Choice of Variables in Vector Autoregressions*

Marek Jarociński
European Central Bank

Bartosz Maćkowiak
European Central Bank and CEPR

Preliminary and incomplete, February 2011

Abstract

Suppose that a dataset with N endogenous variables is available. N_1 of those variables are the variables of interest. You want to estimate a vector autoregression (VAR) with the variables of interest. Which of the remaining $N - N_1$ variables, if any, should you include in the VAR with the variables of interest? We propose a Bayesian methodology to answer this question. This question arises in most applications of VARs, whether in prediction or impulse response analysis. We apply the methodology to predict a vector of macroeconomic variables in the euro area.

*Jarociński: European Central Bank, Kaiserstrasse 29, 60311 Frankfurt am Main, Germany (e-mail: marek.jarocinski@ecb.int); Maćkowiak: European Central Bank, Kaiserstrasse 29, 60311 Frankfurt am Main, Germany (e-mail: bartosz.mackowiak@ecb.int). We thank Chris Sims for a helpful conversation. The views expressed in this paper are solely those of the authors and do not necessarily reflect the views of the European Central Bank.

1 Introduction

Vector autoregressions (VARs), introduced in Sims (1980), are a standard tool in macroeconomics. Macroeconomists use VARs to form out-of-sample, multi-period predictions of interdependent time series. Furthermore, macroeconomists use VARs to identify the dynamic effects of structural shocks such as monetary policy shocks, fiscal policy shocks, technology shocks, and news shocks. Before we run a VAR, we rarely know *a priori* exactly which variables to include. Typically, we begin with a small set of variables that we are interested in and we realize that, in principle, we can include many other variables in the VAR. For example, when we set out to forecast with a VAR the joint path of the price level and GDP, we realize that, in principle, many variables can improve our forecasts. To take another example, when we set out to estimate with a VAR the impulse response of hours worked to a technology shock, we realize that, in principle, the inclusion or exclusion of many "control variables" can affect our estimates.

In this paper, we study the choice of variables in a VAR. We consider the following question. Suppose that a dataset with N endogenous variables is available. Furthermore, suppose that N_1 of those variables are the variables of interest. We want to estimate a VAR with the variables of interest. Which of the remaining $N - N_1$ variables, if any, should we include in the VAR with the variables of interest?

We develop a Bayesian methodology to answer this question. At the center of this methodology lies the concept of block-exogeneity. Consider a partition of a vector of variables y into two subvectors, $y = \{y_i, y_j\}$. Think of the following question: Do the elements of y_j improve a forecast of any variable in y_i, compared with a forecast that is based on lagged values of all the elements of y_i alone? If the answer to this question is "no", y_i is said to be block-exogenous with respect to y_j. It turns out that the decision about which of the remaining variables to include in the VAR with the variables of interest involves evaluating, via marginal likelihood, block-exogeneity restrictions in the VAR with all N variables. If the variables of interest are block-exogenous with respect to some other variable y_k, we can "drop" this other variable y_k from the VAR. That is to say, we can rewrite the VAR in recursive form, so that the variables of interest are explained only by lagged values of themselves and lagged values of other variables in the dataset, but not by lagged values of

y_k. Thereafter, we can use for prediction and impulse response analysis only the part of the VAR including the variables of interest but excluding y_k.

In an important recent paper, Bańbura et al. (2010) show that a large VAR with as many as 131 variables forecasts better out-of-sample than small VARs. This finding appears to suggest that we can simply estimate a VAR with all N variables. However, Bańbura et al. (2010) also show that a VAR with 20 variables achieves much of the improvement in the predictive performance over small VARs. This finding raises the following questions: How do we decide which 20 variables to include in a VAR? How do we decide whether to include 20, 15, or 25 variables? Our methodology addresses these questions in a systematic way.

Cushman and Zha (1997) and Zha (1999) analyze VARs with a block-exogeneity restriction from the Bayesian perspective. Both papers are interested in all variables being modeled and do not use block-exogeneity to justify dropping variables. Furthermore, the authors either impose block-exogeneity *a priori* or test for it using classical methods.

This paper is related to the literature on Bayesian variable selection in VARs initiated in George et al. (2008). See also Jochmann et al. (2010) and Korobilis (2010). This literature considers general patterns of zero restrictions on VAR coefficients and averages over different patterns of zero restrictions. In contrast, we are concerned only with a particular type of zero restrictions on VAR coefficients, namely block-exogeneity restrictions, because block-exogeneity restrictions justify reducing the dimension of a VAR. We do not average over different restrictions, instead picking the single best restriction, because our goal is to choose an optimal VAR of a reduced dimension.

Section 2 states the question that we study and describes the methodology that we propose to answer this question. Section 3 discusses the computational aspects of the methodology. Section 4 presentz an empirical application to euro area data. Section 5 describes the relationship between the methodology that we propose and alternatives. Section 6 concludes.

2 The question and the methodology to answer it

This section states the question that we study and proposes a methodology to answer this question.

Throughout this paper, we consider VAR models all of which have the form

$$y(t) = \gamma + B(L)\, y(t-1) + u(t), \tag{1}$$

where $y(t)$ is a vector of endogenous variables in period t, γ is a constant term, $B(L)$ is a matrix polynomial in the lag operator of order $P-1$, and $u(t)$ is a Gaussian random vector with mean zero and covariance matrix Σ conditional on $y(t-s)$ for all $s \geq 1$.

The question that we study is the following. Suppose that a dataset with N endogenous variables is available. Furthermore, suppose that $N_1 < N$ of those variables are the variables of interest. Let $N_2 = N - N_1$. We want to estimate a VAR with the variables of interest. Which of the remaining N_2 variables, if any, should we include in the VAR with the variables of interest?

The rest of this section describes the methodology that we propose to answer this question. At the center of this methodology lies the concept of block-exogeneity.

2.1 Block-exogeneity

Consider a partition of a vector of variables y into two blocks, $y = \{y_i, y_j\}$, and a conformable partition of the VAR model of y:

$$\begin{pmatrix} y_i(t) \\ y_j(t) \end{pmatrix} = \gamma + \begin{pmatrix} B_{ii}(L) & B_{ij}(L) \\ B_{ji}(L) & B_{jj}(L) \end{pmatrix} \begin{pmatrix} y_i(t-1) \\ y_j(t-1) \end{pmatrix} + u(t). \tag{2}$$

Definition 1 *(Hamilton, 1994, p.309)* ***Block-exogeneity:*** *The group of variables represented by y_i is said to be block-exogenous (in the time series sense) with respect to the variables in y_j if the elements of y_j are of no help in improving a forecast of any variable contained in y_i that is based on lagged values of all the elements of y_i alone.*

In the VAR given in equation (2), y_i is block-exogenous to y_j if and only if $B_{ij}(L) = 0$. We make three observations regarding block-exogeneity.

Relation with Granger causality: The restriction $B_{ij}(L) = 0$ is the same as: (i) the statement that the variables in y_j do not Granger-cause any of the variables in y_i, and (ii) the statement that the variables in y_i are Granger causally prior to the variables in y_j.[1] Furthermore, with this restriction the VAR has a recursive form: current y_i is explained only by lagged values of itself, and current y_j is explained by lagged values of itself and lagged values of y_i.

Bayesian test of block-exogeneity: To test the restriction $B_{ij}(L) = 0$ a Bayesian computes the marginal likelihood of the VAR given in equation (2), $p(Y)$, and the marginal likelihood of that VAR conditional on the restriction $B_{ij}(L) = 0$, $p(Y|B_{ij}(L) = 0)$. When the prior probabilities of the unrestricted and restricted model are equal, the Bayesian prefers the specification with the higher marginal likelihood. Note that, with equal prior probabilities, the Bayes factor

$$\frac{p(Y|B_{ij}(L) = 0)}{p(Y)}$$

is equal to the posterior odds in favor of the restriction.

Relation of the Bayesian test to out-of-sample fit: The outcome of this test depends on out-of-sample fit of the VAR without and with the restriction, because, as is well known, marginal likelihood depends on out-of-sample fit. Furthermore, the outcome of this test depends on how well the model fits both y_i and y_j. The reasons why fitting y_j matters are as follows: (i) in order to predict y_i well out-of-sample, the unrestricted model must also predict y_j well out-of-sample, and (ii) the comparison in the test is made with respect to the unrestricted model. In section 5, we compare this approach with other approaches based on out-of-sample fit of y_i only.

These three observations suggest the following principle that we adopt: *Given a dataset with N variables and given that N_1 of those variables are the variables of interest, the decision about which of the remaining N_2 variables to include in the VAR with the variables of interest involves evaluating, via marginal likelihood, block-exogeneity restrictions in the VAR with all N variables.*

[1] For the definition of Granger causal priority see Sims (2010).

2.2 The family of VARs and the choice of the best VAR

We now define a family of VAR models with block-exogeneity restrictions and cast the choice of variables in a VAR as the choice of one model from this family.

Consider a family of models Ω such that: (i) all models in Ω are VARs with N variables represented by y with zero, one or more block-exogeneity restrictions, and (ii) the variables in y are ordered so that the variables of interest represented by y_1 are in the first block. Note that with zero block-exogeneity restrictions the VAR has one block.

Since our approach is Bayesian, we need to define a prior about parameters in each model from this family. The prior about the parameters in the unrestricted model is $p(B, \Sigma)$, where B is a matrix collecting γ and the parameters in $B(L)$ in equation (1). The prior in a model $\omega \in \Omega$ is given by $p_\omega(B, \Sigma) = p(B, \Sigma | B_\omega = 0)$, where B_ω denotes the parameters in the matrix B that are set to zero because of one or more block-exogeneity restrictions.

We evaluate the marginal likelihood of each model in Ω, $p(Y|\omega)$, and choose the model with the highest marginal likelihood, ω^*.

Note that, in the end, the model of interest is only the first block of the best model ω^*. The reasons are that (i) y_1 is the vector containing the variables of interest, and (ii) once we have found the best model ω^*, given the definition of block-exogeneity, only the first block of the best model ω^* is relevant for modeling y_1.

2.3 A trivariate example

Given that our interest is in y_1, it is not immediately intuitive why we study a family that includes VARs with more that one block-exogeneity restriction. In this subsection, we study an example to illustrate our approach and motivate our choice of model family.

Suppose that $N_1 = 1$ and $N = 3$, i.e. there is one variable of interest and the dataset contains three variables. Setting $N = 3$ we can rewrite equation (1) as follows

$$\begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{pmatrix} = \gamma + \begin{pmatrix} B_{11}(L) & B_{12}(L) & B_{13}(L) \\ B_{21}(L) & B_{22}(L) & B_{23}(L) \\ B_{31}(L) & B_{32}(L) & B_{33}(L) \end{pmatrix} \begin{pmatrix} y_1(t-1) \\ y_2(t-1) \\ y_3(t-1) \end{pmatrix} + u(t) . \tag{3}$$

We think of y_1 as the variable of interest and we think of y_2 and y_3 as the remaining variables. In this example, the question that we study reduces to: Should we include y_1, y_2

and y_3 in the VAR, should we include y_1 and y_2, should we include y_1 and y_3, or should we include only y_1?

We make four observations about this example.

First, "including y_1 and y_2 in the VAR" means estimating the VAR given in equation (3) with the restriction $B_{13}(L) = B_{23}(L) = 0$; "including y_1 and y_3 in the VAR" means estimating that VAR with the restriction $B_{12}(L) = B_{32}(L) = 0$; and "including only y_1 in the VAR" means estimating that VAR with the restriction $B_{12}(L) = B_{13}(L) = 0$.

Note that with each restriction the VAR has a recursive form. For example, consider the restriction $B_{13}(L) = B_{23}(L) = 0$. With this restriction, current y_1 and y_2 are explained only by lagged values of themselves, and current y_3 is explained by lagged values of itself and lagged values of y_1 and y_2.

Second, the restriction $B_{13}(L) = B_{23}(L) = 0$ is the same as the statement that y_1 and y_2 are block-exogenous to y_3; the restriction $B_{12}(L) = B_{32}(L) = 0$ is the same as the statement that y_1 and y_3 are block-exogenous to y_2; and the restriction $B_{12}(L) = B_{13}(L) = 0$ is the same as the statement that y_1 is block-exogenous to y_2 and y_3.[2]

Third, to test the block-exogeneity restrictions, a Bayesian computes the marginal likelihood of the VAR given in equation (3) without and with each block-exogeneity restriction. When the prior probabilities of the unrestricted and restricted models are equal, the Bayesian prefers the specification with the highest marginal likelihood.

The outcome of this procedure depends on how well the model fits y_1, y_2 and y_3. The reasons why fitting y_2 and y_3 matters are as follows: (i) in order to predict y_1 well out-of-sample, the unrestricted model must also predict y_2 and y_3 well out-of-sample, (ii) in order to predict y_1 well out-of-sample, the model with the restriction $B_{13}(L) = B_{23}(L) = 0$ must also predict y_2 well out-of-sample and the model with the restriction $B_{12}(L) = B_{32}(L) = 0$ must also predict y_3 well out-of-sample, and (iii) the restricted models are compared with one another and with the unrestricted model; to make this comparison, we need to keep track of the out-of-sample fit of all variables.

Fourth, one could consider other restrictions. One could be concerned that the result of

[2] We use the term "block-exogeneity" for consistency throughout the paper, even though in the trivariate example some "blocks" include only one variable.

the block-exogeneity tests depends on whether another restriction is imposed. In particular, one could be concerned that modeling the interaction between the remaining variables matters for the result of the block-exogeneity tests.

Our approach to modeling the interaction between the remaining variables is to consider block-exogeneity restrictions among the remaining variables. In the trivariate case, there are five block-exogeneity restrictions: the three restrictions stated earlier and the following two restrictions. The restriction $B_{12}(L) = B_{13}(L) = B_{23}(L) = 0$, which is the same as the statement that y_1 is block-exogenous to y_2 and y_3 *and* y_2 block-exogenous to y_3. Furthermore, the restriction $B_{12}(L) = B_{13}(L) = B_{32}(L) = 0$, which is the same as the statement that y_1 is block-exogenous to y_2 and y_3 *and* y_3 is block-exogenous to y_2. Note that if either one of the two specifications introduced here is selected, the decision is "include only y_1 in the VAR".

2.4 Notation for block-exogeneity and the size of Ω

We use the notation

$$y_i \; BE \; y_j$$

as a stand-in for the statement "the block of variables y_i is block-exogenous to the block of variables y_j".

Block-exogeneity is transitive. That is, $y_i \; BE \; y_j$ and $y_j \; BE \; y_k$ implies that $y_i \; BE \; y_k$.

Each model $\omega \in \Omega$ is fully characterized by the pattern of BE restrictions such that y_1 is always in the first block. That is, each model $\omega \in \Omega$ is fully characterized by the pattern of BE restrictions of the form

$$\{y_1, y_{2.1}\} \; BE \; y_{2.2} \; BE \; ... \; BE \; y_{2.G}, \qquad 0 < G < N_2, \tag{4}$$

where $y_{2.1}, ..., y_{2.G}$ is a partition of y_2 into G subsets. Expression (4) states that the first block of model ω includes the vector $\{y_1, y_{2.1}\}$. This is a general statement, because each of the vectors $y_{2.1}, ..., y_{2.G}$ can be empty. For example, in the VAR with zero block-exogeneity restrictions the vector $y_{2.1}$ is non-empty and each of the vectors $y_{2.2}, ..., y_{2.G}$ is empty.

It may appear that the family of VARs Ω, characterized by expression (4), is small. For example, block-exogeneity restrictions in a VAR are only a small subset of all possible zero restrictions in a VAR. However, it turns out that the family of VARs Ω is very large. In particular, the family of VARs Ω grows very quickly with N_2. To see this, consider the number of all possible block-exogeneity restrictions among N_2 variables, denoted $C(N_2)$. Since block-exogeneity is a transitive relation, $C(N_2)$ is equal to the number of weak orders of N_2 elements. One can show that

$$C(N_2) = \sum_{k=0}^{N_2-1} \binom{N_2}{k} C(k)$$

and asymptotically

$$C(N_2) \approx \frac{N_2!}{2(\ln 2)^{N_2+1}}.$$

See OEIS (2011). The size of the family of VARs Ω is $K(N_2) = 2C(N_2)$, i.e. the size of the family of VARs Ω equals twice the number of block-exogeneity restrictions among N_2 variables in y_2. The reason that the multiplication by two is necessary is that, given each pattern of block-exogeneity restrictions within y_2, we can either have that the block-exogeneity between y_1 and y_2 or not. In terms of expression (4), $y_{2.1}$ can be either empty or non-empty.

Consider a few examples. In our trivariate example, $K(N_2 = 2) = 6$. Next, $K(N_2 = 3) = 26$, $K(N_2 = 4) = 150$, $K(N_2 = 5) = 1082$, $K(N_2 = 6) = 9366$, $K(N_2 = 7) = 94586$, and so on.

3 The methodology: computational aspects

This section discusses the computational aspects of the methodology that we propose in this paper.

3.1 Prerequisites: conjugate prior and posterior

The likelihood of the VAR given in equation (1), conditional on initial observations, is

$$p(Y|B,\Sigma) = (2\pi)^{-NT/2} |\Sigma|^{-T/2} \exp\left(-\frac{1}{2}\operatorname{tr}(Y - XB)'(Y - XB)\Sigma^{-1}\right), \tag{5}$$

where N is the length of the vector $y(t)$, T is the number of observations in the sample,

$$\underset{T\times N}{Y} = \begin{pmatrix} y(1)' \\ y(2)' \\ \vdots \\ y(T)' \end{pmatrix}, \quad \underset{K\times N}{B} = \begin{pmatrix} B_1' \\ \vdots \\ B_P' \\ \gamma' \end{pmatrix},$$

and

$$\underset{T\times K}{X} = \begin{pmatrix} y(0)' & y(-1)' & \dots & y(1-P)' & 1 \\ y(1)' & y(0)' & \dots & y(2-P)' & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ y(T-1)' & y(T-2)' & \dots & y(T-P)' & 1 \end{pmatrix}$$

Throughout this paper, we assume that the prior density of B and Σ is conjugate, i.e. the prior density satisfies

$$p(B,\Sigma) \propto |\Sigma|^{-(\tilde{\nu}+K+N+1)/2} \exp\left(-\frac{1}{2}\operatorname{tr}(\tilde{Y}-\tilde{X}B)'(\tilde{Y}-\tilde{X}B)\Sigma^{-1}\right), \tag{6}$$

where $\tilde{\nu}$, $\tilde{Y}$, and $\tilde{X}$ are prior hyperparameters of appropriate dimensions and $K = NP+1$. Section 4.1 discusses the specification of the prior hyperparameters in this paper's application. Let

$$\tilde{Q} = (\tilde{X}'\tilde{X})^{-1}, \quad \tilde{B} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}, \quad \text{and} \quad \tilde{S} = (\tilde{Y}-\tilde{X}\tilde{B})'(\tilde{Y}-\tilde{X}\tilde{B}).$$

It is straightforward to show that, so long as $\tilde{\nu} > 0$, the prior is proper and satisfies

$$p(B,\Sigma) = p(B|\Sigma)p(\Sigma) = \mathcal{N}\left(\operatorname{vec}\tilde{B}, \Sigma\otimes\tilde{Q}\right)\mathcal{IW}\left(\tilde{S},\tilde{\nu}\right),$$

where $\mathcal{N}$ denotes a multivariate normal density and $\mathcal{IW}$ denotes an inverted Wishart density. See Bauwens et al. (1999), Appendix A, for the definitions of the multivariate normal and inverted Wishart densities.

We obtain the posterior by combining prior (6) with likelihood (5). Let $\bar{\upsilon} = \tilde{\nu} + T$,

$$\bar{Y} = \begin{pmatrix} \tilde{Y} \\ Y \end{pmatrix} \qquad \bar{X} = \begin{pmatrix} \tilde{X} \\ X \end{pmatrix},$$

$$\bar{Q} = \left(\bar{X}'\bar{X}\right)^{-1}, \quad \bar{B} = \left(\bar{X}'\bar{X}\right)^{-1}\bar{X}'\bar{Y}, \quad \text{and} \quad \bar{S} = (\bar{Y}-\bar{X}\bar{B})'(\bar{Y}-\bar{X}\bar{B}).$$

It is straightforward to show that, so long as $\bar{\upsilon} > 0$, the posterior is proper and satisfies

$$p(B,\Sigma) = p(B|\Sigma)p(\Sigma) = \mathcal{N}\left(\operatorname{vec}\bar{B}, \Sigma\otimes\bar{Q}\right)\mathcal{IW}\left(\bar{S},\bar{\upsilon}\right).$$

3.2 Bayes factor for a single block-exogeneity restriction

It turns out that, when the interest is in testing a single block-exogeneity restriction in a VAR, there is an analytical expression for the Bayes factor for this test.

Suppose that we partition the vector modeled in the VAR given in equation (1) into two blocks, $y = \{y_i, y_j\}$, and we want to test if y_i is block-exogenous to y_j. Let α denote the column indices of the variables represented by y_i in matrix Y. Let β denote the column indices of the lags of the variables represented by y_j in matrix X.[3] The block-exogeneity restriction is $B_{\beta,\alpha} = 0$. Let $p_{\{y_i,y_j\}}(B, \Sigma)$ denote the prior on Σ and on the free elements of B in the restricted VAR. Recall that $p(B, \Sigma)$ is the prior in the unrestricted VAR satisfying (6). The following result is available. If

$$p_{\{y_i,y_j\}}(B, \Sigma) = p(B, \Sigma | B_{\beta,\alpha} = 0) \tag{7}$$

then the Bayes factor for the comparison between the VAR with $B_{\beta,\alpha} = 0$ with the unrestricted VAR has the property

$$\frac{p(Y|B_{\beta,\alpha} = 0)}{p(Y)} = \frac{p(B_{\beta,\alpha} = 0|Y)}{p(B_{\beta,\alpha} = 0)}, \tag{8}$$

where $p(B_{\beta,\alpha} = 0)$ is the marginal prior density of $B_{\beta,\alpha}$ in the unrestricted model, evaluated at the point $B_{\beta,\alpha} = 0$, and $p(B_{\beta,\alpha} = 0|Y)$ is the marginal posterior density of $B_{\beta,\alpha}$ in the unrestricted model, evaluated at the point $B_{\beta,\alpha} = 0$. This result is known as the Savage-Dickey result. The right-hand-side of expression (8) is known as the Savage-Dickey ratio. See Dickey (1971) and Verdinelli and Wasserman (1995). The Savage-Dickey result states that the Bayes factor for the test of the restriction $B_{\beta,\alpha} = 0$ against the alternative $B_{\beta,\alpha} \neq 0$ is equal to the ratio of the marginal posterior density of $B_{\beta,\alpha}$ at zero to the marginal prior density of $B_{\beta,\alpha}$ at zero.

When priors are as defined in section 3.1, both marginal densities in the Savage-Dickey ratio are available analytically. The marginal prior density of $B_{\beta,\alpha}$ is

$$p(B_{\beta,\alpha}) = Mt(\tilde{B}_{\beta,\alpha}, (\tilde{Q}_{\beta,\beta})^{-1}, \tilde{S}_{\alpha,\alpha}, \tilde{\nu} - N_2), \tag{9}$$

[3] If we define $\bar{\alpha}$ to be the indices of the variables represented by y_j in matrix Y, we have $\beta = (\bar{\alpha}, N + \bar{\alpha}, 2N + \bar{\alpha}, ...(P-1)N + \bar{\alpha})$.

where Mt denotes the matricvariate Student density. See Bauwens et al. (1999), Appendix A.2.7, for the definition of the matricvariate Student density and the proof of equality (9)). Furthermore, the marginal posterior density is

$$p(B_{\beta,\alpha}|Y) = Mt(\bar{B}_{\beta,\alpha}, (\bar{Q}_{\beta,\beta})^{-1}, \bar{S}_{\alpha,\alpha}, \bar{\nu} - N_2), \tag{10}$$

The analytical expressions (9)-(10) are available due to the fact that picking an intersection of rows β and columns α from B preserves the Kronecker structure of the variance of B.

3.3 Bayes factor for multiple block-exogeneity restrictions

When the interest is in testing multiple block-exogeneity restrictions in a VAR, the Kronecker structure of the variance of the restricted subset of B is destroyed and an analytical expression for the Savage-Dickey ratio is unavailable. However, it is straightforward to compute the Savage-Dickey ratio numerically.

Consider the case of two block-exogeneity restrictions

$$y_i \; BE \; y_j \; BE \; y_k. \tag{11}$$

A generalization to G block-exogeneity restrictions is straightforward. Let α_1 denote the column indices of the variables represented by y_i in matrix Y. Let β_1 denote the column indices of the lags of the variables represented by y_j and y_k in matrix X.[4] Let α_2 denote the column indices of the variables represented by y_j in matrix Y. Let β_2 denote the column indices of the lags of the variables represented by y_k in matrix X.[5] The block-exogeneity restrictions are $B_{\beta_1,\alpha_1} = B_{\beta_2,\alpha_2} = 0$. Suppose that condition (7) is satisfied. Then the Bayes factor for the test of the restriction $B_{\beta_1,\alpha_1} = B_{\beta_2,\alpha_2} = 0$ against the unrestricted model is equal to the ratio of the posterior to prior marginal density of $(B_{\beta_1,\alpha_1}, B_{\beta_2,\alpha_2})$ evaluated at zero. The marginal density of $(B_{\beta_1,\alpha_1}, B_{\beta_2,\alpha_2})$ is not available in closed form. However, the density of $((\text{vec}\, B_{\beta_1,\alpha_1})', (\text{vec}\, B_{\beta_2,\alpha_2})')'$ conditional on Σ is multivariate normal. The conditional prior density is

$$p\left(\begin{pmatrix} \text{vec}\, B_{\beta_1,\alpha_1} \\ \text{vec}\, B_{\beta_2,\alpha_2} \end{pmatrix} |\Sigma\right) = N\left(\begin{pmatrix} \text{vec}\, \tilde{B}_{\beta_1,\alpha_1} \\ \text{vec}\, \tilde{B}_{\beta_2,\alpha_2} \end{pmatrix}, \begin{pmatrix} \Sigma_{\alpha_1,\alpha_1}\tilde{Q}_{\beta_1,\beta_1} & ... \\ \Sigma_{\alpha 2,\alpha_1}\tilde{Q}_{\beta_2,\beta_1} & \Sigma_{\alpha_2,\alpha_2}\tilde{Q}_{\beta_2,\beta_2} \end{pmatrix}\right).$$

[4] If we define $\alpha_{\bar{1}} = (\alpha_2, \alpha_3)$, we have $\beta_1 = (\alpha_{\bar{1}}, N + \alpha_{\bar{1}}, 2N + \alpha_{\bar{1}}, ...(P-1)N + \alpha_{\bar{1}})$.

[5] If we define $\alpha_{\bar{2}} = \alpha_3$, we have $\beta_1 = (\alpha_{\bar{2}}, N + \alpha_{\bar{2}}, 2N + \alpha_{\bar{2}}, ...(P-1)N + \alpha_{\bar{2}})$.

The marginal prior density at zero can be approximated from M Monte Carlo draws of Σ as

$$p\left(\begin{array}{c}\operatorname{vec} B_{\beta_1,\alpha_1} \\ \operatorname{vec} B_{\beta_2,\alpha_2}\end{array} = 0\right) = \frac{1}{M}\sum_{m=1}^{M} p\left(\begin{array}{c}\operatorname{vec} B_{\beta_1,\alpha_1} \\ \operatorname{vec} B_{\beta_2,\alpha_2}\end{array} = 0 \middle| \Sigma^m\right)$$

The approximation of the marginal posterior density at zero is analogous, with the variables with "tildes" replaced by the variables with "upper bars".

[Note that this works also for an arbitrary pattern of zero restrictions in B.]

3.4 Marginal likelihood of the full model

The marginal likelihood of a VAR model with the conjugate prior is available in closed form. The marginal likelihood of the VAR model with our notation is

$$p(Y) = \pi^{-NT/2} \frac{|\tilde{X}'\tilde{X}|^{N/2}}{|\bar{X}'\bar{X}|^{N/2}} \frac{\Gamma_N\left(\frac{\tilde{\nu}+T}{2}\right)}{\Gamma_N\left(\frac{\tilde{\nu}}{2}\right)} \frac{|\tilde{S}|^{\tilde{\nu}/2}}{|\bar{S}|^{(\tilde{\nu}+T)/2}}, \tag{12}$$

where Γ_N denotes the multivariate Gamma function defined in expression (22) in Appendix A. See Appendix A for a derivation of expression (12).

We combine the marginal likelihood of the unrestricted VAR model, $p(Y)$ with the posterior odds ratio between a restricted model and the unrestricted model, $p(Y|restr.)/p(Y)$ to obtain the marginal likelihood of the restricted model, $p(Y|restr.)$. We use the relationship $p(Y|restr.) = p(Y) \times p(Y|restr.)/p(Y)$.

3.5 Finding the best model when Ω is too large to check all models

When there are too many block-exogeneity restrictions for us to evaluate all of them, we search for the best VAR using the Markov Chain Monte Carlo Model Composition (MC^3) algorithm of Madigan and York (1995).

We implement the MC^3 algorithm as follows. Given a particular model, i.e. a pattern of block-exogeneity restrictions (4), we define the *neighborhood* of this pattern. The neighborhood includes the given pattern and all patterns that differ from the given pattern by the position of only one variable. There are possible four differences in the position of a variable. The variables can (i) join the previous block, (ii) join the next block, (iii) become a block on its own prior to its current block, and (iv) become a block on its own posterior to its current block. Only subsets of these four possibilities apply to variables in the first block and the last block as well as to variables that are in a block of only one or two variables. We describe the neighborhood in detail in Appendix C.

The MC^3 chain moves as follows. The neighborhood of pattern ω is denoted $nbr(\omega)$ and there are $\#nbr(\omega)$ patterns in this neighborhood. We attach equal probability to each pattern in the neighborhood and randomly draw the candidate pattern ω' from the neighborhood. We accept this draw with probability

$$\min\left\{1, \frac{\#nbr(\omega)p(Y|\omega')}{\#nbr(\omega')p(Y|\omega)}\right\}$$

We continue drawing models until the chain converges. [To be completed.]

4 Application to the euro area

4.1 Prior

We construct our prior in two steps: (i) we start with an initial prior formulated before seeing any data, and (ii) we combine the initial prior with the training sample. Accordingly, matrices $\tilde{Y}, \tilde{X}$ in expression (6) consist of two blocks.

$$\tilde{Y} = \begin{pmatrix} Y_{SZ} \\ Y_{ts} \end{pmatrix}, \quad \tilde{X} = \begin{pmatrix} X_{SZ} \\ X_{ts} \end{pmatrix},$$

and the terms Y_{SZ}, Y_{ts}, X_{SZ}, X_{ts}, and T_{ts} are defined below.

The initial prior is the prior proposed by Sims and Zha (1998). We implement the Sims-Zha prior by creating dummy observations Y_{SZ} and X_{SZ}. The term $\tilde{\nu}$ in expression (6) also belongs to the initial prior. Appendix B gives the details concerning the initial prior.

In addition to the Sims-Zha prior we add to the prior the information from the pre-EMU period 1989Q1 to 1998Q4. Y_{ts} and X_{ts} denote the matrices with the data from this training sample. We have found that adding this training sample improves the marginal likelihood of the VAR in the Euro Area sample.

4.2 Results

This is a low-dimensional example. In the next draft, this example will be replaced by a high-dimensional example.

The goal is to estimate a VAR with real output (GDP), the price level (HICP) and the short-term interest rate (Eonia) in the euro area. In this example, we study whether it is useful to add the 10-year bond yield and the M3 money aggregate to the VAR when the interest is in modeling the crisis sample, 2007Q3 to 2010Q2. Accordingly, we have $y_1 = \{\text{GDP, HICP, Eonia}\}$ and $y_2 = \{\text{10-year bond yield, M3}\}$. The VAR is specified in levels, it includes a constant term and two lags of the endogenous variables. We use two versions of the Sims-Zha prior: with $\lambda_1 = 0.4$ (looser Sims-Zha prior) and with $\lambda_1 = 0.2$ (tighter Sims-Zha prior), where λ_1 is the hyperparameter controlling the overall tightness of the prior. The remaining hyperparameters of the Sims-Zha prior are given in Appendix B. The training sample prior places a subjective weight of 10 observations on the pre-crisis EMU sample, 1999Q1 to 2007Q2, and zero weight on the pre-EMU sample, 1989Q1 to 1998Q4. We have chosen these weights as they maximize the marginal likelihood of a VAR model of y_1 in the crisis sample. We have tried all combinations of subjective weights on the grid of 0, 5, 10, 15, 20, 30 and 40 observations, and the (10, 0) combination delivered the highest marginal likelihood.

With $N_2 = 2$ there are six VARs in Ω. Table 1 reports the marginal likelihoods of all these models. Panel A shows results obtained with the looser Sims-Zha prior and Panel B shows results obtained with the tighter Sims-Zha prior.

block-exogeneity pattern	log p(y)
1. GDP,HICP,EONIA,M3,10-year bond yield	94.4
2. GDP,HICP,EONIA,M3 *BE* 10-year bond yield	94.6
3. GDP,HICP,EONIA,10-year bond yield *BE* M3	95.8
4. GDP,HICP,EONIA *BE* M3,10-year bond yield	95.3
5. GDP,HICP,EONIA *BE* M3 *BE* 10-year bond yield	95.3
6. GDP,HICP,EONIA *BE* 10-year bond yield *BE* M3	95.8

Table 1 Panel A looser Sims-Zha prior ($\lambda_1 = 0.4$)

block-exogeneity pattern	log p(y)
1. GDP,HICP,EONIA,M3,10-year bond yield	93.6
2. GDP,HICP,EONIA,M3 *BE* 10-year bond yield	93.4
3. GDP,HICP,EONIA,10-year bond yield *BE* M3	94.3
4. GDP,HICP,EONIA *BE* M3,10-year bond yield	93.7
5. GDP,HICP,EONIA *BE* M3 *BE* 10-year bond yield	93.6
6. GDP,HICP,EONIA *BE* 10-year bond yield *BE* M3	94.0

Table 1 Panel B: tighter Sims-Zha prior ($\lambda_1 = 0.2$)

We would like to emphasise two findings. First, the best VAR with the variables of interest does not include M3. The best VAR with the variables of interest is either the VAR with the variables of interest and the 10-year bond yield (model 3) or the VAR with the variables of interest only (model 6). The VARs with M3 have lower marginal likelihood, irrespective of the tightness of the prior.

Second, the VARs with the looser prior fit better than the VARs with the tighter prior. Note that the tighter prior is the standard Sims-Zha prior. This finding agrees with the findings of Giannone et al. (2010) who show, using other data, that looser Sims-Zha priors yield better out-of-sample fit in small VARs. A VAR with five variables is a small VAR.

5 Alternative approaches and why there are less attractive

The Bayesian approach to model comparison requires models to have the same endogenous variables. The statistic used for comparing models in the Bayesian approach is the marginal likelihood, i.e. the model-implied prior predictive density of the endogenous variables evaluated at the actually observed data. Marginal likelihoods are thus only comparable when the endogenous variables are the same in the compared models.

The approach used in this paper is to compare marginal likelihoods of all N variables, $p(Y|\omega) = p(Y_1, Y_2|\omega)$, even though we are ultimately only interested in the first block of variables. $p(Y)$ is affected by the fit of the model for all variables y_2, also those that belong to further blocks and are thus not useful for modeling y_1. This forces us to consider different alternative models of these dropped variables and makes the relevant family of models Ω

larger.

This section compares $p(Y)$ with three statistics that focus on the fit of the model for y_1 only. These three statistics are constructed so that they depend only on the first block of variables, $y_1, y_{2.1|\omega}$. Therefore, these statistics are independent from how we model other variables in y_2, those that are not useful for modeling y_1. This seems attractive. However, we show that the focus on y_1 is bought at the price of discarding useful data evidence.

5.1 Three statistics measuring the fit of the model for y_1

The first statistic about y_1 is the marginal predictive density of Y_1, that is, the marginal likelihood of $(Y_1, Y_{2.1|\omega})$ marginalized with respect to $Y_{2.1|\omega}$

$$p(Y_1|\omega) = \int p(Y_1, Y_{2.1}|\omega) dY_{2.1|\omega}. \tag{13}$$

The second statistic about y_1 is the predictive density of Y_1 conditional on the actually observed $Y_{2.1|\omega}$

$$p(Y_1|Y_{2.1|\omega}, \omega) = \frac{p(Y_1, Y_{2.1}|\omega)}{\int p(Y_1, Y_{2.1}|\omega) dY_{2.1|\omega}}. \tag{14}$$

The third statistic about y_1 is the predictive density score for Y_1 at horizon h (h may be a vector of horizons).

$$g(Y_1, h|\omega) = \prod_{t=1}^{T-\max(h)} p(y_1(t+h)|y(i : i < t), \omega) \tag{15}$$

This statistic is used in Andersson and Karlsson (2007). Similar statistics are discussed in Geweke and Amisano (2011), Eklund and Karlsson (2007).

5.2 Interpretation in terms of probabilities of models

[To be written.]

5.3 Relation with out-of-sample fit

This subsection illustrates the relation of the statistics $p(Y|\omega), p(Y_1|\omega), p(Y_1|Y_{2.1}, \omega)$ and $g(Y_1, h|\omega)$ with the out-of-sample fit of model ω. It is useful here to make it explicit that all the discussed statistics are conditional on the P pre-sample observations $y_1(-P+1, ..., 0)$

and $y_2(-P+1,...,0)$. We follow the discussion in Geweke (2005) p.67 and partition the sequence of dates $1,...,T$ using a strictly increasing sequence of integers $\{s_j\}_{j=0}^{Q}$ with $s_0 = 0$ and $s_Q = T$. Next, we rewrite $p(Y)$ and the three statistics on Y_1 in terms of models' predictive densities for the periods $1...s_1$, $s_1+1...s_2$ etc. up to $s_{Q-1}...T$.

$$p(Y|\omega) = p(y_1(1,...,T), y_2(1,...,T)|y_1(-P+1,...,0), y_2(-P+1,...,0), \omega) = \\ = \prod_{j=1}^{Q} p(y_1(s_{j-1}+1,...,s_j), y_2(s_{j-1}+1,...,s_j)|y_1(-P+1,...,s_{j-1}), y_2(-P+1,...,s_{j-1}), \omega) \tag{16}$$

$$p(Y_1|\omega) = p(y_1(1,...,T)|y_1(-P+1,...,0), y_{2.1|\omega}(-P+1,...,0), \omega) = \\ = \prod_{j=1}^{Q} p(y_1(s_{j-1}+1,...,s_j)|y_1(-P+1,...,s_{j-1}), y_{2.1|\omega}(-P+1,...,0), \omega) \tag{17}$$

$$p(Y_1|Y_{2.1|\omega}, \omega) = p(y_1(1,...,T)|y_1(-P+1,...,0), y_{2.1|\omega}(-P+1,...,T), \omega) = \\ = \prod_{j=1}^{Q} p(y_1(s_{j-1}+1,...,s_j)|y_1(-P+1,...,s_{j-1}), y_{2.1|\omega}(-P+1,...,T), \omega) \tag{18}$$

$$g(Y_1, \{s_j\}_{j=0}^{Q}|\omega) = \prod_{j=1}^{Q} p(y_1(s_{j-1}+1,...,s_j)|y_1(-P+1,...,s_{j-1}), y_{2.1|\omega}(-P+1,...,s_{j-1}), \omega) \tag{19}$$

The following lessons emerge from comparing equations (17)-(19) with equation (16).

First, $p(Y_1|\omega)$ uses least information and thus delivers smallest differences between models. This is so because it only conditions on the variables y_2 up to period 0. To consider an extreme (but not unrealistic) case, suppose that i) all variables in y_2 are normalized so that they all have the same value in period 0, ii) the VAR has one lag and iii) we use a Sims-Zha prior. It is straightforward to show that in this case $p(Y_1|\omega)$ will be the same in all models with the same number of variables in $y_{2.1|\omega}$.

Second, $p(Y_1|Y_{2.1|\omega}, \omega)$ is not really an out-of-sample measure since it conditions on $y_{2.1|\omega}(-P+1,...,T)$. This statistic measures how well model ω captures the relation between

y_1 and y_2. But to forecast y_1 out-of-sample we need to both caputre this relation well and also to forecast $y_{2.1|\omega}$ out-of-sample. $p(Y_1|Y_{2.1|\omega}, \omega)$ does not measure the out-of-sample fit of $y_{2.1|\omega}$.

Third, $g(Y_1, \{s_j\}_{j=0}^{Q}|\omega)$ uses the same information set as $p(Y)$. It is thus free of the disadvantages of $p(Y_1|\omega)$ and $p(Y_1|Y_{2.1|\omega}, \omega)$ discussed above. However, $p(Y_1|\omega)$, $p(Y_1|Y_{2.1|\omega}, \omega)$ and $g(Y_1, \{s_j\}_{j=0}^{Q}|\omega)$ share the following disadvantage: they ignore the evidence on model fit coming from the fit of y_2 forecasts. In all models $\omega \in \Omega$, y_1 is useful for predicting y_2. Thus, models where the first block $y_1, y_{2.1|\omega}$ is well chosen also have an edge in forecasting the remaining variables, $y_{2.g>1|\omega}$. This additional evidence is used in $p(Y)$ to discriminate between models. In contrast, $p(Y_1|\omega), p(Y_1|Y_{2.1}, \omega)$ and $g(Y_1, h|\omega)$ ignore this additional evidence.

Finally, note that any partition $\{s_j\}_{j=0}^{Q}$ leads to the same values of statistics $p(Y|\omega)$, $p(Y_1|\omega)$ and $p(Y_1|Y_2, \omega)$. In contrast, $g(Y_1, \{s_j\}_{j=0}^{Q}|\omega)$ is different for different partitions $\{s_j\}_{j=0}^{Q}$. For example, the best model for forecasting one period ahead may differ from the best model for forecasting four periods ahead. This is a practical problem in using $g(Y_1, \{s_j\}_{j=0}^{Q}|\omega)$.

5.4 Computational issues

[To be written: With conjugate priors, $p(Y)$ is cheaper to compute than any of the alternative measures.]

6 Conclusions

[To be written.]

A Marginal Likelihood of a VAR

In this Appendix we derive expression (12), i.e. the marginal likelihood of VAR model (1) with prior (6). We include this derivation in the paper, because this derivation appears simpler than the derivations available in the literature.

To derive expression (12) we proceed in four steps. First, we rewrite the prior (6) including the normalizing constant. Second, we state and prove a useful lemma. Third, we derive the density $p(Y|\Sigma)$. Fourth, we derive the marginal likelihood $p(Y)$.

Step one. Consider expression (6). The prior density of B and Σ, including the normalizing constant, is given by

$$p(B,\Sigma) = p(B|\Sigma) \times p(\Sigma) \tag{20}$$
$$= (2\pi)^{-NK/2}|\Sigma|^{-K/2}|\tilde{X}'\tilde{X}|^{N/2} \exp\left(-\frac{1}{2}\operatorname{tr}(B-\tilde{B})'\tilde{X}'\tilde{X}(B-\tilde{B})\Sigma^{-1}\right)$$
$$\times C_{\mathcal{IW}}(\tilde{S},\tilde{\nu};N)|\Sigma|^{-(\tilde{\nu}+N+1)/2}\exp\left(-\frac{1}{2}\operatorname{tr}\tilde{S}\Sigma^{-1}\right).$$

The second line is the normal density of B conditional on Σ. The third line is the inverse Wishart density of Σ. The normalizing constant of the N-dimensional inverse Wishart density is given by

$$C_{\mathcal{IW}}(S,\nu;N) = \frac{2^{-\nu N/2}|S|^{\nu/2}}{\Gamma_N\left(\frac{\nu}{2}\right)}. \tag{21}$$

The term $\Gamma_N\left(\frac{\nu}{2}\right)$ is the multivariate Gamma function defined as

$$\Gamma_N\left(\frac{\nu}{2}\right) = \pi^{N(N-1)/4}\prod_{n=1}^{N}\Gamma\left(\frac{\nu+1-n}{2}\right). \tag{22}$$

Step two. The following lemma is useful.

Lemma 1 *Suppose that D is a scalar, X and B are $K\times N$ matrices, C and Σ are $N\times N$ matrices, A is a $K\times K$ matrix, and Σ and A are nonsingular. Then the following equality holds*

$$D\exp\left(-\frac{1}{2}\operatorname{tr}\left(X'AX-2X'B+C\right)\Sigma^{-1}\right)$$
$$= (2\pi)^{-KN/2}|A|^{N/2}|\Sigma|^{-K/2}\exp\left(-\frac{1}{2}\operatorname{vec}(X-A^{-1}B)'\left(\Sigma^{-1}\otimes A\right)\operatorname{vec}(X-A^{-1}B)\right)$$
$$\times(2\pi)^{KN/2}|A|^{-N/2}|\Sigma|^{K/2}D\exp\left(-\frac{1}{2}\operatorname{tr}(-B'A^{-1}B+C)\Sigma^{-1}\right).$$

The proof of the lemma follows from simple algebra:

$$D \exp\left(-\frac{1}{2}\,\mathrm{tr}\left(X'AX - 2X'B + C\right)\Sigma^{-1}\right)$$

$$= (2\pi)^{-KN/2}|A|^{N/2}|\Sigma|^{-K/2}(2\pi)^{KN/2}|A|^{-N/2}|\Sigma|^{K/2}D$$

$$\times \exp\left(-\frac{1}{2}\,\mathrm{tr}(X'AX - 2X'B + B'A^{-1}B - B'A^{-1}B + C)\Sigma^{-1}\right)$$

$$= (2\pi)^{-KN/2}|A|^{N/2}|\Sigma|^{-K/2}\exp\left(-\frac{1}{2}\,\mathrm{vec}(X - A^{-1}B)'\left(\Sigma^{-1}\otimes A\right)\mathrm{vec}(X - A^{-1}B)\right)$$

$$\times (2\pi)^{KN/2}|A|^{-N/2}|\Sigma|^{K/2}D\exp\left(-\frac{1}{2}\,\mathrm{tr}(-B'A^{-1}B + C)\Sigma^{-1}\right).$$

Step three. We derive the density $p(Y|\Sigma)$: (i) we use $p(Y|B,\Sigma)$ and $p(B|\Sigma)$ to derive $p(Y,B|\Sigma)$, and (ii) we integrate out B from $p(Y,B|\Sigma)$. We have:

$$p(Y,B|\Sigma) = p(Y|B,\Sigma)p(B|\Sigma)$$

$$= (2\pi)^{-TN/2}|\Sigma|^{-T/2}\exp\left(-\frac{1}{2}\,\mathrm{tr}(Y - XB)'(Y - XB)\Sigma^{-1}\right)$$

$$\times (2\pi)^{-NK/2}|\Sigma|^{-K/2}|\tilde{X}'\tilde{X}|^{N/2}\exp\left(-\frac{1}{2}\,\mathrm{tr}(B - \tilde{B})'\tilde{X}'\tilde{X}(B - \tilde{B})\Sigma^{-1}\right)$$

$$= (2\pi)^{-TN/2}|\Sigma|^{-T/2}(2\pi)^{-NK/2}|\Sigma|^{-K/2}|\tilde{X}'\tilde{X}|^{N/2}$$

$$\exp\left(-\frac{1}{2}\,\mathrm{tr}\left(B'(X'X + \tilde{X}'\tilde{X})B - 2B'(X'Y + \tilde{X}'\tilde{Y}) + Y'Y + \tilde{Y}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}\right)\Sigma^{-1}\right)$$

$$= (2\pi)^{-TN/2}|\Sigma|^{-T/2}|\tilde{X}'\tilde{X}|^{N/2}|\bar{X}'\bar{X}|^{-N/2}$$

$$\exp\left(-\frac{1}{2}\,\mathrm{tr}\left(-\bar{Y}'\bar{X}(\bar{X}'\bar{X})^{-1}\bar{X}'\bar{Y} + Y'Y + \tilde{Y}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}\right)\Sigma^{-1}\right)$$

$$\times \mathcal{N}(B;\bar{B},\Sigma\otimes(\bar{X}'\bar{X})^{-1}).$$

The first equality follows from the definition of the conditional density. The second equality from the definitions of the likelihood and the prior. After the third equality we rearrange terms. In the fourth equality we use $\bar{X}'\bar{X} = X'X + \tilde{X}'\tilde{X}$, $\bar{X}'\bar{Y} = X'Y + \tilde{X}'\tilde{Y}$, $\bar{B} \equiv (\bar{X}'\bar{X})^{-1}\bar{X}'\bar{Y}$, and we use Lemma 1. We can simplify the expression in the exponent as follows

$$-\bar{Y}'\bar{X}(\bar{X}'\bar{X})^{-1}\bar{X}'\bar{Y} + Y'Y + \tilde{Y}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}$$

$$= Y'Y + \tilde{Y}'\tilde{Y} - \bar{Y}'\bar{X}(\bar{X}'\bar{X})^{-1}\bar{X}'\bar{Y} - \tilde{Y}'\tilde{Y} + \tilde{Y}'\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} = \bar{S} - \tilde{S}.$$

Integrating out B from $p(Y, B|\Sigma)$ yields

$$p(Y|\Sigma) = \int p(Y, B|\Sigma) dB = (2\pi)^{-TN/2} |\Sigma|^{-T/2} \frac{|\tilde{X}'\tilde{X}|^{N/2}}{|\bar{X}'\bar{X}|^{N/2}} \exp\left(-\frac{1}{2} \operatorname{tr}\left(\bar{S} - \tilde{S}\right) \Sigma^{-1}\right) \quad (23)$$

This integral is computed by noting that the normal density of B integrates to one.

Step four. We derive the marginal likelihood $p(Y)$: (i) we use $p(Y|\Sigma)$ and $p(\Sigma)$ to derive $p(Y, \Sigma)$, and (ii) we integrate out Σ from $p(Y, \Sigma)$. We have:

$$\begin{aligned} p(Y, \Sigma) &= p(Y|\Sigma) p(\Sigma) \\ &= (2\pi)^{-TN/2} |\Sigma|^{-T/2} \frac{|\tilde{X}'\tilde{X}|^{N/2}}{|\bar{X}'\bar{X}|^{N/2}} \exp\left(-\frac{1}{2} \operatorname{tr}\left(\bar{S} - \tilde{S}\right) \Sigma^{-1}\right) \\ &\quad \times C_{\mathcal{IW}}(\tilde{S}, \tilde{\nu}; N) |\Sigma|^{-(\tilde{\nu}+N+1)/2} \exp\left(-\frac{1}{2} \operatorname{tr} \tilde{S}\Sigma^{-1}\right) \\ &= (2\pi)^{-TN/2} \frac{|\tilde{X}'\tilde{X}|^{N/2}}{|\bar{X}'\bar{X}|^{N/2}} \frac{C_{\mathcal{IW}}(\tilde{S}, \tilde{\nu}; N)}{C_{\mathcal{IW}}(\bar{S}, \tilde{\nu} + T; N)} C_{\mathcal{IW}}(\bar{S}, \tilde{\nu}+T; N) |\Sigma|^{-(\tilde{\nu}+T+N+1)/2} \exp\left(-\frac{1}{2} \operatorname{tr} \bar{S}\Sigma^{-1}\right). \end{aligned}$$

The first equality follows from the definition of the conditional density. The second equality uses (23) and (20). The third equality follows from simple algebra.

Integrating out Σ from $p(Y, \Sigma)$ yields

$$p(Y) = \int p(Y, \Sigma) d\Sigma = (2\pi)^{-TN/2} \frac{|\tilde{X}'\tilde{X}|^{N/2}}{|\bar{X}'\bar{X}|^{N/2}} \frac{C_{\mathcal{IW}}(\tilde{S}, \tilde{\nu}; N)}{C_{\mathcal{IW}}(\bar{S}, \tilde{\nu} + T; N)}$$

This integral is computed by noting that the inverse Wishart density integrates to one. Using the definition of the normalizing constant, (21), yields

$$\begin{aligned} p(Y) &= (2\pi)^{-TN/2} \frac{|\tilde{X}'\tilde{X}|^{N/2}}{|\bar{X}'\bar{X}|^{N/2}} \frac{2^{-\tilde{\nu}N/2} |\tilde{S}|^{\tilde{\nu}/2}}{\Gamma_N\left(\frac{\tilde{\nu}}{2}\right)} \frac{\Gamma_N\left(\frac{\tilde{\nu}+T}{2}\right)}{2^{-(\tilde{\nu}+T)N/2} |\bar{S}|^{(\tilde{\nu}+T)/2}} \\ &= \pi^{-TN/2} \frac{|\tilde{X}'\tilde{X}|^{N/2}}{|\bar{X}'\bar{X}|^{N/2}} \frac{\Gamma_N\left(\frac{\tilde{\nu}+T}{2}\right)}{\Gamma_N\left(\frac{\tilde{\nu}}{2}\right)} \frac{|\tilde{S}|^{\tilde{\nu}/2}}{|\bar{S}|^{(\tilde{\nu}+T)/2}}, \end{aligned}$$

which is expression (12).

B Initial Prior About B and Σ

This appendix gives the details concerning the initial prior, introduced in Section 4.1, about B and Σ. This initial prior consists of the modified Minnesota prior and the Sims dummy observations prior Sims and Zha (1998). We implement them following Bańbura et al. (2010). This prior is built from four components.

The first component is the modified Litterman prior. The modified Litterman prior is

$$p(\operatorname{vec} B|\Sigma) = \mathcal{N}\left(\operatorname{vec}\begin{pmatrix} I_N \\ 0_{K-N\times N}\end{pmatrix}, \Sigma \otimes WW'\right). \tag{24}$$

Recall that $K = NP + 1$ is the number of right-hand side variables in the VAR. $\mathcal{N}$ denotes the normal density. Furthermore, W is a diagonal matrix of size $K \times K$ such that the diagonal entry corresponding to variable n and lag p equals $\lambda_1/(\hat{\sigma}_n p^{\lambda_2})$. The terms λ_1, λ_2, and $\hat{\sigma}_n$ are hyperparameters. Let $\bar{P} = (1, ..., P)$ and $\hat{\sigma} = (\hat{\sigma}_1, ..., \hat{\sigma}_N)$. Then

$$W^{-1} = \operatorname{diag}\left(\lambda_1^{-1}\bar{P}^{\lambda_2} \otimes \hat{\sigma}, \lambda_3^{-1}\right),$$

where λ_3 is a hyperparameter associated with the constant term γ. We implement the modified Litterman prior with the dummy observations

$$Y_{Litterman} = W^{-1}\begin{pmatrix} I_N \\ 0_{K-N\times N}\end{pmatrix}, \quad X_{Litterman} = W^{-1}.$$

We set $\hat{\sigma}$ equal to sample standard deviations of residuals from univariate autoregressive models with P lags fit to the individual series in the sample.

The second component of the initial prior is the one-unit-root prior. The one-unit-root prior is implemented with the single dummy observation

$$Y_{one-unit-root} = \lambda_4 \bar{y}, \quad X_{one-unit-root} = \lambda_4(\bar{y}, ..., \bar{y}, 1),$$

where λ_4 and $\bar{y}$ are hyperparameters. We set $\bar{y} = (1/P)\sum_{t=0}^{P-1} y_{-t}$, the average of initial values of y. This follows Sims and Zha (1998).

The third component is the no-cointegration prior. The no-cointegration prior is implemented with the N dummy observations

$$Y_{no-cointegration} = \lambda_5 \operatorname{diag}(\bar{y}), \quad X_{no-cointegration} = \lambda_5(\operatorname{diag}(\bar{y}), ..., \operatorname{diag}(\bar{y}), 0),$$

where λ_5 is a hyperparameter.

The fourth component of the initial prior is an inverse Wishart prior about Σ with mean $\operatorname{diag}(\hat{\sigma}^2)$. This prior is

$$\begin{aligned} p(\Sigma) = \mathcal{IW}(ZZ', \nu_0) &\propto |\Sigma|^{-(\nu_0+N+1)/2} \exp\left(-\frac{1}{2}\operatorname{tr}\left(ZZ'\Sigma^{-1}\right)\right) \\ &= |\Sigma|^{-(\nu_0+1)/2}|\Sigma|^{-N/2} \exp\left(-\frac{1}{2}\operatorname{tr}\left(Z' - 0B\right)'\left(Z' - 0B\right)\Sigma^{-1}\right), \end{aligned}$$

where $\mathcal{IW}$ denotes the inverse Wishart density and $Z_{N\times N}$ and ν_0 are hyperparameters. This prior is proportional to a likelihood of N observations with $Y_\Sigma = Z'$ and $X_\Sigma = 0_{N\times K}$, multiplied by the factor $|\Sigma|^{-(\nu_0+1)/2}$. We set $Z = \sqrt{\nu_0 - N - 1}\,\mathrm{diag}(\hat{\sigma}_i)$ which implies that the mean of this prior is

$$E(\Sigma) = \frac{ZZ'}{\nu_0 - N - 1} = \mathrm{diag}(\hat{\sigma}_n^2).$$

We set $\nu_0 = K + N = N(P+1) + 1$. The reason for this choice for the value of ν_0 is as follows. The inverse Wishart density is proper when $\nu_0 > N - 1$. The whole Sims-Zha prior for B and Σ is proper when $\nu_0 > K + N - 1$, because K degrees of freedom are "used up" by the normal density of B. Therefore, as a rule of thumb we use the next integer value of ν_0.

Collecting all dummy observations introduced here yields

$$Y_{SZ} = \begin{pmatrix} Y_{Litterman} \\ Y_{one-unit-root} \\ Y_{no-cointegration} \\ Y_\Sigma \end{pmatrix}, \quad \text{and} \quad X_{SZ} = \begin{pmatrix} X_{Litterman} \\ X_{one-unit-root} \\ X_{no-cointegration} \\ X_\Sigma \end{pmatrix}.$$

The matrices Y_{SZ} and X_{SZ} appear in expression (6).

We use two versions of this prior. In the first version, we set $\lambda_1 = 0.4$, $\lambda_2 = 0.5$, $\lambda_3 = 2$, $\lambda_4 = 0.5$, and $\lambda_5 = 0.5$. In the second version, we set $\lambda_1 = 0.2$ and the other hyperparameters are the same as in the first version.

C MC^3

[To be written.]

D Computation of $p(Y_1|\omega)$

[To be written.]

References

Andersson, M. K. and Karlsson, S. (2007). Bayesian forecast combination for VAR models. Working Paper Series 216, Sveriges Riksbank (Central Bank of Sweden).

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92.

Bauwens, L., Lubrano, M., and Richard, J.-F. (1999). *Bayesian Inference in Dynamic Econometric Models*. Advanced Texts in Econometrics. Oxford University Press, Oxford, first edition.

Cushman, D. O. and Zha, T. (1997). Identifying monetary policy in a small open economy under flexible exchange rates. *Journal of Monetary Economics*, 39(3):433–448.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 42(1):204–223.

Eklund, J. and Karlsson, S. (2007). Forecast combination and model averaging using predictive measures. *Econometric Reviews*, 26(2-4):329–363.

George, E. I., Sun, D., and Ni, S. (2008). Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics*, 142(1):553–580.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*. Wiley Series in Probability and Statistics. John Wiley and Sons, Hoboken, New Jersey, first edition.

Geweke, J. and Amisano, G. (forthcoming 2011). Optimal prediction pools. *Journal of Econometrics*.

Giannone, D., Lenza, M., and Primiceri, G. (2010). Prior selection for vector autoregressions. mimeo, Université libre de Bruxelles.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Jochmann, M., Koop, G., and Strachan, R. W. (2010). Bayesian forecasting using stochastic search variable selection in a var subject to breaks. *International Journal of Forecasting*, 26(2):326 – 347. Special Issue: Bayesian Forecasting in Economics.

Korobilis, D. (2010). VAR forecasting using Bayesian variable selection. Working Paper Series 51-10, Rimini Centre for Economic Analysis.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review / Revue Internationale de Statistique*, 63(2):pp. 215–232.

OEIS (2011). On-line encyclopedia of integer sequences. Sequence A000670, The OEIS Foundation. published electronically at http://oeis.org.

Sims, C. (2010). Lecture notes: causal orderings and exogeneity. Lecture notes, Princeton University. available at http://sims.princeton.ftp.edu/yftp/Times09/AddedNotes/GCP09.pdf.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Sims, C. A. and Zha, T. (1998). Bayesian methods for dynamic multivariate models. *International Economic Review*, 39(4):949–68.

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618.

Zha, T. (1999). Block recursion and structural vector autoregressions. *Journal of Econometrics*, 90(2):291–316.