$. For $\rho_1 = \rho_2 = 1$ this density becomes uniform and noninformative. The parameters s_t are estimated as in Chib (1996).

3.4 Extension to large VARs and comparison with other models

The fact that automatic Bayesian variable selection is stochastic and simulation is needed (Gibbs sampler) implies that it's use is in general prohibitive in VARs with hundreds of dependent variables as in Bańbura, Giannone and Reichlin (2010). Moreover, the disadvantage of variable selection is that in order to allow different variables to enter different equations, the SUR form of the VAR is needed which relies on inverting large matrices (since the RHS data matrix is $z_t = I_m \otimes x_t$ instead of just x_t in the reduced-form VAR). Even so, this subsection discusses some modifications to variable selection that would make its usage in medium-sized VARs possible.

Consider the linear VAR[6] model (1) written compactly as

$$y_t = Bx_t + \varepsilon_t$$

where $x_t = (1, y_{t-1}, ..., y_{t-p})$ and $B = \left[c, B_1', ..., B_p'\right]$ is $m \times k$. Instead of restricting individually each of the $n = mk$ elements of B, when m is "large" we might want to consider restricting only the k columns of B. This simplification implies that a specific RHS variable $y_{i,t-j}$, $i \in [1, m]$, $j \in [1, p]$ either enters simultaneously in all m VAR equations or none. While this results in a loss of modelling flexibility, the implication is that when we model, say, $m = 15$ variables in a VAR with $p = 4$ lags we only need to average across 2^{60} models as opposed to the 2^{900} models available

[6] Obviously treating large nonlinear VARs is not different. However this is not discussed, since large time-varying parameters and structural breaks VARs are computationally intensive. In Korobilis (2011b) I derive efficient computational methods to forecast with VARs of very large dimensions (whether T or m are in the order of thousands) in seconds of computer time.

otherwise. More importantly, we do not need the computationally expensive SUR form to estimate the VAR model, since we can now write the large VAR + model selection model as

$$y_t = \Theta x_t + \varepsilon_t$$

where $\Theta = B\Gamma$ with Γ the $k \times k$ diagonal matrix with the restriction indices γ on its diagonal.

It would be of benefit to relax the assumption that the prior on the indices γ_j is Bernoulli with "uninformative" hyperparameter $\pi_{0j} = 0.5$. It is feasible to impose many restrictions a priori by setting $0 < \pi_{0j} \ll 0.5$[7]. For instance $\pi_{0j} = 0.1$ means that our expectation is that 90% of the coefficients should be restricted. However, we need not impose these restrictions linearly on all parameters. Following the Minnesota tradition we can use a prior which restricts a priori coefficients on more distant lags

$$\pi_{0j} = \begin{cases} 0.5, \text{ for own lags} \\ 1/(r+1), \text{ otherwise} \end{cases}$$

where $r = 1, .., p$.

The idea to restrict the VAR regression coefficients can also be extended to finding restrictions in the covariance matrix of a VAR. In fact, Smith and Kohn (2002) and Wong, Carter and Kohn. (2003), take the Cholesky decomposition $\Sigma^{-1} = A\Omega A'$ of an $m \times m$ covariance matrix, and impose restrictions on the matrix A using indicator variables, say δ. In this decomposition Ω is a diagonal matrix and A is a lower triangular matrix with 1's on the diagonal. Hence model selection proceeds by setting

$$\begin{aligned} \alpha_i &= 0 \text{ if } \delta_i = 0 \\ \alpha_i &\neq 0 \text{ if } \delta_i = 0 \end{aligned}$$

[7]The alternative $\pi_{0j} > 0.5$ imposes the prior belief that not many restrictions are expected in the VAR coefficients. If the researcher is uncertain about these beliefs, a Beta prior can always be placed on π_{0j} which makes this hyperparameter an unknown random variable to be updated from the data.

where α_i is each of the $m(m-1)/2$ non-zero and non-one elements of A. Therefore, similarly to the case of variable selection in the mean equation coefficients, their approach can be easily generalized to a covariance matrix which is stochastic as for example in the popular Heteroskedastic TVP-VARs of Primiceri (2005), Canova and Gambetti (2009) and Cogley and Sargent (2002). Considering covariance matrix selection and assuming different functional forms for the covariance matrix (say time-varying, or structural breaks) will affect forecasts to some extent and would not allow to evaluate the performance of variable selection in the mean VAR equation, which is of prime interest since it has much larger number of coefficients. For that reason, it is better to integrate out the (constant) covariance matrix, as well as the intercepts, using uninformative priors as is the standard practice in the Bayesian Statistics literature when evaluating model selection or shrinkage priors (see among others Park and Casella, 2008; Villani, 2009; and Liang, Paulo, Molina, Clyde and Berger, 2008).

There are several other approaches to automatic Bayesian model selection and shrinkage for univariate regression models which can be generalized to VAR models. The formal "full-Bayes" procedure as it is called, is based on hierarchical Normal priors of the form

$$\begin{array}{c} \beta|\gamma \sim N_n\left(0_{n\times 1}, \gamma \underline{V}\right) \\ \gamma \sim F\left(a,b,c\right) \end{array} \tag{19}$$

where $\underline{V}$ is a prior covariance matrix and $F(\cdot)$ denotes a density function with parameters a, b, c. In this case, if the prior distribution of γ, $F(a,b,c)$, is the $Bernoulli(\pi)$ then γ takes only the values 0 and 1 and we have model selection identical to the one described above (if $\gamma = 1$ the prior is $(\beta|\gamma = 1) \sim N_n(0, \underline{V})$, if $\gamma = 0$ the prior is $(\beta|\gamma = 0) \sim N_n(0_{n\times 1}, 0_{n\times n})$, i.e. a Dirac δ point mass at zero). This is the case of the stochastic search variable selection (SSVS) prior used in George, Sun and Ni (2008), Korobilis (2008) and Jochmann, Koop and Strachan (2010). As discussed in subsection 3.1 if we assume $\underline{V} = I_n$ and we assign a prior for γ of the form $\gamma \sim Gamma(\alpha_1, \alpha_2)$ then we can have shrinkage of β dependent on whether the $\gamma \gg 0$ or $\gamma \to 0$. Additionally, the shrinkage priors have the desirable property that they become variable/model selection priors in models with more predictors than observations; see Korobilis (2011).

From a practitioner's point of view, it must be noted that the SSVS prior as well as adaptive shrinkage priors of this hierarchical form are computationally much faster than variable selection considered in this paper. The main issue with Hierarchical Gaussian priors is that they cannot be used in nonlinear VARs like the TVP-VAR, which are of special interest to academics and practitioners in Central Banks. A hierachical prior like (19) can be potentially applied to the initial condition of the TVP-VAR, which would take the form $\beta_0|\gamma \sim N_n(0, \gamma\underline{V})$. We can immediately observe that for the subsequent time periods, the prior on the time-varying coefficients becomes $\beta_t \sim N_n\left(\beta_{t-1}, Q\right)$ so that dependence on the shrinkage properties of γ is lost, and the prior mean becomes β_{t-1} which in general will be estimated from the likelihood to be other than zero. To the best of my knowledge there are no formal Bayesian model selection or shrinkage estimators for these nonlinear VARs and the focus of this paper is to fill this gap using the methods described so far.

4 Macroeconomic forecasting with VARs

The variable selection techniques described previously are used to provide forecasts of three major U.K. macroeconomic series. These series are: the unemployment rate u_t (Unemployment rate: All aged 16 and over, Seasonally adjusted); the inflation rate π_t (RPI:Percentage change over 12 months: All items); and the interest rate r_t (Treasury bills: average discount rate). The data are obtained from the Office for National Statistics (ONS) website: http://www.statistics.gov.uk/. The available sample runs from 1971Q1 to 2008Q4. All variables are measured originally on a monthly basis, and quarterly series are calculated by the ONS by taking averages over the quarter (for inflation), the value at the mid-month of the quarter (for unemployment), and the value at the last-month of the quarter (for the interest rate), respectively.

Unemployment u_t is specified as a gap from its trend $\widetilde{u}_t$, where the trend is estimated using the one-sided low pass filter $\widetilde{u}_t = \widetilde{u}_{t-1} + 0.2\left(u_t - \widetilde{u}_{t-1}\right)$. This is an approximation to an exponentially weighted moving average filter which is an easy but effective way to estimate the trend in economic time series; see also the discussion

in Cogley, Morozov and Sargent (2005) and references therein. Henceforth, whenever "unemployment" is mentioned, this will be the unemployment gap variable $u_t - \widetilde{u}_t$.

4.1 Forecasting models

Here I provide a summary of all the models presented in the previous section. The models compared in this article are the linear Bayesian VAR with ridge regression (VAR Ridge), Minnesota (VAR Min) and adaptive shrinkage prior (VAR Shrink). The two nonlinear models estimated for the UK data are the time-varying parameters VAR (TVP-VAR) and the structural breaks VAR (SB-VAR), both with a Minnesota prior on the mean coefficients[8]. Additionally a 13-variable linear VAR with Minnesota prior is estimated (Large-VAR). The variables in this model are the ones used in the trivariate VARs above plus 10 major variables for the UK economy including GDP, total employment, £/$ exchange rate and money stock M4 . These models are summarized in Table 1. This gives forecasts from six models with and without variable selection, i.e. a total of 12 model forecasts to assess. All models have an intercept and 4 lags of the dependent variables.

Moreover, we have to decide on selection of the shrinkage coefficient λ for the Minnesota prior. This can be done subjectively as in Litterman (1986), but also searching over a grid of values in a training sample as in Bańbura, Giannone and Reichlin (2010). A value of $\lambda = 0.1$ is used for the trivariate linear and nonlinear VARs. This choice is the one which optimizes the forecasting performance of the TVP-VAR model in particular, compared to competing values of λ in the grid $\{1, 0.5, 0.1, 0.01, 0.01\}$. Note that this "sensitivity analysis" approach is done because the main purpose of this section is to evaluate the performance of variable selection and not which of the various VARs performs the best. It turns out that for the whole grid of values for λ, the conclusions about whether including variable selection improves forecasting or not are qualitatively similar. Following the same procedure, and based on the arguments of Bańbura, Giannone and Reichlin (2010),

[8] A "less tight" ridge regression prior can also be used in the initial condition of the mean coefficients of these two models, say $\beta_0 \sim N_n(0, 9I)$. In that case, variable selection indeed performs much better than no variable selection. In practical situations though, one would realistically use a data-based shrinkage prior in these models (like the Minnesota or the Primiceri, 2005, prior) to reduce the nonlinear parameter space.

who compare VARs of large dimensions, the shrinkage factor on the large linear VAR model is set to a tighter value, i.e. $\lambda = 0.01$.

Table 1: Definition of VAR models for the UK macro series

Model	Description
VAR Ridge	VAR with ridge regression prior, $\lambda = 9$
VAR Min	VAR with Minnesota prior, $\lambda = 0.1$
VAR Shrink	VAR with Normal-Jeffreys prior, $p(\lambda) \propto 1/\lambda$
TVP-VAR	Time-varying VAR with Minnesota prior, $\lambda = 0.1$
SB-VAR	Structural Breaks VAR with Minnesota prior, $\lambda = 0.1$
Large-VAR	Large VAR with Minnesota prior, $\lambda = 0.01$

4.2 Forecast implementation

The initial estimation period is 1971Q1 to 1989Q4 and forecasts are computed iteratively for h quarters ahead, $h = 1, 2, 3, 4$. Then one data point is added at the end of the sample (1990Q1) and forecasting is implemented again for h quarters ahead. This procedure is followed until the sample is exhausted. Estimation is based on 30.000 samples from the posterior after an initial convergence (burn-in) period of 2.000 iterations. Convergence of the Gibbs sampler is excellent in all instances.

Standard results for forecasting with VAR models apply whether or not variable selection is present. The companion form of the standard VAR model is

$$\mathbf{y}_t = \mathbf{c} + \mathbf{B}y_{t-1} + \boldsymbol{\varepsilon}_t$$

where $y_t = \left(y_t', ..., y_{t-p+1}'\right)'$, $\varepsilon_t = (\varepsilon_t', 0, ..., 0)'$, $c = (c', 0, ..., 0)'$ and

$$\mathbf{B} = \begin{bmatrix} B_1 ... B_{p-1} & B_p \\ I_{m(p-1)} & 0_{m(p-1)\times m} \end{bmatrix}.$$

Iterated h-step ahead forecasts can be computed using the formulas

$$\begin{aligned} E(\mathbf{y}_{t+h}) &= \sum_{i=0}^{h-1} \mathbf{B}^i \mathbf{c} + \mathbf{B}^h y_{t-1} \\ var(\mathbf{y}_{t+h}) &= \sum_{i=0}^{h-1} \mathbf{B}^i \boldsymbol{\Sigma} (\mathbf{B}^i)' \end{aligned} \tag{20}$$

Two points have to be clarified here. First, in the case of variable selection, the parameter matrices $B_1, ..., B_p$ are going to be replaced by the respective elements of the restricted parameter vector $\theta = \Gamma\beta$. Second, in the case of the two models with drifting coefficients, predictive simulation can be implemented to forecast breaks in the coefficients out-of-sample. This would mean that we should use the random walk evolution of the mean coefficients in the time-varying parameters and structural breaks VARs and simulate their future path using Monte Carlo; see Bauwens, Koop, Korobilis and Rombouts (2011) for more details. I follow D'Agostino, Gambetti and Giannone (2010) and relax this assumption. In that case, I use the formula (20) where I plug-in the last known values of the coefficients in sample, i.e. $\widehat{\beta}_T$ and $\widehat{\beta}_{s_T}$ respectively for the two nonlinear models.

Using MCMC implies that we sample from the full posterior density of the VAR coefficients, so that instead of a single point forecast $E(\mathbf{y}_{t+h})$ we end up having samples from the full Bayesian predictive density. This also implies that there are two ways to implement the variable selection forecasts. The one is to estimate a specific VAR model using the Gibbs sampler, save the sequence of $S = 30.000$ posterior draws γ^s, $s = 1, ..., S$, and obtain the mean/median $\overline{\gamma}$. Then the "best" model is the one for which β_j is unrestricted (restricted) if $\overline{\gamma} \geq 0.5$ ($\overline{\gamma} < 0.5$), so that we can estimate and forecast only with this best model at a second step. The second way is simply to implement one run of the MCMC and forecast using the current estimates $\theta^s = \Gamma^s\beta^s$ for $s = 1, ..., S$ MCMC samples. That way if we sample $\gamma_j = 1$ 10% of the time (3.000 samples from the posterior) and $\gamma_j = 0$ for the remaining samples, this means that we also use β_j to produce the final forecasts only 10% of the time. The former case provides absolute variable selection of a single optimal model, which is what Barbieri and Berger (2004) call the "median probability model". The second method provides relative variable selection which is equivalent to Bayesian Model Averaging. In previous research (Korobilis, 2008; Koop and Korobilis, 2009) I find that there is no clear dominance of one method over the other in forecasting. In face of this result, I use the second method for forecasting which takes explicitly into account uncertainty about the true model (by giving relative, instead of absolute, weights to each VAR coefficient).

4.3 Forecast evaluation

All models are evaluated using various measures of out-of-sample performance and forecast accuracy. Precision of mean forecasts is evaluated using averages of the Mean Absolute Forecast Error (MAFE) and the Root Mean Squared Forecast Error (RMSFE) over the whole pseudo out-of-sample evaluation period. In particular, for each of the three variables $y_{i,t}$ (i =*inflation, unemployment, interest rate*) of the vector y_t, and conditional on the forecast horizon h and the time period t, these three measures are calculated as

$$\begin{aligned}
\left(\widehat{MAFE}\right)_i^h &= \frac{1}{\tau_1 - h - \tau_0 + 1} \sum_{t=\tau_0}^{\tau_1 - h} \left| \widehat{y}_{i,t+h|t} - y_{i,t+h}^o \right| \\
\left(\widehat{RMSFE}\right)_i^h &= \sqrt{\frac{1}{\tau_1 - h - \tau_0 + 1} \sum_{t=\tau_0}^{\tau_1 - h} \left(\widehat{y}_{i,t+h|t} - y_{i,t+h}^o \right)^2}
\end{aligned}$$

where $\widehat{y}_{i,t+h|t}$ is the time $t+h$ prediction of variable i, made using data available up to time t, and $y_{i,t+h}^o$ is the observed out-of-sample value (realization) of variable i at time $t+h$. In the recursive forecasting exercise, averages over the full forecasting period 1990:Q1 - 2008:Q4 are presented using these formulas where τ_0 is 1989:Q4 and τ_1 is 2008:Q4.

These two measures can help provide a ranking of all the VAR models and give an idea of which model and prior specification performs the best. An interesting question to answer is whether the inclusion of variable selection results in overall improvement of forecasts. A simple measure is to compute the time series of differences between the squared losses of the two models, i.e.

$$d_{t+h} = \left(\epsilon_{t+h}^R\right)^2 - \left(\epsilon_{t+h}^U\right)^2, \tag{21}$$

where $\left(\epsilon_t^R\right)^2$ are the squared forecast errors from the restricted model (with variable selection), and $\left(\epsilon_{t+h}^U\right)^2$ are the squared forecast errors from the unrestricted model (without variable selection). The subscript t runs only for the pseudo out-of-sample period $\tau_1 - h - \tau_0 + 1$. Diebold and Mariano (1995) provide a simple test statistic when the null is that of equal predictive ability, i.e. $E\left(d_{t+h}\right) = 0$. From a Bayesian

point of view, since we have 30.000 samples from the predictive density of our data y_{t+h}, it is easy to construct through equation (21) an equal number of samples from the finite sample density of d_{t+h}. Hence this Bayesian procedure is equivalent, but not identical, to bootstrapping d_t under the assumption of Gaussianity (instead of having to rely on the asymptotic distribution of d_t in the presence of small samples). Subsequently, it is straightforward to get a pairwise measure of overall predictive ability by using the whole posterior density $\Pr(d_{t+h})$, i.e. we can evaluate the following "Bayesian Diebold-Mariano" (BDM) statistic

$$BDM = \frac{1}{\tau_1 - h - \tau_0 + 1} \sum_{t=\tau_0}^{\tau_1 - h} \Pr(d_{t+h} > 0), \tag{22}$$

see also Garratt, Koop, Mise and Vahey (2009). This statistic implies that if $BDM > 0.5$, the unrestricted model performs better than the restricted model, and vice versa.

4.4 In-sample variable selection results

Before proceeding to the forecast evaluation of variable selection, it would be interesting first to obtain a picture of what is the output of variable selection. Since the Gibbs sampler provides a sequence of 0-1 draws from the posterior of γ, once we take an average of these draws we can end up with an average "probability of inclusion in the true model" for the respective VAR coefficients β. Table 2 does exactly that for the six models described earlier. The table is split in three blocks pertaining to each of the three VAR equations (unemployment u_t, inflation π_t and interest rate r_t). Each row corresponds to the lags of the three variables as they appear in each equation. Numerical entries in this table are the averages of the posterior of γ using the full sample 1971:Q1 - 2008:Q4. The prior on γ for the five trivariate VARs is the $Bernoulli(0.8)$ discussed earlier, whilst for the Large VAR model the tighter prior discussed in subsection 3.4 applies.

Variable selection indicates that some variables should always be included, irrespective of the model specification or the priors used. These are the first own lags of each dependent variable, but also the first lag of the interest rate in the inflation equation. Moreover, inflation and interest rates two periods ago seem to affect the

Table 2: Posterior means of the restriction variables γ_j using the full sample

	VAR Ridge	*VAR Min*	*VAR Shrink*	*SB-VAR*	*TVP-VAR*	*Large-VAR*
			VAR equation: u_t			
u_{t-1}	1.00	1.00	1.00	1.00	1.00	1.00
π_{t-1}	0.34	0.26	0.72	0.00	1.00	1.00
r_{t-1}	0.01	0.23	0.63	0.23	1.00	1.00
u_{t-2}	0.23	0.32	0.72	0.29	0.43	0.17
π_{t-2}	0.03	0.47	0.58	0.07	0.00	1.00
r_{t-2}	0.08	0.53	0.59	0.00	0.03	1.00
u_{t-3}	1.00	0.99	1.00	1.00	0.98	0.08
π_{t-3}	0.00	0.46	0.65	0.00	0.00	0.00
r_{t-3}	0.14	0.45	0.74	0.00	1.00	0.00
u_{t-4}	0.10	0.23	0.64	0.17	0.56	0.00
π_{t-4}	0.02	0.53	0.66	0.00	0.00	0.00
r_{t-4}	0.17	0.39	0.61	0.00	0.00	0.00
			VAR equation: π_t			
u_{t-1}	0.36	0.12	0.64	0.59	0.80	1.00
π_{t-1}	1.00	1.00	1.00	1.00	1.00	1.00
r_{t-1}	1.00	1.00	1.00	0.98	1.00	1.00
u_{t-2}	0.43	0.18	0.65	0.56	0.79	0.17
π_{t-2}	0.93	0.98	1.00	0.88	1.00	1.00
r_{t-2}	0.68	0.85	0.82	0.90	0.97	1.00
u_{t-3}	0.42	0.21	0.71	0.60	0.80	0.08
π_{t-3}	0.29	0.39	0.69	0.38	0.84	0.00
r_{t-3}	0.33	0.57	0.70	0.29	0.87	0.00
u_{t-4}	0.33	0.23	0.73	0.46	0.80	0.00
π_{t-4}	0.21	0.38	0.61	0.21	0.71	0.00
r_{t-4}	0.21	0.71	0.82	0.21	0.85	0.00
			VAR equation: r_t			
u_{t-1}	0.62	0.32	0.75	0.68	0.81	1.00
π_{t-1}	0.12	0.13	0.63	0.09	0.41	1.00
r_{t-1}	1.00	1.00	1.00	0.95	1.00	1.00
u_{t-2}	0.60	0.42	0.68	0.59	0.78	0.17
π_{t-2}	0.11	0.29	0.67	0.14	0.66	1.00
r_{t-2}	0.21	0.23	0.66	0.10	0.65	1.00
u_{t-3}	0.62	0.39	0.71	0.69	0.79	0.08
π_{t-3}	0.32	0.53	0.62	0.14	0.84	0.00
r_{t-3}	0.16	0.32	0.67	0.17	0.71	0.00
u_{t-4}	0.45	0.30	0.69	0.59	0.81	0.00
π_{t-4}	0.12	0.39	0.65	0.18	0.79	0.00
r_{t-4}	0.07	0.30	0.66	0.17	0.66	0.00

current level of inflation, as well as the third lag of unemployment affects the current level of unemployment (but only in the small, trivariate VAR models). Lastly, unemployment in the previous quarter is more likely to affect the current level of the interest rate than past inflation.

Other than these few regularities, the posterior probabilities of inclusion of each predictor variable varies a lot between specifications. For the linear VAR model, the relatively uninformative ridge regression prior invites more restrictions from the variable selection algorithm than when the Minnesota and Normal-Jeffrey's priors are present. This is because the last two priors already provide shrinkage of coefficients towards zero. Subsequently it is the case that shrinkage will force more (compared to an uninformative prior) the posterior of the β_j's to move towards the region of zero, so that the respective γ_j's are not identified and they will be drawn randomly from their $Bernoulli(0.8)$ prior. As discussed earlier, this is not a failure of variable selection since what we care about is the combined coefficient $\theta_j = \gamma_j \beta_j$ to be zero, whether it is because $\beta_j = 0$ or $\gamma_j = 0$. An example where this effect happens is for variable π_{t-2} in the unemployment equation, which has only a probability of 8% of inclusion when using the VAR Ridge model, but this probability increases to circa 50% when using the VAR Min and VAR Shrink models. Nevertheless, in these two latter models, the posterior mean of β_j for $j = \pi_{t-2}$ is around 0.002, so that it finally holds that $\theta_j = \gamma_j \beta_j \approx 0$.

For the rest of the VAR models mixed results are present which depend on the nature of each model. Even among the two nonlinear models many differences exist. For instance, π_{t-1} has 0% probability of appearing in the unemployment equation of the structural breaks VAR but 100% probability of appearing in the same equation in the time-varying VAR model. Finally, notice that more restrictions are present in the Large-VAR model since a more restricted form of the prior on γ is used, compared to the one used in the small models. In this Large-VAR setting the right-hand side (RHS) variables have exactly the same probability of appearing in each of the three VAR equations of interest. This is due to the simplifying assumption described in subsection 3.4 which allows computational tractability when the dimensions of the VAR grow large.

4.5 Out-of-sample iterated forecasts

In this subsection the restricted and unrestricted VAR models are evaluated out-of-sample. Tables 3 and 4 present the MAFE and RMSFE statistics over the forecast sample 1990:Q1-2008:Q4. The first column of each table shows the three variables in the vector of interest y_{t+h}, for horizons $h = 1, ..., 4$. The second column of both tables presents the absolute value of the MAFE and RMSFE, respectively, for the driftless random walk model. Consequently the remaining columns present the MAFE and RMSFE statistics from the six Bayesian four-lag VARs with and without variable selection, as a proportion of the respective MAFE and RMSFE of the random walk. For comparison the third column in each table gives the respective statistics from a parsimonious VAR(1) specification estimated with OLS.

The results suggest that all small four-lag VAR models perform better the naïve model in short-term forecasting of unemployment and inflation. The very flexible TVP-VAR provides the lowest mean prediction error (the gains are especially visible during the financial crisis sample 2007-2008), while the Large VAR being quite heavily parametrized gives only the best VAR forecasts for the interest rate. Nevertheless, none of the VAR models can beat the random walk in interest rate forecasting.

In terms evaluating variable selection, the unrestricted VAR(4) model with ridge regression prior (which in this paper is defined to be uninformative, as if using a VAR(4) estimated with least squares) is better at all horizons than the unrestricted, more parsimonious VAR(1) in forecasting unemployment and inflation. In that respect, good performance of the variable selection is translated into expecting substantial restrictions of the VAR(4) Ridge model coefficients only in the interest rate equation since from the VAR(1) it is obvious that using one lag in this equation is always better. At the same time less restrictions are expected in the coefficients in the unemployment and interest rate equation, since the VAR(4) is already doing much better than the VAR(1) for these two equations. Table 2 provided an idea of the restrictions that actually hold in each model, however notice that in a recursive forecasting exercise the posterior probabilities are estimated in real-time as new data become available, so they will not be constant during the forecast evaluation sample.

Table 3: Relative MAFE of unrestricted and restricted VARs: unemployment, inflation and interest rate for $h = 1, 2, 3$ and 4.

	RW	*VAR(1)*	*VAR(4) Ridge*		*VAR(4) Min*		*VAR(4) Shrink*		*SB-VAR(4)*		*TVP-VAR(4)*		*Large-VAR(4)*	
	MAFE	OLS	no VS	with VS	no VS	with VS	no VS	with VS	no VS	with VS	no VS	with VS	no VS	with VS
u_{t+1}	0.1343	1.07	0.88	0.88	0.87	0.86	0.86	0.86	0.88	0.86	0.84	0.84	0.95	0.86
π_{t+1}	0.4882	1.18	0.91	0.88	0.90	0.88	0.88	0.89	0.90	0.90	0.83	0.83	1.24	1.21
r_{t+1}	0.4047	1.31	1.45	1.45	1.43	1.44	1.42	1.41	1.36	1.33	1.33	1.33	1.22	1.17
u_{t+2}	0.2163	1.14	0.94	0.92	0.93	0.93	0.92	0.91	0.94	0.86	0.83	0.84	1.03	0.90
π_{t+2}	0.8173	1.24	1.00	1.00	0.96	0.97	1.00	1.02	0.97	1.01	0.86	0.88	1.36	1.29
r_{t+2}	0.6971	1.38	1.59	1.55	1.57	1.52	1.50	1.47	1.50	1.47	1.39	1.37	1.09	1.01
u_{t+2}	0.2912	1.15	0.96	0.92	0.95	0.94	0.93	0.92	0.95	0.87	0.79	0.81	1.12	0.91
π_{t+3}	1.1014	1.35	1.18	1.16	1.12	1.09	1.13	1.14	1.15	1.11	0.88	0.91	1.64	1.53
r_{t+3}	0.9532	1.45	1.74	1.64	1.70	1.60	1.55	1.51	1.63	1.46	1.47	1.41	1.06	1.03
u_{t+4}	0.3479	1.16	0.99	0.95	0.99	0.98	0.97	0.96	0.99	0.88	0.80	0.83	1.22	0.93
π_{t+4}	1.2863	1.51	1.49	1.46	1.40	1.34	1.37	1.37	1.46	1.45	1.02	1.01	1.91	1.77
r_{t+4}	1.1868	1.52	1.79	1.70	1.74	1.65	1.58	1.54	1.68	1.49	1.49	1.44	1.10	1.04

Note: The second column shows the absolute MAFE of the Random Walk (RW). Remaining columns report MAFEs of each VAR model relative to the MAFE of the RW. No VS/ with VS indicates the presence of variable selection.

Table 4: Relative RMSFE of unrestricted and restricted VARs: unemployment, inflation and interest rate for $h = 1, 2, 3$ and 4.

	RW	*VAR(1)*	*VAR(4) Ridge*		*VAR(4) Min*		*VAR(4) Shrink*		*SB-VAR(4)*		*TVP-VAR(4)*		*Large-VAR(4)*	
	MAFE	OLS	no VS	with VS	no VS	with VS	no VS	with VS	no VS	with VS	no VS	with VS	no VS	with VS
u_{t+1}	0.1712	1.07	0.85	0.84	0.84	0.84	0.83	0.83	0.85	0.85	0.79	0.80	0.92	0.86
π_{t+1}	0.6828	1.10	0.85	0.84	0.85	0.86	0.87	0.88	0.85	0.89	0.79	0.80	1.13	1.11
r_{t+1}	0.6378	1.16	1.24	1.25	1.23	1.23	1.24	1.23	1.17	1.23	1.16	1.18	1.07	1.04
u_{t+2}	0.2933	1.04	0.84	0.81	0.83	0.82	0.82	0.81	0.83	0.81	0.75	0.77	0.91	0.84
π_{t+2}	1.1157	1.22	0.95	0.96	0.93	0.94	0.97	0.98	0.93	1.03	0.80	0.81	1.26	1.19
r_{t+2}	1.0278	1.23	1.37	1.33	1.34	1.32	1.30	1.29	1.29	1.28	1.24	1.24	1.00	0.93
u_{t+2}	0.4047	1.00	0.85	0.81	0.84	0.83	0.82	0.81	0.84	0.81	0.73	0.76	0.99	0.84
π_{t+3}	1.4596	1.37	1.16	1.16	1.12	1.10	1.12	1.13	1.13	1.08	0.89	0.91	1.55	1.47
r_{t+3}	1.3441	1.28	1.47	1.40	1.43	1.37	1.34	1.32	1.39	1.29	1.30	1.30	1.00	0.99
u_{t+4}	0.4986	0.96	0.82	0.79	0.82	0.81	0.80	0.79	0.82	0.79	0.70	0.73	1.05	0.83
π_{t+4}	1.7036	1.54	1.44	1.42	1.37	1.32	1.33	1.33	1.41	1.41	1.03	1.03	1.83	1.72
r_{t+4}	1.6406	1.31	1.49	1.43	1.46	1.40	1.36	1.33	1.41	1.30	1.31	1.28	1.03	1.02

Note: The second column shows the absolute MAFE of the Random Walk (RW). Remaining columns report MAFEs of each VAR model relative to the MAFE of the RW. No VS/ with VS indicates the presence of variable selection.

In fact, variable selection in the VAR(4) Ridge model does improve forecasts of all three variables, especially at longer horizons. For the VAR(4) Min and VAR(4) Shrink (these two models already have shrinkage priors) variable selection only improves the interest rate forecast while there is usually a $\pm 1\%$ gain/loss in MAFE or RMSFE, but this is so small that might also be attributed to sampling and rounding error. The main result is that none of the three unrestricted linear VARs with four lags is forecasting interest rates as the VAR(1) estimated with OLS does, something that is consistently accounted for when adding variable selection[9].

The gains from variable selection for forecasting all three variables of interest are more clear as the model size increases. As forecasting results for the 13-variable Large VAR suggest, when the model dimensions increase, variable selection really helps to prevent overfitting. Although the Minnesota shrinkage parameter is not set optimally, this improvement when using variable selection is robust for a large grid of values of λ (see the discussion in subsection 4.1).

The story behind the structural breaks model SB-VAR(4) is different. There, the gains are quite impressive for longer horizons, but closer examination shows that these are linked only indirectly to variable selection. Estimation of the unrestricted SB-VAR(4) model with maximum number of *possible* breaks equal to 3, indicates that there are actually no breaks[10]. When the SB-VAR(4) model is estimated with variable selection, a break is found (using the full sample) in 2004Q1. This is actually the exact reason why variable selection does much better in mean prediction with the structural breaks model. By restricting the parameter space, a structural break is found that is not otherwise identified when all 39 mean VAR coefficients are unrestricted.

In the TVP-VAR model with Minnesota prior, which is the best performing among all VAR models, variable selection helps improve the MAFE of the interest

[9]Here we can observe that although variable selection improves forecasts of interest rate from the linear VAR(4), these are never as good as the VAR(1)-OLS forecasts. This is due to the fact that our prior expection is that 20% of the parameters should be restricted ($\pi_{0j} = \pi_0 = 0.8$). Subsequently there might be benefit from setting $\pi_{0j} << 0.8$ but only if β_j is a coefficient in the interest rate equation; see also the discussion in the next subsection.

[10]Notice that although no breaks are estimated, the SB-VAR(4) forecasts are not the same as the VAR(4) Min forecasts (these two models have identical Minnesota priors). The reason is computational, but explaining why is beyond the scope of this paper. The reader is advised to consult Bauwens, Koop, Korobilis and Rombouts (2011).

rate in longer horizons. Nevertheless, in this case variable selection increases the absolute and squared forecast error of unemployment and inflation at horizons two to four quarters. Subsequently, the shrinkage prior in this case is sufficient to guarantee optimal mean forecasts, and variable selection is not necessary. Although this observation might be correct for the expected risk of mean forecasts, the Bayesian Diebold-Mariano (BDM) statistic given in equation (22) reveals that there is the case that variable selection provides overall superior predictive ability.

The BDM statistic, which is based on the time series of differences between the squared forecast errors of the restricted and the unrestricted models, is presented in Table 5. A value less than 0.5 shows the probability that the restricted model has better forecasting ability overall compared to the unrestricted model. Table 5 reveals that this is the case for all models apart from the structural breaks VAR. That is because in this model we saw that variable selection indicates one break, while in the unrestricted model no break is found. Thus forecasts from the restricted model with one break have larger variance because all the VAR coefficients in the second regime are estimated using only 19 observations (the break date is 2004Q1). Since the BDM statistic is based on all simulated draws from the posterior predictive densities, parameter uncertainty is included in the evaluation of the quantity $\Pr(d_{t+h} > 0)$. Thus, this fact explains why the unrestricted no-break model does better overall than the restricted model with one break, despite the fact that the MAFE and RMSFE results suggest otherwise. Finally, in Table 5 we can observe again that as the forecast horizon increases the gains from using variable selection also increase.

Table 5: Bayesian Diebold-Mariano statistic, $\frac{1}{T}\sum \Pr\left(d_{t+h} > 0\right)$.

	VAR Ridge	*VAR Min*	*VAR Shrink*	*SB-VAR*	*TVP-VAR*	*Large-VAR*
u_{t+1}	0.481	0.486	0.491	0.535	0.485	0.433
π_{t+1}	0.467	0.467	0.505	0.622	0.495	0.476
r_{t+1}	0.477	0.486	0.473	0.619	0.498	0.441
u_{t+2}	0.470	0.480	0.472	0.522	0.491	0.421
π_{t+2}	0.473	0.472	0.501	0.625	0.489	0.470
r_{t+2}	0.470	0.474	0.456	0.587	0.486	0.473
u_{t+3}	0.458	0.468	0.464	0.525	0.488	0.380
π_{t+3}	0.463	0.460	0.487	0.618	0.481	0.442
r_{t+3}	0.448	0.453	0.444	0.562	0.476	0.483
u_{t+4}	0.463	0.466	0.457	0.528	0.486	0.345
π_{t+4}	0.453	0.449	0.473	0.597	0.472	0.436
r_{t+4}	0.447	0.449	0.433	0.546	0.471	0.485

Note: The Table shows the average values of the statistic $\Pr(d_{t+h}>0)$ where d_{t+h} are the time series of differences between the squared forecast errors from the restricted and unrestricted models; see also equation (22) in the text.

4.6 Sensitivity analysis: Direct forecasts, and expected number of restrictions

In many cases, iterated, multi-step ahead VAR forecasts might not be satisfactory. This is particularly true when the model is misspecified (Marcellino, Stock and Watson, 2006), in which case econometricians estimate a direct VAR using information up to time t to directly predict y_{t+h}, i.e. the model

$$y_{t+h} = Bx_t + \varepsilon_t.$$

Using the above VAR equation, the researcher can use directly the available information x_T to forecast y_{T+h}. This is, additionally, a particularly useful approach when x_t contains exogenous predictors for which forecasts are not available to the econometrician (and hence iterating the VAR h-steps ahead is not possible).

This case is examined analytically in Korobilis (2008) using the SSVS algorithm in large linear VARs with hundreds of predictors. Here I provide results for 4-steps ahead forecasting using the TVP-VAR(4) in the context of a "sensitivity analysis"

with varying degree of prior expected number of restrictions. Restrictions in the VAR models with variable selection can be imposed through the prior hyperparameter π_{0j} of the Bernoulli density in equation (6). Table 6 presents the RMSFE from the unrestricted TVP-VAR(4) in the second column, and the RMSFE of the restricted TVP-VAR(4) with $\pi_{0j} = \pi_0$ for all $j = 1, ..., n$, relative to that of the unrestricted model. The case $\pi_0 = 0.8$ is the one examined previously in the small VARs (but it was relaxed in the Large VAR model) and implies the expectation that 20% of the coefficients should be restricted a priori. Other values shown in this Table can be interpreted in a similar way. The optimal forecasts from the restricted model are obtained when π_0 is 0.7, where gains of up to 8% in forecasting inflation are attained. When more and more restrictions are imposed, the RMSFE are monotonically increasing, suggesting that there is a risk attached to imposing strong prior beliefs in such a small model. For $\pi_0 > 0.7$ the RMSFE also increases, where the limit $\pi_0 = 1$ implies the unrestricted model (where all relative RMSFEs are equal to 1.00).

Table 6: RMSFE of 4-quarter ahead direct forecasts from a TVP-VAR(4)

	TVP-VAR(4)	TVP-VAR(4) with VS					
	no VS	$\pi_0 = .3$	$\pi_0 = .4$	$\pi_0 = .5$	$\pi_0 = .6$	$\pi_0 = .7$	$\pi_0 = .8$
u_{t+4}	0.3569	1.05	1.01	1.00	0.97	0.96	0.97
π_{t+4}	1.7546	0.93	0.94	0.92	0.92	0.92	0.93
r_{t+4}	1.9521	1.03	1.03	1.02	1.01	0.99	0.99

Note: The second column presents the RMSFE of the unrestricted TVP-VAR(4) model. The next columns present the RMSFEs of the restricted model (relative to that of the unrestricted TVP-VAR(4)) for different prior expected number of restrictions on γ.

Although for other direct VAR models and forecast horizons results are mixed as to whether variable selection improves forecasting over the unrestricted model, it is always the case that for small VAR models the RSMFE is a quadratic function of π_0. Consequently, choice of π_0 should not pose a challenge for the applied researcher as soon as the choice of expected restrictions is chosen reasonably, i.e. it is tied to the dimension of the VAR model considered. For instance, in subsection 3.4 an empirical method for tuning the prior expected number of restrictions as the dimension of the VAR increases was introduced. Moreover, if there are actually practical difficulties

in selecting a value for π_0, full Bayes methods can also be used. That means that a hyperpior distribution is placed on π_0 (or even π_{0j} for $j = 1, ..., n$), so that this hyperparameter is estimated from the data and hence it will also vary with the sample size considered.

5 Concluding remarks

Vector autoregressive models have been used extensively over the past for the purpose of macroeconomic forecasting, since they have the ability to fit the observed data better than competing theoretical and large-scale structural macroeconometric models. This paper shows that Bayesian variable selection methods can be used to find restrictions based on the evidence in the data with positive implications in preserving parsimony. It was argued that these types of restrictions are important for long-horizon forecasts as well as forecasts from large VAR systems. Specifically, variable selection i) dominates forecast from VAR models with uninformative priors; ii) competes favourably to shrinkage estimation; and iii) provides more benefits in forecasting as the model size increases.

References

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

Barbieri, M. M., and J. O. Berger. (2004). Optimal predictive model selection. The Annals of Statistics, 32, 870-897.

Bauwens, L., Koop, G., Korobilis, D., and J. Rombouts. (2011). A comparison of forecasting procedures for macroeconomic series: The contribution of structural break models. CIRANO Working Papers 2011s-13, CIRANO.

Canova, F. (1993). Modelling and forecasting exchange rates using a Bayesian time varying coefficient model. Journal of Economic Dynamics and Control,17, 233-262.

Canova, F., and L. Gambetti. (2009). Structural changes in the US economy: Is there a role for monetary policy? Journal of Economic Dynamics and Control, 33, 477-490.

Carter, C., and R. Kohn (1994). On Gibbs sampling for state space models. Biometrika, 81, 541-553.

Chan, J. C. C., Koop, G., Leon-Gonzalez, R., and Strachan, R. W. (2010). Time-varying dimension models. ANU School of Economics Working Papers 2010-523.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. Journal of Econometrics, 75, 79-98.

Chipman, H., George, E. I., and R.E. McCulloch. (2001). The practical implementation of Bayesian model selection. In P. Lahiri (Ed.), Model Selection, (pp. 67-116). IMS Lecture Notes – Monograph Series, vol. 38.

Cogley, T., Morozov, S., and T. Sargent. (2005). Bayesian fan charts for U.K. inflation: Forecasting and sources of uncertainty in an evolving monetary system. Journal of Economic Dynamics and Control, 29, 1893-1925.

Cogley, T., and T. Sargent. (2002). Evolving post-World War II inflation dynamics. NBER Macroeconomics Annual, 16, 331-388.

Clark, T. E., and M. W. McCracken. (2010). Averaging forecasts from VARs with uncertain instabilities. Journal of Applied Econometrics, 25, 5-29.

D'Agostino, A., Gambetti, L., and D. Giannone. (2009). Macroeconomic forecasting and structural change. ECARES Working Paper 2009-020.

Diebold, F. X. and R. S. Mariano. (1995). Comparing predictive accuracy. Journal of Business and Economic Statistics, 13, 253-263.

Doan, T., R. Litterman, and C. A. Sims. (1984). Forecasting and conditional projection using realistic prior distributions. Econometric Reviews, 3, 1-100.

Garratt, A., Koop, G., Mise, E. & S. P. Vahey (2009). Real-time prediction with U.K. monetary aggregates in the presence of model uncertainty. Journal of Business and Economic Statistics, 27, 480-491.

George, E. I., Sun, D. and S. Ni. (2008). Bayesian stochastic search for VAR model restrictions. Journal of Econometrics, 142, 553-580.

Groen, J., Paap, R., and F. Ravazzolo. (2009). Real-time inflation forecasting in a changing world. Unpublished manuscript.

Jochmann, M., Koop, G., and R.W. Strachan. (2010). Bayesian forecasting using stochastic search variable selection in a VAR subject to breaks. International Journal of Forecasting, 26, 326-347.

Kohn, R., Smith, M., and D. Chan. (2001). Nonparametric regression using linear combinations of basis functions. Statistics and Computing, 11, 313-322.

Koop, G., and D. Korobilis. (2009a). Bayesian Multivariate time series methods for empirical macroeconomics. Foundations and Trends in Econometrics, 3, 267-358.

Koop, G., and D. Korobilis. (2009b). Forecasting inflation using dynamic model averaging. RCEA Working Paper 34-09.

Koop, G., Leon-Gonzalez, R., and R. Strachan. (2009). On the evolution of the monetary policy transmission mechanism. Journal of Economic Dynamics and Control, 33, 997-1017.

Koop, G., and S. M. Potter. (2007). Estimation and forecasting in models with multiple breaks. The Review of Economics and Statistics, 74, 763-789.

Korobilis, D. (2008). Forecasting in vector autoregressions with many predictors. Advances in Econometrics, 23, 403-431.

Korobilis, D. (2011). Hierarchical shrinkage priors for dynamic regressions with many predictors. Unpublished manuscript.

Kuo, L., and B. Mallick. (1997). Variable selection for regression models. Shankya: The Indian Journal of Statistics, 60 (Series B), 65-81.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g-priors for Bayesian Variable Selection. Journal of the American Statistical Association, 103, 410-423.

Litterman, R. (1986). Forecasting with Bayesian vector autoregressions - 5 years of experience. Journal of Business and Economic Statistics, 4, 25-38.

Marcellino, M., Stock, J. H. and M. W. Watson. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. Journal of Econometrics, 135, 499-526.

Park, T. and Casella, G. (2008). The Bayesian Lasso. Journal of the American Statistical Association, 103, 681-686.

Primiceri, G. (2005). Time varying structural vector autoregressions and monetary policy. Review of Economic Studies, 72, 821-852.

Sims, C. (1980). Macroeconomics and reality. Econometrica 48, 1-80.

Smith, M., and R. Kohn. (2002). Parsimonious covariance matrix estimation for longitudinal data. Journal of the American Statistical Association, 97, 1141-1153.

Stock, J. H., and M.W. Watson. (2006). Forecasting with Many Predictors." In G Elliott, CWJ Granger, A Timmermann (eds.), Handbook of Economic Forecasting, volume 1, chapter 10, pp. 515-658. Elsevier, Amsterdam.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 58, 267-288.

Tierney, L. (1991). Markov Chains for Exploring Posterior Distributions. University of Minnesota School of Statistics Technical Report No. 560.

Villani, M. (2009). Steady-state priors for vector autoregressions. Journal of Applied Econometrics, 24, 630-650.

Wong, F., Carter, C. K., and R. Kohn. (2003). Efficient estimation of covariance selection models. Biometrika, 90, 809-830.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Series B 67, 301-320.

Technical Appendix

A Posterior inference in the linear VAR with variable selection

In this section I provide exact details on the conditional densities of the restricted VAR model. For simplicity rewrite the priors, which are

$$\begin{aligned} \beta &\sim N_n\left(\underline{b}, \underline{V}\right) && \text{(A.1)} \\ \gamma_j|\gamma_{\setminus -j} &\sim Bernoulli\left(\pi_{0j}\right) && \text{(A.2)} \\ \Sigma &\propto |\Sigma|^{-(m+1)/2} && \text{(A.3)} \end{aligned}$$

A.1 Algorithm 1

Given the prior hyperparameters $(\underline{b}, \underline{V}, \pi_0, \Psi, \alpha)$ and an initial value for γ, Σ, sampling from the conditional distributions proceeds as follows

1. Sample β from the density

$$\beta|\gamma, \Sigma, y, z \sim N_n\left(\widetilde{b}, \widetilde{V}\right) \quad \text{(A.4)}$$

   where $\widetilde{V} = \left(\underline{V}^{-1} + \sum_{t=1}^{T} z_t^{*\prime}\Sigma^{-1}z_t^*\right)^{-1}$ and $\widetilde{b} = \widetilde{V}\left(\underline{V}^{-1}\underline{b} + \sum_{t=1}^{T} z_t^{*\prime}\Sigma^{-1}y_{t+h}\right)$, and $z_t^* = z_t\Gamma$.

2. Sample γ_j, $j = 1, ..., n$, from the density

$$\gamma_j|\gamma_{\setminus -j}, \beta, \Sigma, y, z \sim Bernoulli\left(\widetilde{\pi}_j\right) \quad \text{(A.5)}$$

preferably in random order j, where $\widetilde{\pi}_j = \frac{l_{0j}}{l_{0j}+l_{1j}}$, and

$$l_{0j} = p\left(y|\theta_j, \gamma_{\setminus -j}, \gamma_j = 1\right)\pi_{0j} \quad (A.6)$$

$$l_{1j} = p\left(y|\theta_j, \gamma_{\setminus -j}, \gamma_j = 0\right)(1-\pi_{0j}) \quad (A.7)$$

The expressions $p\left(y|\theta_j, \gamma_{\setminus -j}, \gamma_j = 1\right)$ and $p\left(y|\theta_j, \gamma_{\setminus -j}, \gamma_j = 0\right)$ are conditional likelihood expressions. Define θ^* to be equal to θ but with its $j-th$ element $\theta_j = \beta_j$ (i.e. when $\gamma_j = 1$). Similarly, define θ^{**} to be equal to θ but with the $j-th$ element $\theta_j = 0$ (i.e. when $\gamma_j = 0$). Then in the case of the VAR likelihood of model (4), we can write l_{0j}, l_{1j} analytically as

$$\begin{aligned} l_{0j} &= \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left(y_t - z_t\theta^*\right)'\Sigma^{-1}\left(y_t - z_t\theta^*\right)\right)\pi_{0j} \\ l_{1j} &= \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left(y_t - z_t\theta^{**}\right)'\Sigma^{-1}\left(y_t - z_t\theta^{**}\right)\right)(1-\pi_{0j}). \end{aligned}$$

3. Sample Σ^{-1} from the density

$$\Sigma^{-1}|\beta, \gamma, y, z \sim Wishart\left(T, S^{-1}\right) \quad (A.8)$$

where $S = \sum_{t=1}^{T}\left(y_t - z_t\theta\right)'\left(y_t - z_t\theta\right)$.

A.2 Algorithm 2

In modern matrix programming languages it is more efficient to replace "for" loops with matrix multiplications (what is called "vectorizing loops"). This section provides a reformulation of the VAR, so that the summations in the Gibbs sampler algorithm (A.4) - (A.8) are replaced by matrix multiplications. For example, computing l_{0j} and l_{1j} requires to evaluate $\sum_{t=1}^{T}\left(y_t - z_t\theta^*\right)'\Sigma^{-1}\left(y_t - z_t\theta^*\right)$ for $t = 1, ..., T$. In practice, it is more efficient to use the matrix form of the VAR likelihood:

Begin from formulation (1), and let $y = (y_1', ..., y_T')$, $x = (x_1', ..., x_T')$ and $\varepsilon = (\varepsilon_1', ..., \varepsilon_T')$. A different SUR formulation of the VAR takes the form

$$vec(y) = (I_m \otimes x')\,\Gamma\beta^* + vec(\varepsilon) \tag{A.9}$$

$$Y = W\theta + e \tag{A.10}$$

where $Y = vec(y)$ is a $(Tn) \times 1$ column vector, $W = I_m \otimes x$ is a block diagonal matrix of dimensions $(Tn) \times m$ with the matrix x replicated m times on its diagonal, $\theta = \Gamma\beta^*$ is a $m \times 1$ vector, $\beta^* = vec(B')$ and $e = vec(\varepsilon) \sim N(0, \Sigma \otimes I_T)$. To clarify notation, $vec(\circ)$ is the operator that stacks the columns of a matrix and $\otimes$ is the Kronecker product. In this formulation, $W = I_m \otimes x$ is not equal to $z = (z_1', ..., z_T') = \left((I_m \otimes x_1)', ..., (I_m \otimes x_T)'\right)$ which was defined in (4). Additionally, note that while β and β^* are both $n \times 1$ vectors, they are not equal. It holds that $\beta = vec(B)$ and $\beta^* = vec(B')$.

The priors are exactly the same as the ones described in the main text. The conditional posteriors of this formulation are given by

1. Sample β^* from the density

$$\beta^*|\gamma, \Sigma, Y, W \sim N_n\left(\widetilde{b}, \widetilde{V}\right) \tag{A.11}$$

   where $\widetilde{V} = \underline{V}^{-1} + W^{*\prime}(\Sigma^{-1} \otimes I_T) W^*$ and $\widetilde{b} = \widetilde{V}\left(\underline{V}^{-1}\underline{b} + W^{*\prime}(\Sigma^{-1} \otimes I_T) Y\right)$, and $W^* = W\Gamma$.

2. Sample γ_j, $j = 1, ..., n$, from the density

$$\gamma_j|\gamma_{\setminus -j}, \beta^*, \Sigma, Y, W \sim Bernoulli(\widetilde{\pi}_j) \tag{A.12}$$

   preferably in random order j, where $\widetilde{\pi}_j = \frac{l_{0j}}{l_{0j}+l_{1j}}$, and

$$l_{0j} = \exp\left(-\frac{1}{2}(Y - W\theta^*)'\left(\Sigma^{-1} \otimes I_T\right)(Y - W\theta^*)\right)\pi_{0j}$$

$$l_{1j} = \exp\left(-\frac{1}{2}(Y - W\theta^{**})'\left(\Sigma^{-1} \otimes I_T\right)(Y - W\theta^{**})\right)(1 - \pi_{0j}).$$

3. Sample Σ^{-1} from the density

$$\Sigma^{-1}|\gamma, \beta^*, Y, x \sim Wishart\left(T, S^{-1}\right)$$

where $S = (y - x\Theta)'(y - x\Theta)$, with Θ the $k \times m$ matrix obtained from the vector $\theta = \Gamma\beta^*$, which has elements $(\Theta_{ij}) = \theta_{(j-1)k+i}$, for $i = 1, ..., k$ and $j = 1, ..., m$.

This sampler has slight modifications compared to the one above because of the different specification of the likelihood function, but the two SUR specifications are equivalent and produce the same results. Posterior inference in the TVP-VAR model is just a simple generalization of the VAR case and it is described in the next section.

A.3 Sampling from a VAR with Normal-Jeffreys' prior

The previous results hold for the linear VAR models when the prior covariance matrix $\underline{V}$ is known. If instead a Jeffreys' prior is placed on the diagonal elements λ_j, $j = 1, ..., n$, of $\underline{V}$ as in the case of the prior in (13) one needs to sample these elements using the following step which is added to previous VAR model algorithms

4. Sample λ_j^{-1} for each $j = 1, ..., n$ from the density

$$\frac{1}{\lambda_j}|\beta, \gamma, \Sigma, y, z \sim Gamma\left(\frac{1}{2}, \frac{\beta_j^2}{2}\right)$$

Then sampling of β proceeds conditional on all sampled λ_j's, i.e. whenever $\underline{V}$ shows up in the posterior of β in step 1, we use the matrix $\underline{V} = diag\{\lambda_1, ..., \lambda_n\}$.

B Posterior inference in the TVP-VAR with variable selection

The homoskedastic TVP-VAR with variable selection is of the form

$$y_t = z_t\theta_t + \varepsilon_t \tag{B.1}$$
$$\beta_t = \beta_{t-1} + \eta_t \tag{B.2}$$

where $\theta_t = \Gamma\beta_t$, and $\varepsilon_t \sim N(0, \Sigma)$ and $\eta_t \sim N(0, Q)$ which are uncorrelated with each other at all leads and lags. The priors for this model are:

$$\begin{aligned}
\beta_0 &\sim N_n(\underline{b}, \underline{V}) \\
\gamma_j|\gamma_{\setminus -j} &\sim Bernoulli(\pi_{0j}) \\
Q^{-1} &\sim Wishart\left(\xi, R^{-1}\right) \\
\Sigma &\propto |\Sigma|^{-(m+1)/2}
\end{aligned}$$

Estimating these parameters means sampling sequentially from the following conditional densities

1. Sample β_t for all t, conditioning on data $z_t^* = z_t\Gamma$ with $\Gamma = diag\{\gamma_1, ..., \gamma_n\}$, using the Carter and Kohn (1994) filter and smoother for state-space models (see below)

2. Sample γ_j, $j = 1, ..., n$, from the density

$$\gamma_j|\gamma_{\setminus -j}, \beta, Q, \Sigma, y, z \sim Bernoulli(\widetilde{\pi}_j) \tag{B.3}$$

preferably in random order j, where $\widetilde{\pi}_j = \frac{l_{0j}}{l_{0j}+l_{1j}}$, and

$$l_{0j} = p\left(y|\theta_j^{1:T}, \gamma_{\setminus -j}, \gamma_j = 1\right)\pi_{0j} \tag{B.4}$$
$$l_{1j} = p\left(y|\theta_j^{1:T}, \gamma_{\setminus -j}, \gamma_j = 0\right)(1 - \pi_{0j}) \tag{B.5}$$

The expressions $p\left(y|\theta_j^{1:T}, \gamma_{\setminus -j}, \gamma_j = 1\right)$ and $p\left(y|\theta_j^{1:T}, \gamma_{\setminus -j}, \gamma_j = 0\right)$ are conditional likelihood expressions, where $\theta_j^{1:T} = [\theta_{1,j}, ..., \theta_{t,j}, ..., \theta_{T,j}]'$. Define θ_t^* to be equal to θ_t but with its $j-th$ element $\theta_{t,j} = \beta_{t,j}$ (i.e. when $\gamma_j = 1$). Similarly, define θ_t^{**} to be equal to θ_t but with the $j-th$ element $\theta_{t,j} = 0$ (i.e.

when $\gamma_j = 0$), for all $t = 1, ..., T$. Then in the case of the TVP-VAR likelihood of model (B.1), we can write l_{0j}, l_{1j} analytically as

$$\begin{aligned} l_{0j} &= \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left(y_t - z_t\theta_t^*\right)'\Sigma^{-1}\left(y_t - z_t\theta_t^*\right)\right)\pi_{0j} \\ l_{1j} &= \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left(y_t - z_t\theta_t^{**}\right)'\Sigma^{-1}\left(y_t - z_t\theta_t^{**}\right)\right)\left(1 - \pi_{0j}\right). \end{aligned}$$

3. Sample Q^{-1} from the density

$$Q^{-1}|\beta, \gamma, \Sigma, y, z \sim Wishart\left(\widetilde{\xi}, \widetilde{R}^{-1}\right) \tag{B.6}$$

where $\widetilde{\xi} = T + \xi$ and $\widetilde{R}^{-1} = \left(R + \sum_{t=1}^{T}\left(\beta_t - \beta_{t-1}\right)'\left(\beta_t - \beta_{t-1}\right)\right)^{-1}$.

4. Sample Σ^{-1} from the density

$$\Sigma^{-1}|\beta, Q, \gamma, y, z \sim Wishart\left(T, S^{-1}\right) \tag{B.7}$$

where $S = \sum_{t=1}^{T}\left(y_t - z_t\theta_t\right)'\left(y_t - z_t\theta_t\right)$.

B.1 Carter and Kohn (1994) algorithm:

Consider a general state-space model of the following form

$$y_t = z_t a_t + u_t \tag{B.8a}$$

$$a_t = a_{t-1} + v_t \tag{B.8b}$$

$$u_t \sim N\left(0, R\right), v_t \sim N\left(0, W\right)$$

where (B.8a) is the measurement equation and (B.8b) is the state equation, with observed data y_t and unobserved state a_t. If the errors u_t, v_t are iid and uncorrelated with each other, we can use the Carter and Kohn (1994) algorithm to obtain a draw from the posterior of the unobserved states.

Let $a_{t|s}$ denote the expected value of a_t and $P_{t|s}$ its corresponding variance, using data up to time s. Given starting values $a_{0|0}$ and $P_{0|0}$, the Kalman filter recursions provide us with initial filtered estimates:

$$\begin{aligned} a_{t|t-1} &= a_{t-1|t-1} \\ P_{t|t-1} &= P_{t-1|t-1} + W \\ K_t &= P_{t|t-1} z_t' \left(z_t P_{t|t-1} z_t + R \right)^{-1} \\ a_{t|t} &= a_{t|t-1} + K_t \left(y_t - z_t a_{t|t-1} \right) \\ P_{t|t} &= P_{t|t-1} - K_t z_t P_{t|t-1} \end{aligned} \tag{B.9}$$

The last elements of the recursion are $a_{T|T}$ and $P_{T|T}$ for which are used to obtain a single draw of a_T. However for periods $T-1, ..., 1$ we can smooth our initial Kalman filter estimates by using information from subsequent periods. That is, we run the backward recursions for $t = T-1, ..., 1$ and obtain the smooth estimates $a_{t|t+1}$ and $P_{t|t+1}$ given by the backward recursion:

$$\begin{aligned} a_{t|t+1} &= a_{t|t} + P_{t|t} P_{t+1|t}' \left(a_{t+1} - a_{t|t} \right) \\ P_{t|t+1} &= P_{t|t} - P_{t|t} P_{t+1|t}' P_{t|t} \end{aligned}$$

Then we can draw from the posterior of a_t by simply drawing from a Normal density with mean $a_{t|t+1}$ and variance $P_{t|t+1}$ (for $t = T$ we use $a_{T|T}$ and $P_{T|T}$).

C Posterior inference in the structural breaks VAR with variable selection

Having described the TVP-VAR with variable selection, the structural breaks VAR is a special case of this model and takes the form

$$y_t = z_t \theta_{s_t} + \varepsilon_t \tag{C.1}$$

$$\beta_{s_t} = \beta_{s_{t-1}} + \eta_{s_t} \tag{C.2}$$

The full set of prior distributions for this model are

$$\begin{aligned}
\beta_0 &\sim N_n\left(\underline{b}, \underline{V}\right) \\
\gamma_j|\gamma_{\backslash -j} &\sim Bernoulli\left(\pi_{0j}\right) \\
p_{ii} &\sim Beta\left(\rho_1, \rho_2\right) \\
Q^{-1} &\sim Wishart\left(\xi, R^{-1}\right) \\
\Sigma &\propto |\Sigma|^{-(m+1)/2}
\end{aligned}$$

where $j = 1, ..., n$ and $i = 1, ..., K$.

Estimating these parameters means sampling sequentially from the following conditional densities

1. Sample β_{s_t} for all t, conditioning on data $z_t^* = z_t\Gamma$ with $\Gamma = diag\left\{\gamma_1, ..., \gamma_n\right\}$, using the modified Carter and Kohn (1994) filter and smoother for state-space models (see below)

2. Sample γ_j, $j = 1, ..., n$, from the density

$$\gamma_j|\gamma_{\backslash -j}, \beta, Q, P, \Sigma, y, z \sim Bernoulli\left(\widetilde{\pi}_j\right) \tag{C.3}$$

preferably in random order j, where $\widetilde{\pi}_j = \frac{l_{0j}}{l_{0j}+l_{1j}}$, and

$$l_{0j} = p\left(y|\theta_j, \gamma_{\backslash -j}, \gamma_j = 1\right)\pi_{0j} \tag{C.4}$$

$$l_{1j} = p\left(y|\theta_j, \gamma_{\backslash -j}, \gamma_j = 0\right)\left(1 - \pi_{0j}\right) \tag{C.5}$$

The expressions $p\left(y|\theta_j^{1:s_T}, \gamma_{\backslash -j}, \gamma_j = 1\right)$ and $p\left(y|\theta_j^{1:s_T}, \gamma_{\backslash -j}, \gamma_j = 0\right)$ are conditional likelihood expressions, where $\theta_j^{1:s_T} = [\theta_{s_1,j}, ..., \theta_{s_t,j}, ..., \theta_{s_T,j}]'$. Define $\theta_{s_t}^*$ to be equal to θ_{s_t} but with its $j-th$ element fixed to $\theta_{s_t,j} = \beta_{s_t,j}$ (i.e. when $\gamma_j = 1$). Similarly, define $\theta_{s_t}^{**}$ to be equal to θ_{s_t} but with the $j-th$ element set to $\theta_{s_t,j} = 0$ (i.e. when $\gamma_j = 0$), for all $t = 1, ..., T$. Then in the case of the

TVP-VAR likelihood of model (B.1), we can write l_{0j}, l_{1j} analytically as

$$\begin{aligned} l_{0j} &= \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left(y_t - z_t\theta^*_{s_t}\right)'\Sigma^{-1}\left(y_t - z_t\theta^*_{s_t}\right)\right)\pi_{0j} \\ l_{1j} &= \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left(y_t - z_t\theta^{**}_{s_t}\right)'\Sigma^{-1}\left(y_t - z_t\theta^{**}_{s_t}\right)\right)(1-\pi_{0j}). \end{aligned}$$

3. Sample Q^{-1} from the density

$$Q^{-1}|\beta,\gamma,P,\Sigma,y,z \sim Wishart\left(\widetilde{\xi},\widetilde{R}^{-1}\right) \tag{C.6}$$

where $\widetilde{\xi} = T + \xi$ and $\widetilde{R}^{-1} = \left(R + \sum_{t=1}^{T}\left(\beta_{s_t} - \beta_{s_t-1}\right)'\left(\beta_{s_t} - \beta_{s_t-1}\right)\right)^{-1}$.

4. Sample Σ^{-1} from the density

$$\Sigma^{-1}|\beta,Q,P,\gamma,y,z \sim Wishart\left(T,S^{-1}\right) \tag{C.7}$$

where $S = \sum_{t=1}^{T}\left(y_t - z_t\theta_{s_t}\right)'\left(y_t - z_t\theta_{s_t}\right)$.

5. Sample s_t using Chib's (1996) algorithm.

6. Sample p_{ii} from the density

$$p_{ii}|\beta,Q,\Sigma,\gamma,y,z \sim Beta\left(\rho_1 + T_i, \rho_2 + 1\right)$$

where T_i are the number of observations in regime i (i.e. number of time periods for which $s_t = i$), $i = 1, ..., K$.

C.1 Modified Carter and Kohn (1994) algorithm for structural breaks VAR:

Consider the following special state-space form

$$y_t = z_t a_{s_t} + u_t \tag{C.8a}$$

$$a_{s_t} = a_{s_{t-1}} + v_{s_t} \tag{C.8b}$$

$$u_t \sim N(0, R),\ v_t \sim N(0, W)$$

When structural breaks indicators s_t are present, the Kalman filter and smoother have to be modified. The main idea is that in the standard Kalman filter we have a break in each period, so that $s_t = t$ and at the end of the sample $s_T = T$. Subsequently, when $s_t < t$ (a few breaks model) we run the Kalman filter for $t = 1, ..., T$, with the exception that the second filtering equation in (B.9) takes the form

$$P_{t|t-1} = \begin{cases} P_{t-1|t-1} + W,\ if\ s_t \neq s_{t-1} \\ P_{t-1|t-1},\ otherwise \end{cases}$$

In order to get the smoothed estimates of a_j for $j = 1, ..., s_T$ we run the backward recursions

$$a_{t|t+1} = \begin{cases} a_{t|t} + P_{t|t}P'_{t+1|t}\left(a_{t+1} - a_{t|t}\right),\ if\ s_t \neq s_{t-1} \\ a_{t|t},\ otherwise \end{cases}$$

$$P_{t|t+1} = \begin{cases} P_{t|t} - P_{t|t}P'_{t+1|t}P_{t|t},\ if\ s_t \neq s_{t-1} \\ P_{t|t},\ otherwise \end{cases}$$

for $t = T-1, ..., 1$ and draw $a_{s_t} \sim N\left(a_{t|t+1}, P_{t|t+1}\right)$, $iff\ s_t \neq s_{t-1}$.

D Efficient sampling of the variable selection indicators

In order to sample all the γ_j we need n evaluations of the conditional likelihood functions $p\left(y|..., \gamma_j = 1\right)$ and $p\left(y|..., \gamma_j = 0\right)$ which can be quite inefficient for large n. Kohn, Smith and Chan (2001) replace step 2 of the algorithms above with step 2* below. For notational convenience denote S to be the total number of Gibbs draws, and let the (current) value of γ_j at iteration s of the Gibbs sampler to be denoted by γ_j^s, and the (candidate) draw of γ_j at iteration $s+1$ to be denoted by γ_j^{s+1}. An efficient accept/reject step for generating γ_j is:

2* a) Draw a random number g from the continuous Uniform distribution $U(0,1)$.

b) - If $\gamma_j^s = 1$ and $g > \pi_{0j}$, set $\gamma_j^{s+1} = 1$.

- If $\gamma_j^s = 0$ and $g > 1 - \pi_{0j}$, set $\gamma_j^{s+1} = 0$.

- If $\gamma_j^s = 1$ and $g < \pi_{0j}$ or $\gamma_j^s = 0$ and $g < 1 - \pi_{0j}$, then generate γ_j^{s+1} from the Bernoulli density $\gamma_j|\gamma_{\setminus -j}, b, y, z \sim Bernoulli\left(\widetilde{\pi}_j\right)$, where $\widetilde{\pi}_j = \frac{l_{0j}}{l_{0j}+l_{1j}}$ and l_{0j}, l_{1j} are given in equations (A.6)-(A.7) and (B.4)-(B.5), for the VAR and TVP-VAR models respectively.