# VAR forecasting using Bayesian variable selection

Korobilis, Dimitris

December 2009

# VAR forecasting using Bayesian variable selection

Dimitris Korobilis*

December 2009

**Abstract**

This paper develops methods for automatic selection of variables in forecasting Bayesian vector autoregressions (VARs) using the Gibbs sampler. In particular, I provide computationally efficient algorithms for stochastic variable selection in generic (linear and nonlinear) VARs. The performance of the proposed variable selection method is assessed in a small Monte Carlo experiment, and in forecasting 4 macroeconomic series of the UK using time-varying parameters vector autoregressions (TVP-VARs). Restricted models consistently improve upon their unrestricted counterparts in forecasting, showing the merits of variable selection in selecting parsimonious models.

***Keywords:*** Forecasting; variable selection; time-varying parameters; Bayesian

***JEL Classification:*** C11, C32, C52, C53, E37, E47

*Department of Economics, University of Strathclyde, and the Rimini Center for Economic Analysis. Address for correspondence: 130 Rottenrow, G4 0GE - Glasgow, United Kingdom. Tel: (+44) 141 41.63.555.

# 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 problematic predictions. This pitfall of VAR modelling was recognized early and shrinkage methods have been proposed; see for example the so-called Minnesota prior (Doan et al., 1984). Nowadays the toolbox of applied econometricians includes numerous efficient modelling tools to prevent the proliferation of parameters and eliminate parameter/model uncertainty, like variable selection priors (George et al. 2008), steady-state priors (Villani, 2009), Bayesian model averaging (Andersson and Karlsson, 2008) and factor models (Stock and Watson, 2005), 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 (say estimate all possible model combinations and decide on the best model), the algorithm will visit only the most probable models. 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

$$y_{t+1} = Bx_t + \varepsilon_{t+1} \tag{1}$$

where $y_{t+1}$ is an $m \times 1$ vector of $t = 1, ..., T$ time series observations on the dependent variables, the vector $x_t$ is of dimensions $k \times 1$ and may contain an intercept, lags of the dependent variables, trends, dummies and exogenous regressors, and $B$ is a $m \times k$ matrix of regression coefficients. The errors $\varepsilon_t$ are assumed to be $N(0, \Sigma)$, where $\Sigma$ is an $m \times m$ covariance matrix. The idea behind Bayesian variable selection is to introduce indicators $\gamma_{ij}$ 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 matrix $B$, for $i = 1, .., m$ and $j = 1, ..., k$.

There are various benefits of using this approach over the shrinkage methods mentioned previously. 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 $\gamma_{ij}$ indicate which elements of $B$ should be included or excluded from the final optimal model, thus implementing a selection among all possible $2^{m \times k}$ VAR model combinations, without the need to estimate each and everyone of these models. Second, this form of Bayesian variable selection is independent of the prior assumptions about the parameters $B$. That is, if the researcher has defined any desirable prior for her pa-

rameters of the unrestricted model (1), adopting the variable selection restriction (2) needs no other modification than one extra block in the posterior sampler that draws from the conditional posterior of the $\gamma_{ij}$'s. Finally, unlike other proposed stochastic search variable selection algorithms for VAR models (George et al. 2008, Korobilis, 2008), this form of variable selection may be adopted in many nonlinear extensions of the VAR models.

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, see for example Canova and Gambetti (2009), Cogley and Sargent (2005), Cogley et al. (2005), Koop et al. (2009) and Primiceri (2005). Common feature of these papers is that they all fix the number of autoregressive lags to 2 for parsimony. That is because marginal likelihoods are difficult to obtain in the case of time-varying parameters. Therefore, automatic variable selection is a convenient and fast way to overcome the computational and practical problems associated with models where parameters drift at each point in time. Although the methods described in this paper can be used for structural analysis (by providing data-based restrictions on parameters, useful for identifying monetary policy for instance), the aim is to show how more parsimonious models can be selected with positive impact in macroeconomic forecasting.

In particular, the next section describes the mechanichs behind variable selection in VAR and TVP-VAR models. In Section 3, the performance of the variable selection algorithm is assessed using a small Monte Carlo exercise. The paper concludes by evaluating the out-of-sample forecasting performance VAR models with variable selection, by computing pseudo-forecasts of 4 UK macroeconomic variables over the sample period 1971:Q1 - 2008:Q4.

# 2 Variable selection in vector autoregressions

## 2.1 The standard VAR model

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

$$y_{t+1} = z_t\beta + \varepsilon_{t+1} \tag{3}$$

where $z_t = I_m \otimes x_t'$ 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 $\gamma$ 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

$$y_{t+1} = z_t\theta + \varepsilon_t \tag{4}$$

where $\theta = \Gamma\beta$. Here $\Gamma$ 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 $\Gamma_{jj} = 0$ then $\theta_j$ is restricted and is equal to $\Gamma_{jj}\beta_j = 0$, while for $\Gamma_{jj} = 1$, $\theta_j = \Gamma_{jj}\beta_j = \beta_j$, so that all possible $2^n$ specifications can be explored and variable selection in this case is equivalent to model selection.

The Gibbs sampler provides a natural framework to estimate these parameters, by drawing sequentially from the conditional posterior of each parameter. In fact, sampling the restriction indices $\gamma$ just adds one more block to the Gibbs sampler of the unrestricted VAR model[2]. For example, the full conditional (i.e. conditional on the data and $\Gamma$) densities of $\beta$ and $\Sigma$ are of standard form, assuming the so-called independent Normal-Wishart prior. For the restriction indicators we need to sample the $n$ elemenst in the column vector $\gamma = (\gamma_1, ..., \gamma_n)'$, and then recover the diagonal matrix $\Gamma = diag\{\gamma_1, ..., \gamma_n\}$ only when computations require it. Derivations are simplified if the indicators $\gamma_j$ are 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 $j - th$, so that a conjuagate prior for each $\gamma_j$ is the independent Bernoulli density. In particular, define the priors

$$\begin{aligned} \beta &\sim N_n(b_0, V_0) && (5)\\ \gamma_j|\gamma_{\setminus -j} &\sim Bernoulli(1, \pi_{0j}) && (6)\\ \Sigma^{-1} &\sim Wishart\left(\alpha, S^{-1}\right) && (7) \end{aligned}$$

where $b_0$ is $n \times 1$ and $V_0$ is $n \times n$, $\pi_0 = (\pi'_{01}, ..., \pi'_{0n})$ is $n \times 1$, $\Omega$ is a $m \times m$ matrix, and $\alpha$ a scalar. Note that the algorithm presented below does not depend on the asumption about the prior distribution of $(\beta, \Sigma)$. The Normal-Wishart form is used here only for illustration and because it is a standard conjugate choice in Bayesian analysis which makes computations of the conditional posteriors easier (see Koop and Korobilis, 2009a). Extensions to other prior distributions are straightforward and not affected by variable selection.

Exact expressions for the conditional densities of the parameters are provided in the appendix. Here I provide a pseudo-algorithm which demonstrates that the algorithm for the restricted model (4) actually adds only one block which samples $\gamma$, in the standard algorithm of the unrestricted VAR model (3).

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

[2] See Koop and Korobilis (2009a) for a review of priors and estimation approaches in Bayesian vector autoregressions.

**Bayesian Variable Selection Algorithm**

1. Sample $\beta$ as we would do in the unrestricted VAR in (3), but conditional on data being $Z_t^*$, with $Z_t^* = Z_t\Gamma$.

2. Sample each $\gamma_j$ conditional on $\gamma_{\backslash -j}$, $\beta$, $\Sigma$ and the data from

$$\gamma_j|\gamma_{\backslash -j}, \beta, \Sigma, y, z \sim Bernoulli\left(1, \pi_{0j}\right)$$

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

$$l_{0j} = p\left(y|\theta_j, \Sigma, \gamma_{\backslash -j}, \gamma_j = 1\right)\pi_{0j} \quad (8)$$

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

3. Sample $\Sigma$ as in the unrestricted VAR in (3), where now the mean equation parameters are $\theta = \Gamma\beta$.

In this type of model selection, what we care about is which of the parameters $\theta_j$ are equal to zero, so that identifiability of $\beta_j$ and $\gamma_j$ 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 $\beta_j = 0$ then $\gamma_j$ is identified by drawing from its prior: notice that in this case in equations (8) - (9) it holds that $p\left(y|\theta_j, \gamma_{\backslash -j}, \gamma_j = 1\right) = p\left(y|\theta_j, \gamma_{\backslash -j}, \gamma_j = 0\right)$, so that the posterior probability of the Bernoulli density, $\widetilde{\pi}_j$, will be equal to the prior probability $\pi_{0j}$. Similarly, when $\gamma_j = 0$ then $\beta_j$ is identified from the prior: the $j$-th column of $z_t^* = z_t\Gamma$ will be zero, i.e. the likelihood provides no information about $\beta_j$, and drawing from the posterior of $\beta_j$ collapses to getting a draw from its prior, i.e. $\widetilde{b}_j = b_{0j}$ and $\widetilde{V}_{jj} = V_{0,jj}$[3]. Nevertheless, in both of the above cases the result of interest is that $\theta_j = 0$, whether because $\beta_j = 0$ or because $\gamma_j = 0$, and the respective parameter is restricted.

There are several other approaches to automatic Bayesian model selection for regression models which can be generalized to VAR models. Most of them are based on introducing and sampling indicator variables $\gamma$ as we saw above. For the specific case of the simple VAR the restriction search proposed in Algorithm 1 is computationally more intensive than other approaches, like the variable selection algorithm of George et al. (2008); see also Appendix D. Nevertheless, Algorithm 1 can easily be adopted in the case of nonlinearity in the parameters, or specifications which admit non-conjugate priors. In that respect, the remainder of this paper develops a usefull extension, namely model selection in time-varying parameters

[3]This holds when $V_0$ is diagonal, which is usually the case in practice. If $V_0$ is a full $n \times n$ matrix, then the prior variance of $\beta_j$ is obviously detemined by the $j$-th row and $j$-th column of $V_0$. However the basic result stays unaffected, i.e. when $\gamma_j = 0$ taking a draw from the posterior of $\beta_j$ collapses to taking a draw from its prior.

VARs. Since a different value of the parameters at each time period $t$ occurs in these models, they tend to be non-parsimonious representations of macroeconomic data. Additionally, marginal likelihood calculations may be difficult to obtain (at least in terms of computer time). In this case, variable selection offers a very easy and fast method for selection of lag length and/or exogenous predictors.

## 2.2 Time-varying parameters VAR model

Modern macroeconomic applications increasingly involve the use of VARs with mean regression coefficients and covariance matrices which are time-varying. 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 et al. (1984) and Sims (1989); for a large-scale application in an 11-variable VAR see Canova (1993). However, the developement of accurate Bayesian sampling methods through the '90s (Gibbs sampler), combined with modern computing power has resulted in a recent rising interest in forecasting structural instability using time-varying parameters models. Canova and Ciccarelli (2004), Clark and McCracken (2010) and D'Agostino et al. (2009) are examples of forecasting multiple time-series using TVP-VAR's, while Stock and Watson (2007), Groen et al. (2009) and Koop and Korobilis (2009b) are focusing on univariate predictions but with the use of a large set of exogenous variables.

A time-varying parameters VAR with constant covariance (Homoskedastic TVP-VAR) takes the form

$$y_{t+1} = z_t \beta_t + \varepsilon_{t+1} \quad (10)$$

$$\beta_t = \beta_t + \eta_t \quad (11)$$

where $z_t = I_m \otimes x_t'$ is an $m \times n$ matrix, $\beta_t$ is an $n \times 1$ vector of parameters for $t = 1, ..., T$, $\varepsilon_t \sim N(0, \Sigma)$ with $\Sigma$ an $m \times m$ covariance matrix, and $\eta_t \sim N(0, Q)$ with $Q$ an $n \times n$ covariance matrix. Models of the form described above, and variations of it, have extensively been used for structural analysis (cite papers) and forecasting (cite papers). Obviously, the model in (10) is not a parsimonious representation of our data, and in practice most of the studies mentioned rely on quarterly data while using 2 lags of the dependent variable in order to prevent the proliferation of parameters and the estimation error.

Variable selection in this model is a simple extension of the VAR model with

constant parameters[4]. For that reason replace (10) with

$$y_{t+1} = z_t \theta_t + \varepsilon_{t+1} \tag{12}$$

where, as before, $\theta_t = \Gamma \beta_t$ and $\Gamma$ is the $n \times n$ matrix defined in (4). For this model, the priors on $\Sigma$ and $\gamma_j$ are the same as in the VAR case, i.e. $\Sigma^{-1} \sim Wishart\left(\alpha, S^{-1}\right)$ and $\gamma_j | \gamma_{\setminus -j} \sim Bernoulli\left(1, \pi_{0j}\right)$ respectively. For the time varying parameters, a prior on the initial condition is necessary which is of the form $\beta_0 \sim N_n\left(b_0, V_0\right)$. The random walk evolution of $\beta_t$, it would be desirable to restrict their prior variance in order to avoid explosive behavior. The (implied) priors for $\beta_1$ to $\beta_T$ are provided by the state equation (11), and they are of the form $\beta_t | \beta_{t-1}, Q \sim N\left(\beta_{t-1}, Q\right)$. The covariance matrix $Q$ is considered to be unknown, so it will have its own prior of the form $Q^{-1} \sim Wishart\left(\xi, R^{-1}\right)$. A Gibbs sampler for the unrestricted TVP-VAR model exists, so that sampling from the TVP-VAR with model selection, requires only one extra block which samples $\gamma$, in the spirit of Algorithm 1 of the previous section. Full details are provided in the appendix.

Due to the random walk assumption on the evolution of $\beta_t$, it is imperative need to restrict its covariance $Q$ otherwise draws of $\beta_t$ will enter the explosive region which might affect forecasting negatively. Primiceri (2005, Section 4.4.1.), who gives a detailed description of this issue, proposes prior hyperparameters for $Q$ based on the OLS quantities obtained from a constant-parameters VAR on a training sample. D'Agostino et al. (2009) adopt this idea in forecasting with TVP-VAR models, and following Cogley ad Sargent (2005) and Cogley et al. (2005) they also request that only stationary draws of $\beta_t$ are accepted. Stationarity restrictions in TVP-VAR models are satisfied if the roots of the reverse characteristic VAR polynomial defined by $\beta_t$ lie outside the complex unit circle for *each and every* $t = 1, ..., T$. This restriction can hardly be satisfied in VARs with more than 3 variables and 2 lags (see also Koop and Potter, 2008). Additionally, in many cases a training sample might not be available due to shortage of observations.

In that respect, in this forecasting exercise I use a Minnesota-based prior to elicit the prior hyperparameters of the TVP-VAR model. This results to an Empirical Bayes prior that can be tuned using the full sample, without the need to waste useful observations in a training sample. In order to avoid explosive draws, I subjectively choose the hyperparameters for the initial condition $\beta_0$ and the covariance matrix $Q$, in order to get a very tight prior. This prior allows to overcome stationarity restrictions which make the MCMC sampler inefficient. The reader should note that I use only a single choice of hyperparameters for the Minnesota prior, without searching and comparing other choices for the TVP-VAR model. The main purpose

[4]Note that variable selection is parsimonious and implies that a coefficient $\beta_{jt}$ will either be selected or discarted from the "true" model at all time periods $1, .., T$. For different approaches which allow different coefficients to enter or exit the "true" model at different points in time, see Koop and Korobilis (2009b) and Chan, Koop and Strachan (2010).

of the paper is to compare the unrestricted to the restricted (with variable selection) model, so only one benchmark prior is used for the shake of this comparison. Nevertheless, examining the forecasting performance of different priors in models with time-varying parameters is a challenging, but very important idea for future research.

### 2.3 Prior elicitation for variable selection

The performance of the variable selection is affected by the hyperparameters which affect the mean and variance of the mean equation coefficients $\beta$ or $\beta_t$. In the case of the VAR, we already discussed that when $\gamma_j = 0$ and a parameter $\beta_j$ is restricted we just take a draw from each prior. That means that the prior variance $V_0$ cannot be very large (to go to $\infty$) because this would imply that no predictors are selected. Kuo and Mallick (1997) propose to set $b_0 = (0, ..., 0)'$ and $V_0 = d \times I_n$, where $I_n$ is the identity matrix of dimensions $n \times n$. Then reasonable values for $d$ would be in the range $[0.25, 25]$. At first, this may seem like a restrictive assumption, but for VARs where the variables are approximately stationary, a prior variance on the regression coefficients of the form $V_0 = 10 \times I_n$ is fairly uninformative. In TVP-VARs, as explained previously, it is common practice to use an Emprical Bayes prior on the variance $Q$. These priors are by definition informative. For the purpose of comparison, I use a benchmark Minnesota-type prior.

Finally, it should be noted that variable selection is also affected by the hyperparameter of the Bernoulli prior of $\gamma_j$. The hyperparameters $\pi_{0j}$ can be tuned according to the researcher's beliefs about the number of expected restrictions in a model. As a rule of thumb, if the researcher expects or wants to impose as many restrictions as possible (for example, due to a degrees of freedom problem) then she can set $0 < \pi_{0j} \ll 0.5$[5]. Less restrictions are implied by setting $\pi_{0j} > 0.5$. The choice $\pi_{0j} = 0.5$ is used in practice as the uninformative choice, although it implies a priori that exactly 50% of the predictors should be included; see Chipman et al. (2001) for more details.

## 3 Simulated numerical examples

In order to assess the performance of the model selection algorithm, this section presents the results of two examples using simulated datasets.

**Example 1: Constant parameters VAR.** The first exercise is the one considered in George, Sun and Ni (2008). Consider a 6-variable VAR with a constant and one

[5] An insightful application of Bayesian variable selection by imposing many restrictions a priori, can be found in Brown, Vanucci and Fearn (2002). In this paper, the authors forecast with regression models using more predictors than observations.

lag and constant parameters[6]

$$B = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \Psi = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{13}$$

where it holds that $\Sigma = (\Psi\Psi')^{-1}$. 100 samples of size $T = 50$ are generated, using a random initial condition for the dependent variables $y_{-1} \sim U(0, 1)$. Hence 100 VAR models are estimated using the generated samples, saving 30.000 draws from the posterior of the estimated parameters after discarding an initial 20.000 draws to ensure convergence. The prior hyperparameters (see equations (5) - (7)) are set to be uninformative: $b_0 = 0_{n\times 1}$, $V_0 = 9 \times I_n$, $\pi_{0j} = 0.5$ for all $j = 1, ..., n$, $\alpha = 0$ and $S = 0 \times I_m$. The intercepts are always included in each model, so that the variable selection applies only to lags of the dependent variables.

The unrestricted VAR is just a special case of variable selection, where we impose all $\gamma_j = 1$. Thus, a fully unrestricted estimate of $B$ can be obtained using the variable selection priors and imposing always $\Gamma$ to be the identity matrix $I$ of dimensions $n \times n$. The average over the 100 samples of the posterior mean (posterior standard deviations of non-zero parameters are in parentheses) of $B$ using the unrestricted prior is

$$\widehat{B}_{UN} = \begin{pmatrix} 1.05 \\ (.51) & 1.06 \\ (.56) & 1.07 \\ (.57) & 0.91 \\ (.56) & 0.79 \\ (.56) & 0.96 \\ (.56) \\ 0.81 \\ (.08) & .12 & .10 & .11 & .10 & .09 \\ .05 & 0.71 \\ (.12) & .05 & .01 & .03 & .05 \\ .04 & .02 & 0.74 \\ (.11) & .01 & .01 & .06 \\ .00 & .06 & .06 & 0.75 \\ (.12) & .05 & .02 \\ .06 & .01 & .00 & .05 & 0.69 \\ (.12) & .03 \\ .00 & .06 & .08 & .03 & .07 & 0.72 \\ (.12) \end{pmatrix}$$

[6]The SUR transformation requires to estimate $\beta$ and $\gamma$ which are the parameters in vectorized form. For clarity in presentation, the parameters in these simulation study are given in their usual VAR matrix form ($B$ and $\boldsymbol{\gamma}$, where it holds that $\beta = vec(B)$ and $\boldsymbol{\gamma} = vec(\gamma)$).

The posterior means of the unrestricted model are identical to the MLE estimate. Now define $\boldsymbol{\gamma}$ to be the $k \times m$ matrix obtained from the column vector $\gamma$, which also includes the unrestricted constants. The average, over the 100 samples, of the posterior mean of the variable selection indices (in matrix form) $\widehat{\boldsymbol{\gamma}}$ are:

$$\widehat{\boldsymbol{\gamma}} = \begin{pmatrix} 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 \\ 1.00 & .09 & .09 & .05 & .06 & .07 \\ .05 & 1.00 & .05 & .09 & .05 & .04 \\ .06 & .09 & 1.00 & .14 & .10 & .08 \\ .07 & .11 & .06 & 0.99 & .15 & .05 \\ .05 & .13 & .05 & .11 & 1.00 & .10 \\ .03 & .06 & .07 & .08 & .06 & 1.00 \end{pmatrix}$$

and the average of the variable selection posterior mean, $\widehat{B}_{VS}$ are

$$\widehat{B}_{VS} = \begin{pmatrix} \begin{matrix} 0.99 \\ (.32) \end{matrix} & \begin{matrix} 1.10 \\ (.40) \end{matrix} & \begin{matrix} 1.05 \\ (.43) \end{matrix} & \begin{matrix} 1.02 \\ (.42) \end{matrix} & \begin{matrix} 0.98 \\ (.42) \end{matrix} & \begin{matrix} 1.00 \\ (.41) \end{matrix} \\ \begin{matrix} 0.98 \\ (.03) \end{matrix} & .01 & .00 & .00 & .01 & .00 \\ .00 & \begin{matrix} 0.93 \\ (.06) \end{matrix} & .04 & .05 & .04 & .01 \\ .01 & .01 & \begin{matrix} 0.94 \\ (.06) \end{matrix} & .03 & .01 & .02 \\ .01 & .02 & .03 & \begin{matrix} 0.92 \\ (.07) \end{matrix} & .04 & .02 \\ .01 & .02 & .01 & .01 & \begin{matrix} 0.87 \\ (.10) \end{matrix} & .01 \\ .00 & .02 & .03 & .01 & .03 & \begin{matrix} 0.96 \\ (.04) \end{matrix} \end{pmatrix}.$$

The elements of the $\widehat{\boldsymbol{\gamma}}$ matrix can be interpreted as "probabilities of inclusion" of a certain parameter. The top row of $\widehat{\boldsymbol{\gamma}}$ is $1$ by default, since it refers to the VAR intercept which was left unrestricted. Variable selection picks the correct restrictions in small samples, resulting in more accurate estimates of the VAR regression coefficients (compare $\widehat{B}_{UN}$ and $\widehat{B}_{VS}$). This is true also for the unrestricted intercepts, which are much closer to their true values. Additionally, the covariance matrix resulting from the variable selection is also more accurate than the MLE of the covariance matrix (results not reported here).

George et al. (2008) have used the exact same setup to evaluate the efficiency of the SSVS algorithm described in Section 3.1. Comparing the matrix of restrictions, $\widehat{\boldsymbol{\gamma}}$, reported above, with their equivalent matrix it is obvious that their reported probabilities of inclusion of the parameters are more decisive. In their case, the highest probability which a paramter which should be restricted gets is equal to

$0.5$ in only a few cases. This happens because the SSVS restricts a parameter if it is too low[7], while variable selection in this paper requires a parameter to be restricted exactly to zero. However, using both algorithms model selection implies that the optimal predictive model is the one which has parameters with probability of inclusion higher than 0.5 (see Barbieri and Berger (2004) for a proof).

**Example 2: Homoskedastic TVP-VAR.** In the second example, 100 samples of size $T = 100$ are generated from a 4-variable Homoskedastic TVP-VAR with one autoregressive lag (no constant). The covariance matrix $\Sigma$ is set equal to the upper left 4x4 block of the covariance matrix specified in Example 1 (in equation (13)), and $\beta_t = vec(B_t)$ is let to evolve according to the random walk specification (11), by setting initial condition, $B_0 = \{B_t\}_{t=0}$, and a simple diagonal covariance matrix $Q$, of the form

$$B_0 = \begin{pmatrix} .7 & 0 & .35 & 0 \\ 0 & .7 & 0 & 0 \\ 0 & .45 & .7 & 0 \\ .4 & 0 & 0 & .7 \end{pmatrix}, Q_{j,j} = \begin{cases} 0 & \text{, if } B_{0,ij} = 0 \\ 0.01 & \text{, if } B_{0,ij} > 0.5 \\ 0 & \text{, if } B_{0,ij} < 0.5 \end{cases}.$$

This specification implies that the diagonal elements of $B_t$ are time-varying with initial condition $0.7$, while the non-zero non-diagonal elements $0.4, 0.45, 0.35$ (which are lower than $0.5$) remain constant for all $t$ (and, of course, the zero non-diagonal elements remain zero for all $t$). The goal here is to examine the efficiency of variable selection when in the true model the R.H.S. variables affect the dependent variable through a combination of constant and time-varying coefficients, but the (misspesified) model we are estimating assumes all coefficients to be time varying.

The usual practice in the TVP-VAR models is to use tight databased priors. However, for the purpose of this excercise realatively uninformative priors are defined

$$\begin{aligned} \beta_0 &\sim N_n(b_0, V_0) \\ Q^{-1} &\sim Wishart\left(\xi, R^{-1}\right) \\ \gamma_j|\gamma_{\setminus -j} &\sim Bernoulli(1, \pi_{0j}) \\ \Sigma^{-1} &\sim Wishart\left(\alpha, S^{-1}\right) \end{aligned}$$

where the hyperparameters are set to the values $b_0 = 0_n$, $V_0 = 10I_n$, $\xi = 16$, $R = 1$, $\pi_{0j} = 0.5$ for all $j = 1, ..., n$, $\alpha = 4$ and $S = 1$ . Full details, like means and variances of the posteriors of the parameters, are difficult to present here. However, the average of the restriction indices is again very informative about the efficiency of the the variable selection algorithm to find the correct restrictions:

[7]In particular, they set their prior hyperparameters in such a way, that parameters which are lower than $0.5$ should be restricted and shrunk towards zero (but as explained, never equal to zero).

$$\widehat{\gamma} = \begin{pmatrix} 1.00 & .04 & 1.00 & .04 \\ .01 & 1.00 & .04 & .01 \\ .02 & 1.00 & 1.00 & .05 \\ 1.00 & .04 & .02 & 1.00 \end{pmatrix}.$$

As in the simple VAR case above, posterior means are more accurate and posterior standard deviations are smaller (results available upon request).

# 4 Macroeconomic Forecasting with VARs

## 4.1 Data and set-up

The variable selection techniques described in Section 2 are used to provide forecasts of four macroeconomic series of the U.K. economy. In particular, the variables included in these models are the inflation rate $\Delta\pi_t$ (RPI:Percentage change over 12 months: All items), unemployment rate $u_t$ (Unemployment rate: All aged 16 and over, Seasonally adjusted), the annual growth rate of GDP $gdp_t$ (Gross Domestic Product: Quarter on quarter previous year: Chain volume measure, Seasonally adjusted) and the interest rate $r_t$ (Treasury bills: average discount rate). It is customary in VAR models to include only one measure of economic activity, i.e. either GDP or unemployment. The assumption here is that policymakers are interested individually, at least in the short-run, in forecasts of both GDP and unemployment. There are many reasons for having individual forecasts for unemployment and GDP growth, for example as of January 2010 it is the case that many economies are out of the global recession according to initial GDP growth rate estimates. However in the same countries (including US) unemployment is not getting lower, and price inflation is below its target level.

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. Inflation, unemployment and interest rates are measured on a monthly basis. 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 interest rate), respectively. The data are plotted in Figure 1.

The VAR and Homoskedastic TVP-VAR models include a constant and a maximum of 2 lags of the dependent variables. One could argue that the use of variable selection would allow to define a higher maximum lag length, say 4 lags, and then let the data decide on the optimal number of lags in each VAR equation. However, given the fact that the total observations are only 152, we would be asking too much from the variable selection algorithm in a recursive forecasting exercise. The forecast horizons used for comparison are $h = 1$,$4$ and $8$. The sample 1971:Q1 -

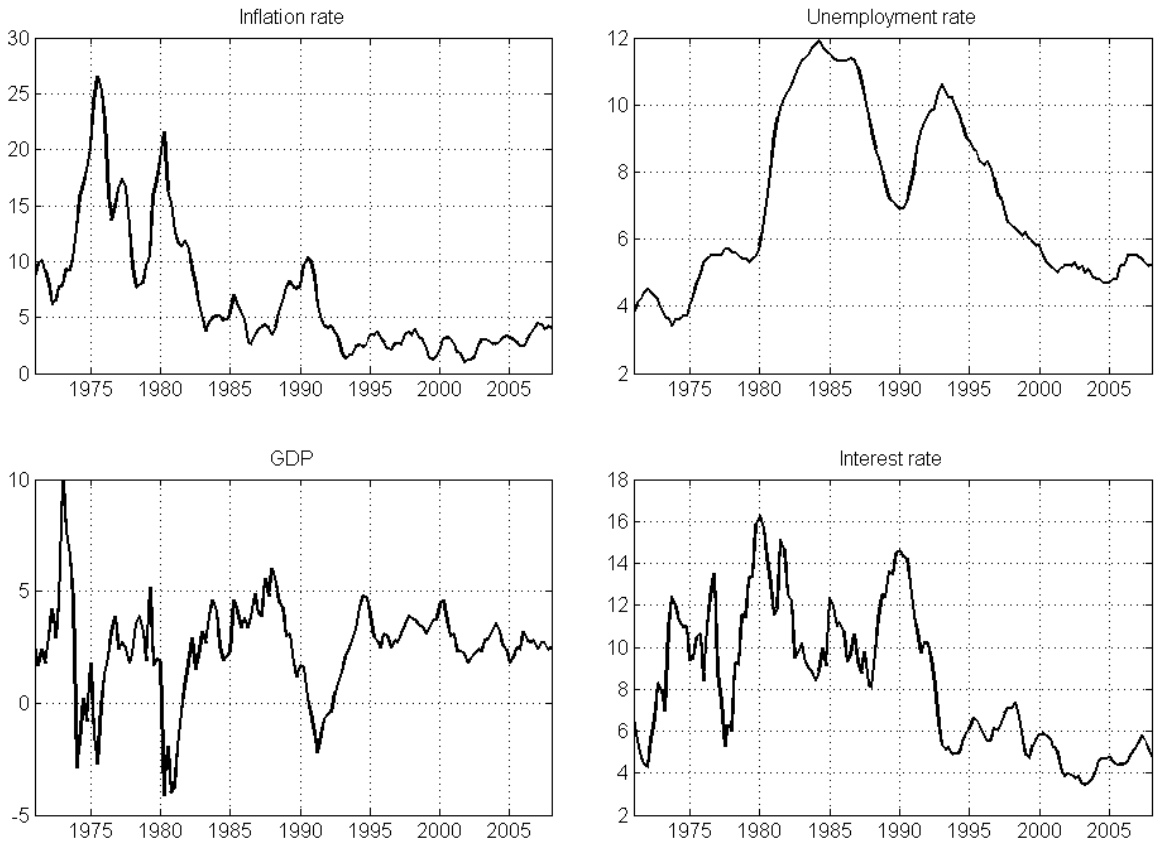


Figure 1: Graph of the data for UK inflation, unemployment, GDP, and interest rate.

1988:Q4 is used for initial estimation, and the forecasts are computed recursively with expansion of the estimation sample each quarter. Subsequently for the period 1989:Q1 - 2008:Q4 we obtain a total of 80-$h$ forecasts.

Iterated forecasts are obtained by estimating the models (4) and (10), writing the models in companion (VAR(1)) form, and iterating the forward up to $h = 8$ periods ahead in order to obtain $[\widehat{y}_{T+1}, ..., \widehat{y}_{T+h}]$, where $T$ is the last observation of the sample. Direct forecasts are obtained from the VAR and the TVP-VAR specifications in (4) and (10) respectively by estimating separately for $h = 1$,4 and 8 the models with the dependent variable $y_{t+1}$ replaced by $y_{t+h}$, while the R.H.S. variables are still measured up to, and including, time $t$. When $h = 1$ direct and iterated forecasts are exactly the same, since we are estimating and forecasting with exactly the same specifications.

When forecasting, the parameters of the VAR model remain constant in the out-of-sample period. However this is not the case for the autoregressive coefficients of the TVP-VAR model. The iterative nature of the Gibbs sampler allows to easily

simulate the out-of-sample path of $\beta_t$. At each iteration, conditional on obtaining a draw of $\beta_T$ and the covariances $Q$, we can use the random walk evolution equation (11) to simulate $\left[\widehat{\beta}_{T+1}, ..., \widehat{\beta}_{T+h}\right]$ in a recursive manner. Then conditional on knowing these parameters, calculation of direct or iterated forecasts can be computed as in the VAR case, separately for each out-of-sample period.

## 4.2 Forecasting models

Dependent on the model specification and the choice of prior hyperparameters, there are 5 competing forecasting models. Common place in all models, in order to evaluate the performance of variable selection of mean equation coefficients, is that the covariance matrix is integrated out using an uninformative prior of the form $p(\Sigma) \propto |\Sigma|^{-(m+1)/2}$ which is equivalent to the Whishart prior defined in (7) with the additional restriction that $\alpha = 0$ and $S^{-1} = 0_{m \times m}$.

The 5 models are

1. VAR with variable selection (VAR VS)

   The priors are $\gamma_j | \gamma_{\setminus -j} \sim Bernoulli(1, 0.5)$ for all $j = 1, ..., n$, and $\beta_j \sim N(0, 10^2)$ if $\beta_j$ is an intercept, and $\beta_j \sim N(0, 3^2)$ otherwise.

2. VAR with Minnesota prior (VAR MIN)

   The Minnesota prior for $\beta$ is of the form. $\beta \sim N\left(b^{MIN}, V^{MIN}\right)$ where

$$V_{i,l}^{MIN} = \begin{cases} g_1/p & \text{for parameters on own lags} \\ g_3/s_i^2 & \text{for intercepts} \\ \frac{g_2 s_i^2}{p s_l^2} & \text{for parameters } j \text{ on variable } l \neq i, l, i = 1, .., m \end{cases} \tag{14}$$

   Here $s_i^2$ is the residual variance from the $p$-lag univariate autoregression for variable $i$. The prior mean vector $b_{MIN}$ is set equal to $0.9$ for parameters on the first own lag of each variable and zero otherwise. The hyperparameters are set to the values $g_1 = 0.5$, $g_2 = 0.005$ and $g_3 = 100$.

3. Benchmark VAR

   The priors are the same as the VAR with variable selection (VAR VS), where now we do not sample $\gamma_j$ (or equivalently, restrict $\gamma_j = 1$ for all $j$)

4. TVP-VAR with variable selection (TVP-VAR VS)

   The initial condition is set to $\beta_0 \sim N(0, 4^2 V_{MIN})$, and $\gamma_j | \gamma_{\setminus -j} \sim Bernoulli(1, 0.5)$. The covariance $Q$ of the varying coefficients has the prior $Q^{-1} \sim Wishart(\xi, R)$ where $\xi = n + 1$ and $R^{-1} = 0.001(n+1) V^{MIN}$, where $V^{MIN}$ is the matrix defined in (14) above.

5. Benchmark TVP-VAR

   The priors are the same as in the TVP-VAR case with variable selection, but in this case $\gamma_j = 1$ for all $j = 1, ..., n$.

The priors for the VAR are fairly uninformative, however the TVP-VAR prior is quite tight. Althernatively we can assign to $\beta_0$ a large variance, say $100I$, on the basis that this is a desirable uninformative choice. However doing so, means that we increase the probability that the whole sequence of draws for $\beta_t$ will be in the nonstationary region. This approach can be computationally cumbersome as many draws may be required as in the case of Cogley et al. (2005) who use 100.000 draws, discard the first 50.000 and save every 10-th draw. Given the dimension of the parameter space and the compuational demands of a recursive forecasting exercise, the informative variance $4^2 V_{MIN}$ on the initial state is used here in order enhance the efficiency of the Gibbs sampler. All models are based on a run of 20.000 draws from the posterior, discarding the first 10.000 draws.

The choice of the hyperparameter $R^{-1}$ is based on the variance of the Minnesota prior as well, with a scaling constant equal to $0.001\,(n+1)$. This might not be the optimally elicited hyperparameter of this prior for forecasting purposes, and other choices exist which the researcher ought to examine. However the purpose of this paper is, for a given prior, to compare the unrestricted model with the same model with variable selection added. Subsequently, while a specific prior can be a subject of critisism if the ulitmate purpose was to compare the performance of the TVP-VAR with that of other models (like random walk, and nonlinear models like a Markov Switching or Structural Breaks VAR), this critisism should not apply here.

## 4.3 Forecast evaluation

All models are evaluated using the Mean Squared Forecast Error (MSFE) and the Mean Absolute Forecast Error (MAFE). In particular, for each of the 4 variables $y_i$ of $y$ and conditional on the forecast horizon $h$ and the time period $t$, the two measures are computed as

$$\begin{aligned} MSFE^h_{i,t} &= \sqrt{\left(\widehat{y}_{i,t+h|t} - y^o_{i,t+h}\right)^2} \\ MAFE^h_{i,t} &= \left|\widehat{y}_{i,t+h|t} - y^o_{i,t+h}\right| \end{aligned}$$

where $\widehat{y}_{i,t+h|t}$ is the time $t+h$ prediction of variable $i$ (inflation, unemployment, gdp or interest rate), made using data available up to time $t$. $y^o_{i,t+h}$ is the observed value (realisation) of variable $i$ at time $t+h$. In the recursive forecasting exercise, averages over the full forecasting period 1989:Q1 - 2008:Q4 are presented using

the formulas

$$\left(\widehat{MSFE}\right)_i^h = \frac{1}{\tau_1 - h - \tau_0} \sum_{t=\tau_0}^{\tau_1 - h} MSFE_{i,t}^h$$

$$\left(\widehat{MAFE}\right)_i^h = \frac{1}{\tau_1 - h - \tau_0} \sum_{t=\tau_0}^{\tau_1 - h} MAFE_{i,t}^h$$

where $\tau_0$ is 1989:Q1 and $\tau_1$ is 2008:Q4.

## 4.4 In-sample variable selection results

Tables 1 and 2 present variable selection results for the VAR-VS and TVP-VAR-VS models using the full sample. Entries in these tables are the means of the posterior draws of the indices $\gamma$ for the two models. Draws from the posterior of $\gamma$ is just a sequence of 1's and 0's, so that the mean can be simply interpreted as a probability of inclusion of each variable. Note that while $\gamma$ is a column vector, results are presented in the table in matrix form, where the dependent variables are in columns and the R.H.S variables are in rows. For $h = 1$ the model for direct forecasts has the same specification as the model for iterated forecasts, so columns 1-4 in the tables refer to both models. However notice that for longer horizons we need to specify a different model for direct forecasts and columns 5-12 in the tables refer only to the this model specification, for horizons $h = 4$ and $8$.

The first thing to observe from Tables 1 and 2 is that for both models and for all horizons variable selection imposes many restrictions. This result is not surprising, both from an empirical and thoeretical point of view. The most recent own lag of each variable is important in most cases for all forecast horizons. Other than that, variable selection indicates only a few extra variables as important in each VAR equation, leading to quite parsimonious models. This pattern complies with the empirical results of Korobilis (2008) and Jochmann et al. (2009) using the SSVS algorithm for VAR models (see Section 3.1 above).

It is obvious that when the posterior mean of $\gamma$ is exactly equal to 0 or 1, then a specific predictor variable should just respectively be exit or enter the best model. An interesting question is how to decide and classify a predictor when the associated probability is 0.6 or 0.3 for example. In fact, Barbieri and Berger (2004) show that the optimal model in model/variable selection for prediction purposes is the median probability model. Subsequently their proposed rule is only to select variables which have probability of inclusion in the best model higher than 0.5.

A comparison of the parameters of the VAR models with the respective parameters of the TVP-VAR models, reveals quite a few differences, but also many similarities at the same time, as to which variables are selected to enter the "best" model. For example, in the VAR strong (probability equal to $1$) predictors for 1-quarter

ahead inflation ($\Delta\pi_{t+1}$) are current inflation ($\Delta\pi_t$) and interest rate($r_t$), as well as inflation in the previous quarter ($\Delta\pi_{t-1}$). In the TVP-VAR it is only $\Delta\pi_t$ and $\Delta\pi_{t-1}$ which are selected, and the current level of interest rate has only probability of 0.28. In the same equation, there is weaker evidence that $gdp_t$, $u_{t-1}$and $r_{t-1}$ are good predictors, which vanishes in the TVP-VAR case (for example $r_{t-1}$ has a probability of 0.61 of entering the VAR model, but only a probability of 0.41 of entering the TVP-VAR model). Similar inference can be made for the rest of variables and equations.

An interesting question is whether any differences in the inclusion probabilities of the predictors in the VAR and the same predictors in the TVP-VAR, are due to the fact that the models are different or because of the different priors. This is a difficult question to answer, since this would require to place exactly the same priors (for instance a flat prior on all parameters) in both specification and do the comparison. As explained in this paper, flat priors on all the hyperparameters of the TVP-VAR model are not possible.

Table 1. Average of posterior of restrictions $\gamma$, for $h = 1, 4, 8$ (VAR model)

| | $\Delta\pi_{t+1}$ | $u_{t+1}$ | $gdp_{t+1}$ | $r_{t+1}$ | $\Delta\pi_{t+4}$ | $u_{t+4}$ | $gdp_{t+4}$ | $r_{t+4}$ | $\Delta\pi_{t+8}$ | $u_{t+8}$ | $gdp_{t+8}$ | $r_{t+8}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Intercept | 0.14 | 0.02 | **0.91** | 0.18 | **0.96** | 0.06 | **1.00** | **0.59** | **0.83** | 0.10 | **1.00** | 0.25 |
| $\Delta\pi_t$ | **1.00** | 0.00 | 0.06 | 0.02 | **1.00** | 0.01 | 0.22 | 0.10 | **1.00** | 0.04 | 0.10 | 0.20 |
| $u_t$ | 0.29 | **1.00** | **0.61** | 0.00 | **0.97** | **1.00** | 0.37 | 0.07 | **0.57** | **1.00** | **0.77** | **0.81** |
| $gdp_t$ | **0.61** | 0.05 | **1.00** | 0.06 | 0.47 | 0.52 | 0.03 | 0.47 | 0.14 | 0.04 | 0.06 | 0.08 |
| $r_t$ | **1.00** | 0 | 0.19 | **1.00** | **1.00** | 0.02 | 0.03 | **1.00** | 0.16 | **1.00** | **1.00** | 0.46 |
| $\Delta\pi_{t-1}$ | **1.00** | 0.00 | 0.02 | 0.03 | 0.07 | 0.00 | 0.69 | 0.55 | 0.06 | 0.34 | 0.03 | **0.84** |
| $u_{t-1}$ | **0.53** | **1.00** | **0.62** | 0.01 | 0.47 | **1.00** | 0.80 | 0.09 | **0.60** | **1.00** | **0.63** | **1.00** |
| $gdp_{t-1}$ | 0.38 | 0.00 | 0.04 | 0.47 | 0.07 | 0.01 | 0.03 | 0.28 | **1.00** | 0.03 | **0.72** | **0.97** |
| $r_{t-1}$ | **0.64** | **1.00** | **0.72** | 0.04 | 0.30 | **1.00** | **1.00** | 0.06 | 0.22 | 0.01 | 0.13 | 0.43 |



Table 2. Average of posterior of restrictions $\gamma$, for $h = 1, 4, 8$ (TVP-VAR model)

| | $\Delta\pi_{t+1}$ | $u_{t+1}$ | $gdp_{t+1}$ | $r_{t+1}$ | $\Delta\pi_{t+4}$ | $u_{t+4}$ | $gdp_{t+4}$ | $r_{t+4}$ | $\Delta\pi_{t+8}$ | $u_{t+8}$ | $gdp_{t+8}$ | $r_{t+8}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Intercept | 0.25 | **0.51** | **0.94** | **0.82** | **0.86** | 0.32 | **1.00** | **1.00** | **1.00** | 0.42 | **0.91** | **1.00** |
| $\Delta\pi_t$ | **1.00** | 0.49 | 0.46 | **1.00** | 0.05 | 0.46 | 0.40 | **0.77** | **1.00** | 0.47 | **0.54** | **0.73** |
| $u_t$ | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 | 0.02 | **1.00** | 0.24 | **1.00** |
| $gdp_t$ | 0.31 | **0.52** | **1.00** | **0.95** | **0.99** | 0.58 | **1.00** | 0.44 | **0.75** | 0.41 | 0.00 | 0.28 |
| $r_t$ | 0.28 | **0.50** | 0.30 | **1.00** | 0.50 | 0.50 | 0.44 | **1.00** | 0.17 | **0.57** | 0.38 | 0.18 |
| $\Delta\pi_{t-1}$ | **1.00** | 0.49 | 0.41 | 0.38 | 0.00 | 0.49 | 0.45 | **0.65** | 0.00 | 0.49 | **0.65** | 0.47 |
| $u_{t-1}$ | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | **1.00** | 0.06 | **1.00** | 0.00 | **1.00** | 0.17 | **1.00** |
| $gdp_{t-1}$ | 0.21 | **0.54** | 0.02 | **0.73** | 0.20 | **0.71** | 0.12 | **0.99** | 0.24 | **0.82** | 0.00 | 0.36 |
| $r_{t-1}$ | 0.41 | 0.43 | **0.77** | 0 | **0.87** | **0.75** | 0.39 | 0.00 | **0.56** | 0.48 | 0.29 | 0.00 |

## 4.5 Out-of-sample forecasting results

In this subsection the restricted and unrestricted VAR models are evaluated out-of-sample. Tables 3 to 5 present the MSFE and MAFE statistics over the forecasting sample 1989:Q1-2008:Q4. Although the aim of this forecasting exercise is to assess the gains from using variable selection in VAR models, for ease of comparison the MSFE and MAFE of a naive forecasting model are presented. This model is the random walk estimated for each individual time series, over the three different forecast horizons. Note that in Table 3 there is one set of results for direct and iterated forecasts, since for $h = 1$ the model specifications are exactly the same.

The results indicate that on average we are much better off when using variable selection than when using the unrestricted models. For individual series it is the case that variable selection would either offer large improvements or it would give predictions similar to the unrestricted model. This would not be surprising as soon as the restrictions imposed are the correct ones. That is, it is expected that a correctly restricted model will perform at least as well as the unrestricted model. However an incorectly restricted model will most probably give predictions which are really worse than the unrestricted model (dependent on the importance of the relevant variables which are incorrectly restricted). Subsequently, the improvement in forecasting suggests that Bayesian variable selection picks correct restrictions, which lead to useful parsimonious models.

For short-term forecasts ($h = 1$) the multivariate VAR models, whether restricted or unrestricted, offer more accurate forecasts compared to the parsimonious naive forecasts. However this picture is reversed for more distant forecasts and the performance varies substantially for each variable, dependent on whether iterated or direct forecasts have been calculated. Previous results (see for example Pesaran, Pick and Timmermann 2009, and references therein) suggest that iterated forecasts *may* dominate direct forecasts in small samples and for large forecast horizons, while direct forecats *may* dominate when the dynamics of the model are misspecified. For 4- and 8-step ahead forecasts of inflation it is obvious that the direct model performs much better. It is well known that the dynamics of inflation are other than linear which implies why the VAR model performs poorly for this variable (always compared to the naive forecast). The nonlinear TVP-VAR model hugely improves over the VAR forecasts for 4- and 8-step ahead horizons, however the direct model specification is more accurate. The reader should note that in this paper the assumption is that the out-of-sample parameters have to be simulated (instead of, for instance, fixing their values at the value estimated at time $T$), which might explain an accumulated uncertainty in the parameters over longer horizons. Even though this paper argues that stationarity restrictions are computationally inefficient in TVP-VAR models for the estimation of the parameters $[\beta_1, ..., \beta_T]$, the applied researcher might want to combine a tight prior on the estimated parameter with stationarity restriction imposed in the out-of-sample simulated parameters $\left[\beta_{T+1}, ..., \beta_{T+h}\right]$.

While MSFE and MAFE measures are very informative in our case, since the purpose is just to evaluate point forecasts, full predictive densities can be compared using predictive likelihoods. In fact predictive likelihoods averaged on all 4 dependent variables suggest that the restricted models (whether VAR and TVP-VAR variable selection or the Minnesota VAR prior) by reducing uncertainty about the paramaeters, tend to also reduce the uncertainty regarding predictions. Finally, note that in order to have a complete picture of the performance of variable selection, we should additionally compare the restricted models with the respective unrestricted models with one lag. The restricted models have a *maximum* lag of two and it might be the case that the "true" data generating process is a model with one lag which variable selection is not able to capture. It turns out that unrestricted VAR and TVP-VAR models with only one lag consistently forecast worse than the unrestricted models with two lags, at all forecast horizons. For the shake of brevity results on predictive likelihoods, and VAR models with different lags are not presented here but are available upon request[8].

The reader can replicate the results in this paper using MATLAB code available in http://personal.strath.ac.uk/gcb07101/code.html.

Table 3. Forecast evaluation, $h = 1$

| | MSFE | | | | MAFE | | | |
|---|---|---|---|---|---|---|---|---|
| | $\Delta\pi_t$ | $u_t$ | $gdp_t$ | $r_t$ | $\Delta\pi_t$ | $u$ | $gdp_t$ | $r_t$ |
| *Naive Model:* | | | | | | | | |
| RW | 0.576 | 0.108 | 0.501 | 0.394 | 0.624 | 0.262 | 0.594 | 0.484 |
| *VAR Models:* | | | | | | | | |
| | Direct/Iterated forecasts | | | | | | | |
| VAR | 0.285 | 0.027 | 0.247 | 0.239 | 0.423 | 0.132 | 0.406 | 0.355 |
| VAR-MIN | 0.300 | 0.029 | 0.267 | 0.241 | 0.432 | 0.135 | 0.419 | 0.356 |
| VAR-VS | 0.208 | 0.030 | 0.164 | 0.152 | 0.354 | 0.134 | 0.324 | 0.291 |
| TVP-VAR | 0.475 | 0.033 | 0.302 | 0.185 | 0.595 | 0.153 | 0.437 | 0.346 |
| TVP-VAR-VS | 0.419 | 0.035 | 0.273 | 0.157 | 0.542 | 0.149 | 0.360 | 0.318 |

[8]It turns out that among the unrestricted VAR and TVP-VAR models with up to four lags, the specifications with two lags perform the best at all horizons and for both iterated and direct forecasts.

Table 4. Forecast evaluation, $h = 4$

| | MSFE | | | | MAFE | | | |
|---|---|---|---|---|---|---|---|---|
| | $\Delta\pi_t$ | $u_t$ | $gdp_t$ | $r_t$ | $\Delta\pi_t$ | $u$ | $gdp_t$ | $r_t$ |
| *Naive Model:* | | | | | | | | |
| RW | 1.190 | 1.223 | 1.439 | 1.213 | 0.874 | 0.932 | 0.902 | 0.930 |
| *VAR Models:* | | | | | | | | |
| | | | | Direct forecasts | | | | |
| VAR | 9.714 | 1.382 | 1.374 | 2.761 | 2.695 | 0.989 | 0.874 | 1.421 |
| VAR-MIN | 3.674 | 1.307 | 1.347 | 2.928 | 1.696 | 0.966 | 0.864 | 1.465 |
| VAR-VS | 5.110 | 1.289 | 0.818 | 1.563 | 1.958 | 0.925 | 0.751 | 1.082 |
| TVP-VAR | 2.068 | 1.058 | 1.259 | 0.775 | 1.188 | 0.911 | 0.917 | 0.779 |
| TVP-VAR-VS | 1.965 | 1.046 | 0.903 | 0.675 | 1.150 | 0.912 | 0.814 | 0.644 |
| | | | | Iterated forecasts | | | | |
| VAR | 8.150 | 0.231 | 1.228 | 2.456 | 2.376 | 0.422 | 0.882 | 1.209 |
| VAR-MIN | 7.948 | 0.230 | 1.215 | 2.577 | 2.318 | 0.422 | 0.876 | 1.255 |
| VAR-VS | 7.730 | 0.208 | 1.263 | 1.303 | 2.025 | 0.361 | 0.697 | 0.843 |
| TVP-VAR | 3.157 | 1.243 | 1.715 | 1.983 | 1.388 | 0.896 | 1.082 | 1.106 |
| TVP-VAR-VS | 3.680 | 1.083 | 1.552 | 1.617 | 1.326 | 0.717 | 1.026 | 0.973 |

Table 5. Forecast evaluation, $h = 8$

| | MSFE | | | | MAFE | | | |
|---|---|---|---|---|---|---|---|---|
| | $\Delta\pi_t$ | $u_t$ | $gdp_t$ | $r_t$ | $\Delta\pi_t$ | $u$ | $gdp_t$ | $r_t$ |
| *Naive Model:* | | | | | | | | |
| RW | 2.011 | 3.730 | 1.752 | 3.545 | 1.258 | 1.558 | 1.076 | 1.547 |
| *VAR Models:* | | | | | | | | |
| | | | | Direct forecasts | | | | |
| VAR | 10.535 | 7.238 | 2.984 | 9.435 | 3.121 | 2.2579 | 1.303 | 2.626 |
| VAR-MIN | 8.800 | 4.419 | 2.525 | 9.648 | 2.336 | 1.8006 | 1.211 | 2.684 |
| VAR-VS | 2.957 | 4.604 | 2.255 | 6.270 | 1.655 | 1.9645 | 1.112 | 2.177 |
| TVP-VAR | 1.533 | 2.831 | 2.913 | 3.928 | 1.072 | 1.2794 | 1.578 | 1.748 |
| TVP-VAR-VS | 1.251 | 1.679 | 2.907 | 4.063 | 0.918 | 1.0735 | 1.561 | 1.757 |
| | | | | Iterated forecasts | | | | |
| VAR | 30.870 | 0.790 | 0.849 | 7.903 | 5.069 | 0.750 | 0.687 | 2.471 |
| VAR-MIN | 29.863 | 0.771 | 0.820 | 7.262 | 4.959 | 0.743 | 0.675 | 2.335 |
| VAR-VS | 22.996 | 0.727 | 0.866 | 2.570 | 3.619 | 0.706 | 0.681 | 1.326 |
| TVP-VAR | 13.822 | 1.457 | 1.298 | 5.124 | 2.642 | 0.982 | 0.945 | 1.826 |
| TVP-VAR-VS | 4.126 | 1.043 | 1.554 | 2.380 | 1.604 | 0.849 | 1.004 | 1.259 |

# 5 Concluding remarks

Vector autoregressive models have been used extensively over the past for the purpose of macroeconomic forecasting, since they can fit the observed data better than competing theoretical and large-scale structural macroeconometric models. Nowadays, Bayesian dynamic stochastic general equilibrium (DSGE) models like the one of Smets and Wouters (2003) have been shown to challenge the forecasting performance of unrestricted VAR models, while at the same time having all the advantages of being structural, i.e. connected to economic theory. While DSGE models provide restrictions based on theory, this paper shows that Bayesian variable selection methods can be used to find restrictions based on the evidence in the data, and at the same time improve over the forecasts of unrestricted VAR models as well. Additionally, Bayesian variable selection methods for vector autoregressions can be used for structural analysis, like measureing monetary policy shocks in identified VARs. A different route for VAR variable selection algorithms would be to uncover empirically the relationship between variables, which could potentially help in the developement of new theoretical relationships.

# Technical Appendix

# A Posterior inference in the 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

$$\begin{aligned} \beta &\sim N_n\left(b_0, V_0\right) && \text{(A.1)} \\ \gamma_j|\gamma_{\setminus -j} &\sim Bernoulli\left(1, \pi_{0j}\right) && \text{(A.2)} \\ \Sigma^{-1} &\sim Wishart\left(\alpha, S^{-1}\right) && \text{(A.3)} \end{aligned}$$

## A.1 Algorithm 1

Given the prior hyperparameters $(b_0, V_0, \pi_0, \Psi, \alpha)$ and an initial value for $\gamma$, $\Sigma$, sampling from the conditional distributions proceeds as follows

1. Sample $\beta$ from the density

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

where $\widetilde{V} = \left(V_0^{-1} + \sum_{t=1}^{T} z_t^{*\prime}\Sigma^{-1}z_t^*\right)^{-1}$ and $\widetilde{b} = \widetilde{V}\left(V_0^{-1}b_0 + \sum_{t=1}^{T} z_t^{*\prime}\Sigma^{-1}y_{t+h}\right)$, and $z_t^* = z_t\Gamma$.

2. Sample $\gamma_j$ from the density

$$\gamma_j|\gamma_{\setminus -j}, \beta, \Sigma, y, z \sim Bernoulli\left(1, \widetilde{\pi}_j\right) \tag{A.5}$$

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

$$\begin{aligned} l_{0j} &= p\left(y|\theta_j, \gamma_{\setminus -j}, \gamma_j = 1\right)\pi_{0j} && \text{(A.6)} \\ l_{1j} &= p\left(y|\theta_j, \gamma_{\setminus -j}, \gamma_j = 0\right)\left(1 - \pi_{0j}\right) && \text{(A.7)} \end{aligned}$$

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 $\theta^*$ to be equal to $\theta$ but with its $j-th$ element $\theta_j = \beta_j$ (i.e. when $\gamma_j = 1$). Similarly, define $\theta^{**}$ to be equal to $\theta$ 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

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

3. Sample $\Sigma^{-1}$ from the density

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

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

## 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

$$vec\left(y\right) = \left(I_m\otimes x'\right)\Gamma b + vec\left(\varepsilon\right) \tag{A.9}$$

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

where $Y=vec\left(y\right)$ 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 $e=vec\left(\varepsilon\right)\sim N\left(0,\Sigma\otimes I_T\right)$. To clarify notation, $vec\left(\circ\right)$ 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 $z=(z_1',...,z_T')=\left(\left(I_m\otimes x_1\right)',...,\left(I_m\otimes x_T\right)'\right)$ which was defined in (4). Additionally, note that while $\beta$ and $b$ 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 $b$ from the density

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

where $\widetilde{V} = V_0^{-1} + W^{*\prime}\left(\Sigma^{-1} \otimes I_T\right) W^*$ and $\widetilde{b} = \widetilde{V}\left(V_0^{-1} b_0 + W^{*\prime}\left(\Sigma^{-1} \otimes I_T\right) Y\right)$, and $W^* = W\Gamma$.

2. Sample $\gamma_j$ from the density

$$\gamma_j|\gamma_{\setminus -j}, \beta^*, \Sigma, Y, W \sim Bernoulli\left(1, \widetilde{\pi}_j\right) \tag{A.12}$$

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

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

3. Sample $\Sigma^{-1}$ from the density

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

where $\widetilde{\alpha} = T + \alpha$ and $\widetilde{S}^{-1} = \left(S + (Y - x\Theta)'(Y - x\Theta)\right)^{-1}$, where $\Theta$ is the $k \times n$ matrix obtained from the vector $\theta = \Gamma\beta^*$, which has elements $(\Theta_{ij}) = \theta_{(j-1)k+i}$, for $i = 1, ..., k$ and $j = 1, ..., n$.

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. Unfortunately it is not possible to formulate a TVP-VAR in the form (A.9), in order to take advantage of matrix computations.

# B Posterior inference in the TVP-VAR with variable selection

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

$$y_t = z_t\theta_t + \varepsilon_t \quad \text{(B.1)}$$
$$\beta_t = \beta_{t-1} + \eta_t \quad \text{(B.2)}$$

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:

$$\begin{aligned}
\beta_0 &\sim N_n(b_0, V_0) \\
\gamma_j|\gamma_{\setminus -j} &\sim Bernoulli(1, \pi_{0j}) \\
Q^{-1} &\sim Wishart\left(\xi, R^{-1}\right) \\
\Sigma^{-1} &\sim Wishart\left(\alpha, S^{-1}\right)
\end{aligned}$$

Estimating these parameters means sampling sequentially from the following conditional densities

1. Sample $\beta_t|\beta_{t-1}, Q, \Sigma, y_t, z_t^*$ for all $t$, where $z_t^* = z_t\Gamma$ and $\Gamma = diag\{\gamma_1, ..., \gamma_n\}$, using the Carter and Kohn (1994) filter and smoother for state-space models (see below)

2. Sample $\gamma_j$ from the density

$$\gamma_j|\gamma_{\setminus -j}, \beta, \Sigma, y, z \sim Bernoulli(1, \widetilde{\pi}_j) \quad \text{(B.3)}$$

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

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

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 $\theta_t^*$ to be equal to $\theta_t$ but with its $j-th$ element $\theta_{t,j} = \beta_{t,j}$ (i.e. when $\gamma_j = 1$). Similarly, define $\theta_t^{**}$ to be equal to $\theta_t$ 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

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

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 $\Sigma^{-1}$ from the density

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

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

## B.1 Carter and Kohn (1994) algorithm:

Consider a state-space model of the following form

$$\begin{aligned} y_t &= z_t a_t + u_t && \text{(B.8a)}\\ a_t &= a_{t-1} + v_t && \text{(B.8b)}\\ u_t &\sim N\left(0,R\right), v_t \sim N\left(0,W\right) \end{aligned}$$

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 recurssions 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 recurssions for $t = T-1,...,1$ and obtain the smooth estimates $a_{t|t+1}$ and

$P_{t|t+1}$ given by the backward recurssion:

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

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 Efficient sampling of the variable selection indicators

In order to sample all the $\gamma_j$ 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 $\gamma_j$ at iteration $s$ of the Gibbs sampler to be denoted by $\gamma_j^s$, and the (candidate) draw of $\gamma_j$ at iteration $s+1$ to be denoted by $\gamma_j^{s+1}$. An efficient acceptance/rejection step for generating $\gamma_j$ is:

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

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 $\gamma_j^{s+1}$ from the Bernoulli density $\gamma_j|\gamma_{\backslash -j}, b, y, z \sim Bernoulli\left(1, \widetilde{\pi}_j\right)$, 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.

# D Further extensions of the basic model

## D.1 Alternative restriction search for VARs with constant parameters

George, Sun and Ni (2008) extend the stochastic-search variable selection (SSVS) algorithm of George and McCullogh (1993) to VAR models. Their approach also assumes the full model with all $n$ parameters and restricts each parameter $\beta_j$ to

be approximately zero when $\gamma_j = 0$, for $j = 1, ..., n$, by restricting the prior of $\beta_j$ towards (but not exactly equal to) a point mass at zero. Specifically, the indicators $\gamma$ do not appear on the VAR equation, but enter in the priors of the VAR regression coefficients in a hierarchical manner

$$\beta_j|\gamma_j \sim \left(1-\gamma_j\right) N\left(0, c_{1j}^2\right) + \gamma_j N\left(0, c_{2j}^2\right). \tag{D.1}$$

In this prior specification $c_{1j}^2 \to 0$ (or $c_{1j}^2 = 0$) while $c_{2j}^2 \to \infty$, so that when $\gamma_j = 0$ the prior for $\beta_j$ is $N\left(0, c_{1j}^2\right)$ which will shrink the posterior towards $0$ (since $c_{1j}^2$ is set very low). When $\gamma_j = 1$, the prior for $\beta_j$ is $N\left(0, c_{2j}^2\right)$ which leaves its posterior unrestricted. There are sensitivity issues arising with the choice of $\left(c_{1j}^2, c_{2j}^2\right)$ in this prior, but these are easily addressed using Empirical Bayes methods, or assigning a hyperprior on these parameters; see George and McCullogh (1997), Chipman et al. (2001) and George and Foster (2000). Korobilis (2008) shows that in an 8-variable VAR with 124 predictors, SSVS selects very parsimonious models which forecast better than conventional model selection approaches based on information criteria. Jochmann, Koop and Strachan (2008) also find that SSVS results in improvement in forecasts coming from a 3-variable VAR with structural breaks.

## D.2 Restrictions on the VAR covariance matrix

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), 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 $\delta$. In this decomposition $\Omega$ is a diagonal matrix and $A$ is a lower triangular matrix with 1's on the diagonal. Hence model selection proceeds by setting

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

where $\alpha_i$ is each of the $m\left(m-1\right)/2$ non-zero and non-one elements of $A$. In a similar attempt, Wong, Carter and Kohn (2003) use the SSVS prior (D.1) to implement restrictions on the elements $\alpha_i$. George, Sun and Ni (2008) apply these methods to the covariance matrix of a VAR. The Smith and Kohn (2002) algorithm however, is not based on specific assumptions about the form of the prior in order to sample the elements of $A$, $T$ and $\delta$. 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 time varying as in the popular Heteroskedastic TVP-VARs of Primiceri (2005), Canova and Gambetti (2009) and Cogley and Sargent (2001, 2005).

## D.3 Selection between constant and time-varying parameters

Shively and Kohn (1997) show how to use Bayesian variable selection in the problem of determining which coefficients are time-varying or constant, in a univariate regression problem. Their model can be written in terms of the TVP-VAR model consisting of equations (10) and (11), where now the time-varying coeffiecients $b_t$ are defined as

$$\beta_t = \beta_{t-1} + \eta_t \tag{D.2}$$
$$\eta_t \sim N(0, \Delta' Q \Delta) \tag{D.3}$$

In this specification $\Delta$ is - similarly to the matrix $\Gamma$ defined for selection of coefficients - an $n \times n$ diagonal matrix with elements $\delta_i$, $i = 1, ..., n$. When $\delta_i = 0$, the $i - th$ column and row of the covariance matrix of $\eta_t$ is zero, and the coefficient in the $i - th$ equation of (D.2) is updated each time-period as $b_{it} = b_{it-1}$, which implies that the parameter remains constant over the full sample $T$. Note however that in this specification - no matter how usefull it might be - analytical derivations of the conditional posterior densities of $\Delta$ are not readily available. When $\delta_i = 0$, $\Delta$ (and subsequently the covariance $\Delta' Q \Delta$) is not invertible. Shively and Kohn (1997) use numerical integration to evaluate the posterior. Nevertheless, this approximation might become cumbersome in multivariate forecasting models with large $m$, while at the same time lacks the simplicity of the expressions used to restrict the regression coefficients and the covariance matrix.

# References


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

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 M. Ciccarelli. (2004). Forecasting and turning point predictions in a Bayesian panel VAR model. *Journal of Econometrics*, 120, 327-359.

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.

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. (2005). Drifts and volatilities: Monetary policies and outcomes in the post WWII U.S.. *Review of Economic Dynamics*, 8, 262-302.

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

Cremers, K. (2002). Stock return predictability: A Bayesian model selection perspective. *Review of Financial Studies*, 15, 1223-1249.

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

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

Dellaportas, P., Foster, J. J., and I. Ntzoufras. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12, 27-36.

George, E. I., and D.P. Foster. (2000). Calibration and empirical bayes variable selection. *Biometrika*, 87, 731-747.

George, E. I., and R.E. McCulloch. (1997). Approaches to Bayesian variable selection. *Statistica Sinica*, 7, 339-379.

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. (2008). Bayesian forecasting using stochastic search variable selection in a VAR subject to breaks. Unpublished manuscript.

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. RCEA Working Paper 47-09.

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. (2008). Time-varying VARs with inequality restrictions. Unpublished manuscript.

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

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

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

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

Shively, T. S., and R. Kohn. (1997). A Bayesian approach to model selection in stochastic coefficient regression models and structural time series models. *Journal of Econometrics*, 76, 39-52.

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

Smets, F., and R. Wouters. (2003). An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area. *Journal of the European Economic Association*, 1, 1123-1175.

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 Mark W. Watson. (2005). Implications of dynamic factor models for VAR analysis. Unpublished paper, Princeton University.

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.