Karlsson, Sune

Working Paper

Forecasting with Bayesian Vector Autoregressions

Working Paper, No. 12/2012

Provided in Cooperation with:
Örebro University School of Business

Suggested Citation: Karlsson, Sune (2012) : Forecasting with Bayesian Vector Autoregressions, Working Paper, No. 12/2012, Örebro University School of Business, Örebro

This Version is available at:
https://hdl.handle.net/10419/244486

WORKING PAPER

12/2012

Forecasting with Bayesian Vector Autoregressions

Sune Karlsson
Statistics

ISSN 1403-0586

http://www.oru.se/Institutioner/Handelshogskolan-vid-Orebro-universitet/Forskning/Publikationer/Working-papers/
Örebro University School of Business
701 82 Örebro
SWEDEN

Forecasting with Bayesian Vector Autoregressions

Sune Karlsson*
Department of Statistics, Örebro University Business School

August 4, 2012

Abstract

Prepared for the Handbook of Economic Forecasting, vol 2

This chapter reviews Bayesian methods for inference and forecasting with VAR models. Bayesian inference and, by extension, forecasting depends on numerical methods for simulating from the posterior distribution of the parameters and special attention is given to the implementation of the simulation algorithm.

JEL-codes: C11, C32, C53
Keywords: Markov chain Monte Carlo; Structural VAR; Cointegration; Conditional forecasts; Time-varying parameters; Stochastic volatility; Model selection; Large VAR

*Sune.Karlsson@oru.se

Contents

1 Introduction

Vector autoregressions (VARs) have become the workhorse model for macroeconomic forecasting. The initial use in economics was to a large degree motived by Sims (1980) critique of the "incredible restrictions" used by the large macroeconometric models developed in the 1970s and much effort was put into tools for policy analysis based on VAR models. This role of the VAR model has to some degree been taken over by the current crop of DSGE models, a new generation of theory based models which are – at times – ill at ease with the data. The role of the VAR model as the baseline, serious, model for economic forecasting is, however, unchallenged. The popularity stems in part from it's relative simplicity, flexibility and ability to fit the data but, of course, also from it's success as a forecasting device.

The flexibility and ability to fit the data comes from the rich parameterization of VAR models brings with it a risk of overfitting the data, of imprecise inference and large uncertainty about the future paths projected by the model. This is essentially the frequentist argument for Bayesian VAR models and one reason why Bayesian VAR models forecast better than VARs estimated with frequentist techniques. The widely used Minnesota prior introduced by Litterman (1979) is a set of data centric prior beliefs that shrinks the parameters towards a stylized representation of macroeconomic data thereby reducing parameter uncertainty and improving forecast accuracy. The Bayesian argument is different. The Minnesota prior captures widely held beliefs about the long run properties of the data, properties that are not readily apparent in the short samples typically used for estimation. Bayes theorem then provides the optimal way of combining these two sources of information leading to sharper inference and more precise forecasts. The development of efficient numerical techniques for evaluating posterior distributions is also a contributing factor to the attractiveness of Bayesian methods. It is now possible to tackle more complex problems under realistic assumptions when we no longer are limited to problem formulations that lead to analytical solutions.

This chapter surveys Bayesian approaches to inference in VAR models with a focus on forecasting. One important feature of the chapter is that it gathers many algorithms for simulating from the posterior distribution of the parameters, some of which have not been clearly stated previously. This provides the necessary tools for analyzing the posterior and predictive distributions and forecast with the models and priors that are studied in the chapter. Koop and Korobilis (2009) and DelNegro and Schorfheide (2011) provides complementary reviews of Bayesian VAR models, Koop and Korobilis (2009) with a focus on models that allows for time-varying parameters and stochastic volatility while DelNegro and Schorfheide (2011) has the broader remit of Bayesian macroeconometrics.

Section 2 lays the foundations by placing the task of forecasting in a Bayesian context and reviews modern simulation techniques for exploring posterior and predictive distributions. Section 3 provides the basic building blocks for forecasting with Bayesian VAR models by introducing the Minnesota prior beliefs in the context of reduced form VAR models and reviews families of prior distributions that have been found useful for expressing the prior beliefs. The more general issue of model specification is also discussed and one important message that emerges is that, in line with a general conclusion in the forecasting literature, simple methods works quite well.

The remaining sections can largely be read independently. Section 4 reviews Bayesian

analysis of a VAR in structural form (SVAR) and section 5 studies the vector error correction (VECM) form of a VAR model. Both SVAR and VECM models have the potential to improve forecast performance if the hard restrictions they impose on the model are at least aproximately correct but they have seen relatively little use in forecasting applications, in particular in their Bayesian flavor. This is partly because it is only recently that satisfactory procedures for posterior inference in these models have become available.

Section 6 consider forecasts conditional on future events. This can be a useful tool for incorporating judgement and other late breaking information that is (perhaps due to the slow release of data) not in the information set used by the model. In a policy setting conditional forecasts are useful for what-if analysis and for producing forecasts that are consistent with the current policy.

Section 7 relaxes the constant parameter assumption and shows how to allow for time-varying parameters and stochastic volatility in Bayesian VAR models. There are encouraging studies that indicate that both time-varying parameters and stochastic volatility can improve the forecast performance but both can also lead to a dramatic increase in the number of parameters in a model. There is consequently a greater risk of overfitting the data. The methods for model and variable selection discussed in Section 8 can then be useful in addition to the Bayesian shrinkage that is routinely applied through the prior. Section 8 provides tools both for selecting the variables to include as left hand side variables in a VAR model and for reducing the number of parameters by effectively excluding some variables and lags from the right hand side. This touches on the issue of model averaging and forecast combination which is not discussed here in spite of this being a natural extension of the Bayesian framework for treating parameter uncertainty. The reader is instead referred to Geweke and Whiteman (2006) and Timmermann (2006).

The final section 9 considers the task of forecasting in a data rich environment where several hundred potential predictors may be available. Recent work shows that Bayesian VARs can be competitive in this setting as well and important recent developments are reviewed.

2 Bayesian Forecasting and Computation

This section provides a brief overview of the underlying principles of Bayesian inference and forecasting. See Geweke and Whiteman (2006) for a more complete discussion and, for example, Gelman, Carlin, Stern and Rubin (2003), Geweke (2005) or Koop (2003) for a text book treatment of Bayesian inference.

2.1 Bayesian Forecasting and Inference

The fundamental object in Bayesian forecasting is the (posterior) predictive distribution, the distribution $p\left(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T\right)$ of future datapoints, $\mathbf{y}_{T+1:T+H} = \left(\mathbf{y}'_{T+1}, \ldots, \mathbf{y}'_{T+H}\right)'$ conditional on the currently observed data, $\mathbf{Y}_T = \{\mathbf{y}_t\}_{t=1}^T$. By itself the predictive distribution captures all relevant information about the unknown future events. It is then up to the forecaster or user of the forecast which features of the predictive distribution are relevant for the situation at hand and should be reported as *the* forecast. This could,

for example, be the mean, mode or median of the predictive distribution together with a probability interval indicating the range of likely outcomes.

Formally this is a decision problem which requires the specification of a problem dependent loss function, $\mathcal{L}\left(\mathbf{a}, \mathbf{y}_{T+1:T+H}\right)$, where $\mathbf{a}$ is the action taken, the vector of real numbers to report as the forecast, and $\mathbf{y}_{T+1:T+H}$ represents the unknown future state of nature. The Bayesian decision is to choose the action (forecast) that minimizes the expected loss conditional on the available information $\mathbf{Y}_T$,

$$E\left[\mathcal{L}\left(\mathbf{a}, \mathbf{y}_{T+1:T+H}\right)|\mathbf{Y}_T\right] = \int \mathcal{L}\left(\mathbf{a}, \mathbf{y}_{T+1:T+H}\right) p\left(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T\right) d\mathbf{y}_{T+1:T+H}.$$

For a given loss function and predictive distribution, $p\left(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T\right)$ the solution to the minimization problem is a function of the data, $\mathbf{a}\left(\mathbf{Y}_T\right)$. For specific loss functions $\mathbf{a}\left(\mathbf{Y}_T\right)$ takes on simple forms. With quadratic loss function, $\left(\mathbf{a} - \mathbf{y}_{T+1:T+H}\right)' \left(\mathbf{a} - \mathbf{y}_{T+1:T+H}\right)$, the solution is the conditional expectation, $\mathbf{a}\left(\mathbf{Y}_T\right) = E\left(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T\right)$, and with an absolute value loss function, $\sum |a_i - \omega_i|$, the conditional mode.

It remains to specify the form of the predictive distribution. This requires the specification of three different distributions that completes the description of the problem, the distribution of the future observations conditional on unknown parameter values, $\boldsymbol{\theta}$, and the observed data, $p\left(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T, \boldsymbol{\theta}\right)$, the distribution of the observed data – that is, the model or likelihood – conditional on the parameters, $L\left(\mathbf{Y}_T|\boldsymbol{\theta}\right)$, and the prior distribution, $\pi\left(\boldsymbol{\theta}\right)$, representing our prior notions about likely or "reasonable" values of the unknown parameters, $\boldsymbol{\theta}$. In a time series and forecasting context, the likelihood $L\left(\mathbf{Y}_T|\boldsymbol{\theta}\right)$ usually takes the form

$$L\left(\mathbf{Y}_T|\boldsymbol{\theta}\right) = \prod_{t=1}^{T} f\left(\mathbf{y}_t|\mathbf{Y}_{t-1}, \boldsymbol{\theta}\right)$$

with the history in $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ suitably extended to include initial observations that the likelihood is conditional on.[1] The distribution of future observations is of the same form,

$$f\left(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T, \boldsymbol{\theta}\right) = \prod_{t=T+1}^{T+H} f\left(\mathbf{y}_t|\mathbf{Y}_{t-1}, \boldsymbol{\theta}\right).$$

With these in hand straightforward application of Bayes Rule yields the predictive distribution as

$$p\left(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T\right) = \frac{p\left(\mathbf{y}_{T+1:T+H}, \mathbf{Y}_T\right)}{m\left(\mathbf{Y}_T\right)} = \frac{\int f\left(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T, \boldsymbol{\theta}\right) L\left(\mathbf{Y}_T|\boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}}{\int L\left(\mathbf{Y}_T|\boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}}. \quad (1)$$

In practice an intermediate step through the posterior distribution of the parameters,

$$p\left(\boldsymbol{\theta}|\mathbf{Y}_T\right) = \frac{L\left(\mathbf{Y}_T|\boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\right)}{\int L\left(\mathbf{Y}_T|\boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}} \propto L\left(\mathbf{Y}_T|\boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\right), \quad (2)$$

is used with the predictive distribution given by

$$p\left(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T\right) = \int f\left(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T, \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}|\mathbf{Y}_T\right) d\boldsymbol{\theta}. \quad (3)$$

[1] It is, of course, in many cases also possible to complete the likelihood with the marginal distribution for the first, say p, observations.

Note that the latter form of the predictive distribution makes it clear how Bayesian forecasts accounts for both the inherent uncertainty about the future embodied by $f\left(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T,\boldsymbol{\theta}\right)$ and the uncertainty about the true parameter values described by the posterior distribution $p\left(\boldsymbol{\theta}|\mathbf{Y}_T\right)$.

While the posterior distribution of the parameters may be available in closed form in special cases when conjugate prior distributions are used closed form expressions for the predictive distribution are generally unavailable when lead times greater than 1 are considered. This makes the form (3) of the predictive distribution especially attractive. Marginalizing out the parameters of the joint distribution of $\mathbf{y}_{T+1:T+H}$ and $\boldsymbol{\theta}$ analytically may be difficult or impossible, on the other hand (3) suggests a straightforward simulation scheme for the marginalization. Supposing that we can generate random numbers from the posterior $p\left(\boldsymbol{\theta}|\mathbf{Y}_T\right)$, for each draw of $\boldsymbol{\theta}$ generate a sequence of draws of $\mathbf{y}_{T+1},\ldots,\mathbf{y}_{T+H}$ by repeatedly drawing from $f\left(\mathbf{y}_t|\mathbf{Y}_{t-1},\boldsymbol{\theta}\right)$ and adding the draw of $\mathbf{y}_t$ to the conditioning set for the distribution of $\mathbf{y}_{t+1}$. This gives a draw from the joint distribution of $(\boldsymbol{\theta},\mathbf{y}_{T+1},\ldots,\mathbf{y}_{T+H})$ conditional on $\mathbf{Y}_T$ and marginalization is achieved by simply discarding the draw of $\boldsymbol{\theta}$. Repeating this R times gives a sample from the predictive distribution that can be used to estimate $E\left(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T\right)$ or any other function or feature $\mathbf{a}\left(\mathbf{Y}_T\right)$ of the predictive distribution of interest.

The denominator in (1) and (2),

$$m\left(\mathbf{Y}_T\right)=\int L\left(\mathbf{Y}_T|\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)d\boldsymbol{\theta}, \tag{4}$$

is known as the *marginal likelihood* or *prior predictive distribution* and plays a crucial role in Bayesian hypothesis testing and model selection. Consider two alternative models, $\mathcal{M}_1$ and $\mathcal{M}_2$, with corresponding likelihoods $L\left(\mathbf{Y}_T|\boldsymbol{\theta}_1,\mathcal{M}_1\right)$, $L\left(\mathbf{Y}_T|\boldsymbol{\theta}_2,\mathcal{M}_2\right)$ and priors $\pi\left(\boldsymbol{\theta}_1|\mathcal{M}_1\right)$, $\pi\left(\boldsymbol{\theta}_2|\mathcal{M}_2\right)$. Supposing that one of $\mathcal{M}_1$ and $\mathcal{M}_2$ is the true model but that we are not certain which of the competing hypothesis or theories embodied in the models is the correct one we can assign prior probabilities, $\pi\left(\mathcal{M}_1\right)$ and $\pi\left(\mathcal{M}_2\right)=1-\pi\left(\mathcal{M}_1\right)$, that each of the models is the correct one. With these in hand Bayes Rule yields the posterior probabilities that the models are correct as

$$p\left(\mathcal{M}_i\right)=\frac{m\left(\mathbf{Y}_T|\mathcal{M}_i\right)\pi\left(\mathcal{M}_i\right)}{m\left(\mathbf{Y}_T|\mathcal{M}_1\right)\pi\left(\mathcal{M}_1\right)+m\left(\mathbf{Y}_T|\mathcal{M}_2\right)\pi\left(\mathcal{M}_2\right)} \tag{5}$$

with $m\left(\mathbf{Y}_T|\mathcal{M}_i\right)=\int L\left(\mathbf{Y}_T|\boldsymbol{\theta}_i,\mathcal{M}_i\right)\pi\left(\boldsymbol{\theta}_i|\mathcal{M}_i\right)d\boldsymbol{\theta}_i$. The posterior odds for model 1 against model 2 is given by

$$\frac{p\left(\mathcal{M}_1\right)}{p\left(\mathcal{M}_2\right)}=\frac{m\left(\mathbf{Y}_T|\mathcal{M}_1\right)\pi\left(\mathcal{M}_1\right)}{m\left(\mathbf{Y}_T|\mathcal{M}_2\right)\pi\left(\mathcal{M}_2\right)}=\frac{m\left(\mathbf{Y}_T|\mathcal{M}_1\right)}{m\left(\mathbf{Y}_T|\mathcal{M}_2\right)}\times\frac{\pi\left(\mathcal{M}_1\right)}{\pi\left(\mathcal{M}_2\right)}$$

the Bayes factor $BF_{1,2}=m\left(\mathbf{Y}_T|\mathcal{M}_1\right)/m\left(\mathbf{Y}_T|\mathcal{M}_2\right)$ comparing $\mathcal{M}_1$ to $\mathcal{M}_2$ times the prior odds. Model choice can be based on the posterior odds but it is also common to use the Bayes factors directly, implying equal prior probabilities. The Bayes factor captures the data evidence and can be interpreted as measuring how much our opinion about the models have changed after observing the data. The choice of model should of course take account of the losses associated with making the wrong choice. Alternatively, we can avoid conditioning on one single model being the correct one by averaging over the models

with the posterior model probabilities as weight. That is, instead of basing our forecasts on the predictive distribution $p(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T, \mathcal{M}_1)$ and conditioning on $\mathcal{M}_1$ being the correct model we conduct *Bayesian Model Averaging* (BMA) to obtain the marginalized (with respect to the models) predictive distribution

$$p(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T) = p(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T, \mathcal{M}_1)\pi(\mathcal{M}_1) + p(\mathbf{y}_{T+1:T+H}|\mathbf{Y}_T, \mathcal{M}_2)\pi(\mathcal{M}_2)$$

which accounts for both model and parameter uncertainty.

The calculations involved in (5) are non-trivial and the integral $\int L(\mathbf{Y}_T|\boldsymbol{\theta}_i, \mathcal{M}_i)\pi(\boldsymbol{\theta}_i|\mathcal{M}_i)\, d\boldsymbol{\theta}_i$ is only well defined if the prior is *proper*. That is, if $\int \pi(\boldsymbol{\theta}_i|\mathcal{M}_i)\, d\boldsymbol{\theta}_i = 1$. For *improper* priors, such as a uniform prior on the whole real line, the integral is not convergent and the scale is arbitrary. For the uniform prior we can write $\pi(\boldsymbol{\theta}_i|\mathcal{M}_i) = k_i$ and it follows that $m(\mathbf{Y}_T) \propto k_i$ and the Bayes factors and posterior probabilities are arbitrary.[2] There is, however, one circumstance where improper prior can be used. This is when there are parameters that are common to all models, for example an error variance. We can then partition $\boldsymbol{\theta}_i = \left(\widetilde{\boldsymbol{\theta}}_i, \sigma^2\right)$ and use proper priors for $\widetilde{\boldsymbol{\theta}}_i$ and an improper prior, such as $\pi(\sigma^2) \propto 1/\sigma^2$, for the variance since the common scale factor cancels in the calculation of posterior model probabilities and Bayes factors.

2.1.1 Vector Autoregressions

To illustrate the concepts we consider the VAR model with m variables

$$\begin{aligned}\mathbf{y}_t' &= \sum_{i=1}^{p} \mathbf{y}_{t-i}'\mathbf{A}_i + \mathbf{x}_t'\mathbf{C} + \mathbf{u}_t' \\ &= \mathbf{z}_t'\boldsymbol{\Gamma} + \mathbf{u}_t'\end{aligned} \tag{6}$$

with $\mathbf{x}_t$ a vector of d deterministic variables, $\mathbf{z}_t' = \left(\mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-p}', \mathbf{x}_t'\right)$ a $k = mp + d$ dimensional vector and $\boldsymbol{\Gamma} = \left(\mathbf{A}_1', \ldots \mathbf{A}_p', \mathbf{C}'\right)'$ a $k \times m$ matrix and normally distributed errors, $\mathbf{u}_t \sim N(0, \boldsymbol{\Psi})$. That is, $f(\mathbf{y}_t|\mathbf{Y}_{t-1}, \boldsymbol{\theta}) = N(\mathbf{y}_t; \mathbf{z}_t'\boldsymbol{\Gamma}, \boldsymbol{\Psi})$. For simplicity we take the prior to be uninformative (diffuse), a uniform distribution for $\boldsymbol{\Gamma}$ and a Jeffreys' prior for $\boldsymbol{\Psi}$,[3]

$$\pi(\boldsymbol{\Gamma}, \boldsymbol{\Psi}) \propto |\boldsymbol{\Psi}|^{-(m+1)/2}.$$

Using (2) we see that the joint posterior distribution of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$ is proportional to the likelihood function times the prior. Stacking the data in the usual way we can write the model as

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{U}$$

[2]Note that this does not affect the posterior distribution as long as the integral $\int L(\mathbf{Y}_T|\boldsymbol{\theta}_i, M_i)\pi(\boldsymbol{\theta}_i|M_i)\, d\boldsymbol{\theta}_i$ is convergent since the arbitrary scale factor cancels in (2).

[3]This is an improper prior and the use of improper prior distributions is not always advisable as this can lead to improper posterior distributions. In the normal regression model with this prior the posterior will be proper if the matrix of explanatory variables has full column rank, i.e. when the OLS estimate is unique.

and the likelihood as

$$
\begin{aligned}
L\left(\mathbf{Y}|\mathbf{\Gamma},\mathbf{\Psi}\right) &= \left(2\pi\right)^{-mT/2}\left|\mathbf{\Psi}\right|^{-T/2}\exp\left\{-\frac{1}{2}\sum\left(\mathbf{y}_t'-\mathbf{z}_t'\mathbf{\Gamma}\right)\mathbf{\Psi}^{-1}\left(\mathbf{y}_t'-\mathbf{z}_t'\mathbf{\Gamma}\right)'\right\} \\
&= \left(2\pi\right)^{-mT/2}\left|\mathbf{\Psi}\right|^{-T/2}\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(\mathbf{Y}-\mathbf{Z}\mathbf{\Gamma}\right)\mathbf{\Psi}^{-1}\left(\mathbf{Y}-\mathbf{Z}\mathbf{\Gamma}\right)'\right]\right\} \\
&= \left(2\pi\right)^{-mT/2}\left|\mathbf{\Psi}\right|^{-T/2}\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\mathbf{\Psi}^{-1}\left(\mathbf{Y}-\mathbf{Z}\mathbf{\Gamma}\right)'\left(\mathbf{Y}-\mathbf{Z}\mathbf{\Gamma}\right)\right]\right\}
\end{aligned}
\tag{7}
$$

where $\mathbf{Y}$ and $\mathbf{U}$ are $T\times m$ matrices and $\mathbf{Z}$ is $T\times k$. Adding and subtracting $\mathbf{Z}\widehat{\mathbf{\Gamma}}$ for $\widehat{\mathbf{\Gamma}}=\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y}$, the OLS estimate, and multiplying with the prior we have the joint posterior as

$$
\begin{aligned}
p\left(\mathbf{\Gamma},\mathbf{\Psi}|\mathbf{Y}_T\right) \propto \left|\mathbf{\Psi}\right|^{-T/2}&\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\mathbf{\Psi}^{-1}\left(\mathbf{Y}-\mathbf{Z}\widehat{\mathbf{\Gamma}}\right)'\left(\mathbf{Y}-\mathbf{Z}\widehat{\mathbf{\Gamma}}\right)\right]\right\} \\
&\times\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\mathbf{\Psi}^{-1}\left(\mathbf{\Gamma}-\widehat{\mathbf{\Gamma}}\right)'\mathbf{Z}'\mathbf{Z}\left(\mathbf{\Gamma}-\widehat{\mathbf{\Gamma}}\right)\right]\right\}\left|\mathbf{\Psi}\right|^{-(m+1)/2}.
\end{aligned}
$$

Focusing on the part involving $\mathbf{\Gamma}$ and noting that

$$
\operatorname{tr}\left[\mathbf{\Psi}^{-1}\left(\mathbf{\Gamma}-\widehat{\mathbf{\Gamma}}\right)'\mathbf{Z}'\mathbf{Z}\left(\mathbf{\Gamma}-\widehat{\mathbf{\Gamma}}\right)\right]=\left(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}}\right)'\left(\mathbf{\Psi}^{-1}\otimes\mathbf{Z}'\mathbf{Z}\right)\left(\boldsymbol{\gamma}-\widehat{\boldsymbol{\gamma}}\right) \tag{8}
$$

for $\boldsymbol{\gamma}=\operatorname{vec}\left(\mathbf{\Gamma}\right)$ and $\widehat{\boldsymbol{\gamma}}=\operatorname{vec}\left(\widehat{\mathbf{\Gamma}}\right)=\left[\mathbf{I}_m\otimes\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\right]\mathbf{y}$[4] we recognize this as a the kernel of a multivariate normal distribution conditional on $\mathbf{\Psi}$ with mean $\widehat{\boldsymbol{\gamma}}$ and variance-covariance matrix $\mathbf{\Psi}\otimes\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}$,

$$
\boldsymbol{\gamma}|\mathbf{Y}_T,\mathbf{\Psi}\sim N\left(\widehat{\boldsymbol{\gamma}},\mathbf{\Psi}\otimes\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\right).
$$

With the special Kronecker structure of the variance-covariance matrix this is a matric-variate normal[5] distribution for $\mathbf{\Gamma}$ and we can also write the conditional posterior as $\mathbf{\Gamma}|\mathbf{Y}_T,\mathbf{\Psi}\sim MN_{km}\left(\widehat{\mathbf{\Gamma}},\mathbf{\Psi},\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\right)$. Integrating out $\boldsymbol{\gamma}$ from the joint posterior is trivial using the properties of the normal distribution and we have the marginal posterior distribution for $\mathbf{\Psi}$ as

$$
p\left(\mathbf{\Psi}|\mathbf{Y}_T\right)\propto\left|\mathbf{\Psi}\right|^{-(T+m+1-k)/2}\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\mathbf{\Psi}^{-1}\mathbf{S}\right]\right\}
$$

with $\mathbf{S}=\left(\mathbf{Y}-\mathbf{Z}\widehat{\mathbf{\Gamma}}\right)'\left(\mathbf{Y}-\mathbf{Z}\widehat{\mathbf{\Gamma}}\right)$. This can be recognized as the kernel of an inverse Wishart distribution with $T-k$ degrees of freedom,

$$
\mathbf{\Psi}|\mathbf{Y}_T\sim iW_m\left(\mathbf{S},T-k\right).
$$

[4] $\widehat{\boldsymbol{\gamma}}$ is the GLS estimate in the univariate regression model for $\mathbf{y}=\operatorname{vec}\left(\mathbf{Y}\right)=\left(\mathbf{I}_m\otimes\mathbf{Z}\right)\boldsymbol{\gamma}+\mathbf{u}$ with $V\left(\mathbf{u}\right)=\mathbf{\Psi}\otimes\mathbf{I}_T$. That is $\widehat{\boldsymbol{\gamma}}=\left[\left(\mathbf{I}_m\otimes\mathbf{Z}\right)'\left(\mathbf{\Psi}\otimes\mathbf{I}_T\right)^{-1}\left(\mathbf{I}_m\otimes\mathbf{Z}\right)\right]^{-1}\left(\mathbf{I}_m\otimes\mathbf{Z}\right)'\left(\mathbf{\Psi}\otimes\mathbf{I}_T\right)^{-1}\mathbf{y}=\left[\left(\mathbf{\Psi}^{-1}\otimes\mathbf{Z}'\mathbf{Z}\right)\right]^{-1}\left(\mathbf{\Psi}^{-1}\otimes\mathbf{Z}'\right)\mathbf{y}=\left[\mathbf{\Psi}\otimes\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\right]\left(\mathbf{\Psi}^{-1}\otimes\mathbf{Z}'\right)\mathbf{y}=\left[\mathbf{I}_m\otimes\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\right]\mathbf{y}$.

[5] See Appendix C for a review of some multivariate distributions.

We refer to the joint posterior of $\mathbf{\Gamma}$ and $\mathbf{\Psi}$ as a normal-Wishart distribution.

Alternatively, we can integrate out $\mathbf{\Psi}$ of the joint posterior. With the Kronecker variance matrix of the conditional normal distribution this yields a matricvariate t-distribution with $T-k$ degrees of freedom as the marginal posterior for $\mathbf{\Gamma}$,

$$\mathbf{\Gamma}|\mathbf{Y}_T \sim Mt_{km}\left(\widehat{\boldsymbol{\gamma}}, \mathbf{Z}'\mathbf{Z}, \mathbf{S}, T-k\right). \tag{9}$$

This is the natural generalization of the scalar variance case where $\mathbf{x}|\sigma \sim N\left(\mu, \sigma^2\mathbf{V}\right)$ with σ^{-2} Gamma distributed with shape parameter $v/2$ and scale parameter $1/2$ (or χ^2 with v degrees of freedom) yields a marginal t-distribution for $\mathbf{x}$ with v degrees of freedom.

For later reference note that the product of the prior and likelihood (7) has the form of an inverse Wishart distribution for $\mathbf{\Psi}$ conditional on $\mathbf{\Gamma}$,

$$p\left(\mathbf{\Psi}|\mathbf{Y}_T, \mathbf{\Gamma}\right) \propto |\mathbf{\Psi}|^{-(T+m+1)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\mathbf{\Psi}^{-1}\left(\mathbf{Y}-\mathbf{Z}\mathbf{\Gamma}\right)'\left(\mathbf{Y}-\mathbf{Z}\mathbf{\Gamma}\right)\right]\right\}$$
$$\mathbf{\Psi}|\mathbf{Y}_T, \mathbf{\Gamma} \sim iW\left(\left(\mathbf{Y}-\mathbf{Z}\mathbf{\Gamma}\right)'\left(\mathbf{Y}-\mathbf{Z}\mathbf{\Gamma}\right), T\right).$$

Turning to the forecasts, recursive substitution in (6) with $p=2$ yields

$$\mathbf{y}'_{T+1} = \mathbf{y}'_T\mathbf{A}_1 + \mathbf{y}'_{T-1}\mathbf{A}_2 + \mathbf{x}'_{T+1}\mathbf{C} + \mathbf{u}'_{T+1}$$
$$\mathbf{y}'_{T+2} = \mathbf{y}'_T\left(\mathbf{A}_1^2 + \mathbf{A}_2\right) + \mathbf{y}'_{T-1}\mathbf{A}_2\mathbf{A}_1 + \mathbf{x}'_{T+2}\mathbf{C} + \mathbf{x}'_{T+1}\mathbf{C}\mathbf{A}_1 + \mathbf{u}'_{T+2} + \mathbf{u}'_{T+1}\mathbf{A}_1$$

etc. The one-step ahead predictive distribution for $\mathbf{y}'_{T+1}$ can be shown to be matricvariate t, $Mt_{1m}\left(\mathbf{z}'_{T+1}\widehat{\mathbf{\Gamma}}, \left(1+\mathbf{z}'_{T+1}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{z}_{T+1}\right)^{-1}, \mathbf{S}, T-k\right)$. For higher lead times we have increasingly non-linear functions of the parameters and no closed form expressions for the predictive distribution are available. Instead the simulation scheme for generating a sample from the predictive distribution described above can be used. Simulating from the posterior and predictive distributions is particularly straightforward in this case and the procedure for simulating from the predictive distribution is given as Algorithm 1.

2.2 Bayesian Computation and Simulation

Having a simulated sample, $\widetilde{\mathbf{y}}^{(j)}_{T+1}, \ldots \widetilde{\mathbf{y}}^{(j)}_{T+H}$, of size R from the predictive distribution in hand it is straightforward to estimate features, such as probability intervals, expectations, etc., of the predictive distribution that we wish to report. An estimate of the minimum mean square error (MSE) h period ahead forecast, $\mathbf{y}_T(h) = E\left(\mathbf{y}_{T+h}|\mathbf{Y}_T\right)$, is given by the simple average of the simulated forecasts,

$$\widehat{\mathbf{y}}_T(h) = \frac{1}{R}\sum_{j=1}^{R}\widetilde{\mathbf{y}}^{(j)}_{T+h}. \tag{11}$$

With direct sampling and hence iid draws, as in the previous section, this is guaranteed to be a consistent and asymptotically normal estimator if $V\left(\mathbf{y}_{T+h}|\mathbf{Y}_T\right)$ exists,

$$\sqrt{R}\left(\widehat{\mathbf{y}}_T(h) - \mathbf{y}_T(h)\right) \xrightarrow{d} N\left(0, V\left(\mathbf{y}_{T+h}|\mathbf{Y}_T\right)\right).$$

Asymptotically motivated error bounds are thus readily available as $1-\alpha$ confidence intervals. Analogous results apply to any function of $\mathbf{y}_{T+h}$ with finite second moment.

Algorithm 1 Simulating the predictive distribution with a normal-Wishart posterior

For $j = 1, \ldots R$

1. Generate $\mathbf{\Psi}^{(j)}$ from the marginal posterior $\mathbf{\Psi}|\mathbf{Y}_T \sim iW_m(\mathbf{S}, T-k)$ distribution using, e.g. the algorithm of Geweke (1988).

2. Generate $\mathbf{\Gamma}^{(j)}$ from the conditional posterior $\mathbf{\Gamma}|\mathbf{Y}_T, \mathbf{\Psi}^{(j)} \sim N\left(\widehat{\mathbf{\Gamma}}, \mathbf{\Psi}^{(j)}, (\mathbf{Z}'\mathbf{Z})^{-1}\right)$

3. Generate $\mathbf{u}_{T+1}^{(j)}, \ldots, \mathbf{u}_{T+H}^{(j)}$ from $\mathbf{u}_t \sim N\left(0, \mathbf{\Psi}^{(j)}\right)$ and calculate recursively

$$\widetilde{\mathbf{y}}_{T+h}^{(j)\prime} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}_{T+h-i}^{(j)\prime} \mathbf{A}_i^{(j)} + \sum_{i=h}^{p} \mathbf{y}_{T+h-i}^{\prime} \mathbf{A}_i^{(j)} + \mathbf{x}_{T+h}^{\prime} \mathbf{C}^{(j)} + \mathbf{u}_{T+h}^{(j)\prime}. \tag{10}$$

Discarding the parameters yields $\left\{\widetilde{\mathbf{y}}_{T+1}^{(j)}, \ldots \widetilde{\mathbf{y}}_{T+H}^{(j)}\right\}_{j=1}^{R}$ as a sample of independent draws from the joint predictive distribution.

Note that it only a factor $\mathbf{P}$ of $\mathbf{\Psi}^{(j)} = \mathbf{P}\mathbf{P}'$ is needed for step 3 and steps 1 and 2 can be replaced by Algorithm 22 which avoids the explicit computation of $\mathbf{\Psi}^{(j)}$.

Estimates of prediction intervals are readily obtained by selecting the appropriate quantiles from the simulated predictive distribution. Let $\widetilde{\mathbf{y}}_{T+h,(i)}$ denote the i^{th} order statistic, a $1-\alpha$ prediction interval is then give by $\left(\widetilde{\mathbf{y}}_{T+h,(l)}, \widetilde{\mathbf{y}}_{T+h,(u)}\right)$ for $l = \lfloor R\alpha/2 \rfloor$ and $u = \lfloor R(1-\alpha/2) \rfloor$ where $\lfloor \cdot \rfloor$ denotes the integer part. $\widetilde{\mathbf{y}}_{T+h,(l)}$ is an estimate of the $\alpha/2$ quantile $\xi_{\alpha/2}$ of the predictive distribution, assessing the precision of this estimate is somewhat more involved than for simple averages. For continuous distributions $f(x)$, the sample order statistic $X_{(m)}$ for $m = \lfloor nq \rfloor$ is a consistent and asymptotically normal estimator of the population quantile ξ_q but the asymptotic variance depends on the underlying distribution, $\sqrt{n}\left(X_{(m)} - \xi_q\right) \xrightarrow{d} N\left(0, q(1-q)/f\left(\xi_q\right)^2\right)$, and requires an estimate of the density at ξ_q in order to be operational.

An alternative procedure based on order statistics can be used to produce distribution free confidence intervals for the population quantile. We seek order statistics $X_{(r)}$ and $X_{(s)}$ that satisfies $P\left(X_{(r)} < \xi_q < X_{(s)}\right) \approx 1-\alpha$. Noting that the probability statement $P\left(X_{(r)} < \xi_q < X_{(s)}\right)$ is equivalent to the statement

$$P(\text{at least } r \text{ but no more than } s-1 \text{ observations satisfy } X_i < \xi_q)$$

this can be evaluated as a Binomial probability,

$$P\left(X_{(r)} < \xi_q < X_{(s)}\right) = \sum_{k=r}^{s-1} \binom{n}{k} q^k (1-q)^{n-k},$$

and for small n it is straightforward to determine values of r and s that gives (approximately) the desired confidence level. For large n the Binomial distribution can be ap-

proximated by a normal distribution and r and s obtained as

$$r = \left\lfloor nq - z_{1-\alpha/2}\sqrt{nq\,(1-1)} \right\rfloor$$
$$s = \left\lceil nq + z_{1-\alpha/2}\sqrt{nq\,(1-1)} \right\rceil .$$

2.2.1 Markov chain Monte Carlo

In general a simulation strategy similar to the one discussed in section 2.1 can be devised to generate a sample from the predictive distribution. The main difficulty is how to generate draws from the posterior distribution of the parameters when, unlike Algorithm 1, it is not possible to sample directly from the posterior. The two most common procedures for solving this problem is importance sampling (Kloek and van Dijk (1978) and Geweke (1989)) and Markov chain Monte Carlo (MCMC). Here we will focus on MCMC methods as these are, in general, quite straightforward to implement with VAR models. Geweke and Whiteman (2006), Chib and Greenberg (1995) and Geweke (1999) gives a more in-depth discussion and book length treatments include Gamerman (1997) and Robert and Casella (1999).

The idea behind MCMC techniques is to construct a Markov chain for the parameters $\boldsymbol{\theta}$ which has the posterior distribution as it's (unique) stationary distribution and fulfills the additional requirement that we can generate random number from the conditional distribution, $f\left(\boldsymbol{\theta}^{(j+1)}|\boldsymbol{\theta}^{(j)}\right)$ that defines the transition kernel. If the initial draw, $\boldsymbol{\theta}^{(0)}$, could somehow be drawn from the posterior distribution, all the following draws will also be from the posterior distribution by virtue of this being the stationary distribution of the chain. But this is, of course, not possible (or we would not need to resort to MCMC methods) and the issue of *convergence* becomes important. Will the distribution of the draws from the Markov chain converge to the posterior distribution if we start the chain at an arbitrary point in the parameter space? And if so, how many draws are required before the distribution of the draws is "close enough" to the posterior distribution?

A precise answer to the first question involves highly technical conditions (see Tierny (1994)). It is, however, possible to state stronger conditions that are sufficient for convergence and relatively easy to check (e.g. Geweke (2005, section 4.5)). One such condition is that, loosely speaking, $P\left(\boldsymbol{\theta}^{(j+1)} \in A|\boldsymbol{\theta}^{(j)}\right) > 0$ for all $\boldsymbol{\theta}^{(j)}$ and any set A with positive probability under the posterior distribution. The Markov chain is then ergodic and allows consistent estimation of posterior quantities. The second question does not have a precise answer and would be unimportant if we could generate an infinite number of draws from the chain. In practice we will only have a finite number of draws available and including draws from the beginning of the chain, before it has converged to the posterior distribution, can give very bad estimates of posterior quantities. As a practical matter it is thus important to discard a sufficiently large number, B, of initial draws (the burn-in). Lacking a precise answer, the choice of the size of the burn-in is subjective and it is better to err on the side of caution. Diagnostics that are useful in determining B are discussed below.

The performance of the Markov chain and the precision of estimates is related to the issue of convergence. Even if the Markov chain is convergent it might move very slowly through the parameter space (mix slowly) with high autocorrelation between the draws and a very large number of draws might be needed in order to explore the parameter space.

Even a well performing Markov chain will by construction have some, typically positive, autocorrelation in the draws which tends to impact the precision of estimates negatively with a larger variance of estimates of, say the posterior mean, than if direct sampling had been possible. For an assessment of the precision of estimates and probabilistic error bounds a central limit theorem is needed. This in turn requires a rate condition on the speed of convergence to the posterior distribution. Let $g(\boldsymbol{\theta})$ be an arbitrary function of the parameters and $\overline{g}$ the average over R draws from the chain. If the chain is geometrically ergodic then

$$\sqrt{R}\left(\overline{g} - E\left(g|\mathbf{Y}_T\right)\right) \xrightarrow{d} N\left(0, \sigma_{MC}^2\right)$$

if $E\left[g^{2+\delta}|\mathbf{Y}_T\right] < \infty$ for $\delta > 0$ and

$$\sigma_{MC}^2 = V\left(g|\mathbf{Y}_T\right) + 2\sum_{k=1}^{\infty} Cov\left[g\left(\boldsymbol{\theta}^{(j)}\right), g\left(\boldsymbol{\theta}^{(j+k)}\right)|\mathbf{Y}_T\right]. \tag{12}$$

If, in addition, the chain is uniformly ergodic then the result holds for $\delta = 0$.

Technical details for the implementation of Gibbs and Metropolis-Hastings samplers are given in Appendix A, including how to assess convergence and how to estimate the variance, σ_{MC}^2, of the Monte Carlo estimate. It should be clear to the reader that these methods come with a health warning: Naive use without careful assessment of the behavior and convergence property of the Markov chain may lead to completely misleading results.

3 Reduced Form VARs

For forecasting purposes reduced form Bayesian VARs, that is models that essentially leaves the parameter matrices and the variance-covariance matrix of $\mathbf{u}_t$ in the VAR $\mathbf{y}_t' = \sum_{i=1}^{p} \mathbf{y}_{t-i}'\mathbf{A}_i + \mathbf{x}_t'\mathbf{C} + \mathbf{u}_t'$ unrestricted have proven to be quite successful. While having a long tradition in time series analysis their use in economic forecasting was limited until Sims (1980) influential critique of the "incredible" identifying restrictions used in the large scale macroeconometric models of the day. Instead Sims argued in favour of VAR-models built essentially on considerations of the time series properties of the data. While being powerful forecast devices that can fit the data well, VAR-models may require relatively large lag lengths p in order to match the time series properties of the data which, with the many parameters to estimate can cause poor forecasting performance. One possible solution to the problems caused by the rich parameterization is to consider the larger class of VARMA-models (see Lütkepohl (2006)) which may be able to represent the data in a more parsimonious fashion.

3.1 The Minnesota prior beliefs

Taking a different route, Litterman(1979, 1980) argued from a largely frequentist view point, using the analogy with Ridge regression and shrinkage estimation, that the precision of estimates and forecasting performance can be improved by incorporating "restrictions" in the form of a prior distribution on the parameters. Litterman's prior formulation is essentially based on stylized facts about his data, macroeconomic variables for the US that could be well characterized by unit root processes and he proposed shrinking towards a univariate random walk for each variable in the VAR.

Recall the multivariate regression formulation for the VAR with m variables, $\mathbf{y}_t' = \mathbf{z}_t'\mathbf{\Gamma} + \mathbf{u}_t'$ for $\mathbf{z}_t' = \left(\mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-p}', \mathbf{x}_t'\right)$ and elements γ_{ij} of $\mathbf{\Gamma}$. The shrinkage towards univariate random walks corresponds the setting the prior mean of $\mathbf{\Gamma}$ to

$$\underline{\gamma}_{ij} = E\left(\gamma_{ij}\right) = \begin{cases} 1, \text{ first own lag, } i = j \\ 0, \ i \neq j \end{cases} . \tag{13}$$

Litterman suggested applying a harder shrinkage towards zero for longer lags, reflecting the prior notion that more distant observations are less influential. In addition, a different amount of shrinkage is applied to lags of the dependent variable than to lags of other variables in the same equation. Typically more shrinkage is applied to lags of other variables to reinforce the univariate random walk nature of the prior. Specifically, Litterman suggested setting the prior standard deviations to

$$\tau_{ij} = sd\left(\gamma_{ij}\right) = \begin{cases} \pi_1 / l^{\pi_3}, \text{ lag } l \text{ of the dependent variable, } i = (l-1)\, m + j \\ \left(\pi_1 \pi_2 s_j\right) / \left(l^{\pi_3} s_r\right), \text{ lag } l \text{ of variable } r \neq j, \ i = (l-1)\, m + r \\ \infty, \text{ deterministic variables, } i = mp + 1, \ldots, k \end{cases} . \tag{14}$$

Here s_j / s_r is a scale factor accounting for the different variances of the dependent and explanatory variables, π_1 is referred to as the "overall tightness", π_2 the "relative tightness of other variables" and π_3 the "lag decay rate". The infinite standard deviations for the coefficients on the deterministic variables $\mathbf{x}_t$ corresponds to an improper uniform prior on the whole real line and could, without affecting the results, be replaced with an arbitrary large value to obtain a proper prior. The prior is completed by using independent normals for each regression coefficient on the lags, $\gamma_{ij} \sim N\left(\underline{\gamma}_{ij}, \tau_{ij}^2\right)$.

To reduce the computational burden Litterman proceeded to estimate the VAR equation by equation rather than as a system of equations. The likelihood is normal and the error variances are assumed to be known, $V\left(u_{tj}\right) = s_j^2$, where s_j^2 is the OLS residual variance for equation j in the VAR or a univariate autoregression for variable j. Equation by equation estimation is, in fact, appropriate if the variance of $\mathbf{u}_t$ is diagonal, $\mathbf{\Psi} = \operatorname{diag}\left(s_1^2, \ldots, s_m^2\right)$ but, as noted by Litterman, suboptimal if the error terms are correlated. Taking the error variances to be known (although data based) is, of course, also a simplification motivated by computational expediency.

For computational purposes, as well as a way to think about the prior in terms of implications for the data, it is useful to note that the prior $\gamma_{ij} \sim N\left(\underline{\gamma}_{ij}, \tau_{ij}^2\right)$ can be restated as

$$\frac{s_j}{\tau_{ij}}\underline{\gamma}_{ij} = \frac{s_j}{\tau_{ij}}\gamma_{ij} + \widetilde{u}_{ij}$$

where $\widetilde{u}_{ij} \sim N\left(0, s_j^2\right)$. The prior information for equation j can thus be written as pseudo data,

$$\mathbf{r}_j = \mathbf{R}_j \boldsymbol{\gamma}_j + \widetilde{\mathbf{u}}_j$$

with element i of $\mathbf{r}_j$ set to $\underline{\gamma}_{ij} s_j / \tau_j$ and element r, s of $\mathbf{R}_j$ zero for $r \neq s$ and $s_j / \tau_{i,j}$ for $r = s = 1, \ldots, mp$. One can then apply the mixed estimation technique of Theil and Goldberger (1960), that is apply OLS to the augmented regression equation

$$\begin{pmatrix} \mathbf{y}_j \\ \mathbf{r}_j \end{pmatrix} = \begin{pmatrix} \mathbf{Z} \\ \mathbf{R}_j \end{pmatrix} \boldsymbol{\gamma}_j + \begin{pmatrix} \mathbf{u}_j \\ \widetilde{\mathbf{u}}_j \end{pmatrix} \tag{15}$$

with known variance error variance s_j^2. This yields the estimate

$$\overline{\boldsymbol{\gamma}}_j = \left(\mathbf{Z}'\mathbf{Z} + \mathbf{R}_j'\mathbf{R}_j\right)^{-1}\left(\mathbf{Z}'\mathbf{y} + \mathbf{R}_j'\mathbf{r}\right)$$

with variance

$$\overline{\mathbf{V}}_j = s_j^2\left(\mathbf{Z}'\mathbf{Z} + \mathbf{R}_j'\mathbf{R}_j\right)^{-1}$$

which corresponds to the posterior mean and variance, $\boldsymbol{\gamma}_j|\mathbf{Y}_T \sim N\left(\overline{\boldsymbol{\gamma}}_j, \overline{\mathbf{V}}_j\right)$ under the assumption that $\mathbf{u}_j \sim N\left(\mathbf{0}, s_j^2\mathbf{I}\right)$ with an (improper) normal prior for $\boldsymbol{\gamma}_j$ with mean $\underline{\boldsymbol{\gamma}}_j$ and precision (inverse variance) $\frac{1}{s_j^2}\mathbf{R}_j'\mathbf{R}_j$. To see this note that $\mathbf{r}_j = \mathbf{R}_j\underline{\boldsymbol{\gamma}}_j$ and that applying the Bayesian calculations directly leads to $\overline{\mathbf{V}}_j = \left(\frac{1}{s_j^2}\mathbf{Z}'\mathbf{Z} + \frac{1}{s_j^2}\mathbf{R}_j'\mathbf{R}_j\right)^{-1}$ and $\overline{\boldsymbol{\gamma}}_j = \overline{\mathbf{V}}\left(\frac{1}{s_j^2}\mathbf{Z}'\mathbf{Z}\widehat{\boldsymbol{\gamma}} + \frac{1}{s_j^2}\mathbf{R}_j'\mathbf{R}_j\underline{\boldsymbol{\gamma}}\right) = \overline{\mathbf{V}}\left(\frac{1}{s_j^2}\mathbf{Z}'\mathbf{y}_j + \frac{1}{s_j^2}\mathbf{R}_j'\mathbf{r}_j\right)$ for $\widehat{\boldsymbol{\gamma}} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{y}_j$.

It remains to specify the prior hyperparameters π_1, π_2 and π_3. Litterman(1979, 1980) conducted several exercises to evaluate the effect of the hyperparameters on the out of sample forecast performance. Suitable choices for his data appear to be $\pi_1 \approx 0.2$, $\pi_2 \approx 0.2$ and $\pi_3 = 1$. These are also close to the hyperparameters used in true out of sample forecast results reported in Litterman (1986).

The actual forecasts produced by Litterman were not based on the predictive distribution (3), in an additional bow to the limited computation resources of the time Litterman approximated the mean of the predictive distribution by calculating the forecasts using the posterior means $\overline{\boldsymbol{\gamma}}_j$ of the parameters and the chain rule of forecasting.

In the remainder of this chapter we will refer to priors with moments similar to (13) and (14) as Minnesota type priors or as priors based on the Minnesota prior beliefs. The term Litterman prior is reserved for the combination of these prior beliefs with the assumption of a diagonal and known error variance matrix.

Forecasting performance Using the Minnesota prior and the forecasting procedure outlined above Litterman started issuing monthly forecasts from a six variable VAR with real GNP, the GNP price deflator, real business fixed investments, the 3-month treasury bill, the unemployment rate and the money supply in 1980. Five years later, and with the model essentially unchanged Litterman (1986) and McNees (1986) report on the forecast accuracy of these true out of sample forecasts compared to commercial forecasts based on large scale macroeconometric models. There is no clear winner in this comparison, the BVAR forecasts dominated for the real variables (real GNP, investments and unemployment) but were among the worst for inflation and the T-bill rate.

3.1.1 Variations on the Minnesota prior

Many variations on the Minnesota prior have been suggested, common ones include

- Stationary variables: For variables believed to be stationary the prior mean on the first lag can be set to a value less than 1, for example $\underline{\gamma}_{jj} = 0.9$ if the variable is believed to be relatively persistent.
- Deterministic variables: Set the prior standard deviations to $\tau_{ij} = \pi_1\pi_4 s_j$, this has the advantage of leading to a proper prior for the coefficients on deterministic variables while still being uninformative about γ_{ij} by setting π_4 (moderately) large.

- "Exogenous" variables: Set the prior standard deviation to $\tau_{ij} = (\pi_1 \pi_5 s_j) / (l^{\pi_3} s_r)$, for lag l of the "endogenous" variable r in the equation for the "exogenous" dependent variable j, $i = (l-1)\, m + r$. This is, for example, useful when modelling a small open economy with "rest of the world" variables included in the model. Forecasting is simplified if these variables are included in $\mathbf{y}_t$ as no external forecasts are needed. Setting π_5 small shrinks γ_{ij} aggressively towards zero and allows us to express that the rest of the world variables are essentially exogenous to the domestic economy.

- Sum of coefficients prior: This prior (introduced by Doan, Litterman and Sims (1984)) expresses the prior notion that the sum of coefficients on own lags is 1 and the sum of coefficients on the lags of each of the other variables is 0 as well as the idea that the recent average of the variable should be a reasonable forecast. To implement this add m rows to $\mathbf{R}_j$ which are zero except for the p positions in the i^{th} row corresponding to variable $i = 1, \ldots, m$, i.e. row i is given by $\left(\overline{\mathbf{w}}_i' \otimes \mathbf{j}_p', 0\right)$ where the zeros correspond to the deterministic variables, element i of $\overline{\mathbf{w}}_i$ is $\overline{y}_{0,i} s_i / (\pi_1 \pi_6 s_j)$ for $\overline{y}_{0i} = \frac{1}{p} \sum_{t=1-p}^{0} y_{t,i}$ the average of the initial conditions for variable i and the remaining $m-1$ elements zero, $\mathbf{j}_p$ is a $p \times 1$ vector of ones. In addition add m elements to $\mathbf{r}_j$ with the j^{th} element equal to $\overline{y}_{0,j} / (\pi_1 \pi_6)$. The prior induces correlation between the coefficients on the same variable (the prior precision $\frac{1}{s_j^2} \mathbf{R}_j' \mathbf{R}_j$ is no longer a diagonal matrix) and forces the model towards a random walk with possible drift for variable j as $\pi_6 \to 0$.

- Dummy initial observations prior: Add a row $\left(\overline{\mathbf{y}}_0' \otimes \mathbf{j}_p', \overline{\mathbf{x}}_0'\right) / (\pi_1 \pi_7 s_j)$ to $\mathbf{R}_j$ and $\overline{y}_{0j} / (\pi_1 \pi_7 s_j)$ to $\mathbf{r}_j$. This prior also implies that the initial observations is a good forecast without enforcing specific parameter values and induces prior correlation among all parameters in the equation. Sims (1993) argues that the dummy initial observations prior is preferable to the sum of coefficients prior. As $\pi_7 \to 0$ the prior implies that either all variables are stationary with mean $\overline{\mathbf{y}}_0$ or that there are unit root components without drift (if there are no trends in $\mathbf{x}_t$).

3.2 Flexible prior distributions

The basic setup of Litterman has been generalized in several directions, attempting to relax some of the more restrictive assumptions that were motivated by the computational limitations of the time or that allows different ways of expressing the prior beliefs. Common to these works is that they maintain the basic flavour of the Minnesota prior as a data centric specification that embodies stylized facts about the time series properties of the data.

Kadiyala and Karlsson (1993,1997) relaxes the assumption of a known diagonal error variance-covariance matrix, $\mathbf{\Psi}$, and studies the effect of varying the family of distribution used to parameterize the prior beliefs. They considered the diffuse prior (which we have already encountered in section 2.1.1), the conjugate normal-Wishart prior, the normal-diffuse prior and an adaption of the extended natural conjugate (ENC) prior originally proposed by Drèze and Morales (1976) in the context of simultaneous equation models. Kadiyala and Karlsson (1993) focuses on the forecasting performance and conducts three small forecasting horse races comparing the forecasting performance of the "new" priors with the Minnesota prior and forecasts based on OLS estimates. With the exception of

the diffuse prior the priors are specified to embody prior beliefs about $\mathbf{\Gamma}$ that are similar to the Minnesota prior. With the Minnesota prior and OLS the forecasts are calculated using the chain rule based whereas Monte Carlo methods are used to evaluate the expected value of the predictive distribution with the other priors. There is no clear cut winner, priors that allow for correlation between equations tend to do better.

Kadiyala and Karlsson (1997) studies the same four priors but this time the focus is on the implementation and efficiency of Monte Carlo methods for evaluating the expected value of the predictive distribution. Importance samplers and Gibbs samplers are developed for the posterior distributions arising from the normal-diffuse and ENC priors. Kadiyala and Karlsson concludes that Gibbs sampling is more efficient than importance sampling, in particular for larger models. The evaluation is done in the context of two forecasting exercises, one using a small bivariate model for the Swedish industrial production index and unemployment rate and one using the seven variable model of Litterman (1986). In terms of forecast performance there is no clear winner, the diffuse, normal-Wishart priors and forecasts based on the OLS estimates does best with the Swedish data and the Minnesota, normal-Wishart and normal-diffuse does best with the Litterman model. In the following we will focus on the normal-Wishart and normal-diffuse priors as the ENC prior is quite complicated to work with and did not perform significantly better than the other priors in terms of forecasting performance.

Departing from the normal-Wishart prior Giannone, Lenza and Primiceri (2012) suggests a hierarchical prior structure that allows the choice of prior hyperparameters to be influenced by the data and, in a sense, makes the procedure more "objective".

3.2.1 The normal-Wishart prior

The normal-Wishart prior is the natural conjugate prior for normal multivariate regressions. It generalizes the original Litterman prior by treating the error variance-covariance matrix, $\mathbf{\Psi}$, as an unknown positive definite symmetric matrix rather than a fixed diagonal matrix. By allowing for correlation between the equations this also leads to computationally convenient system estimation instead of the equation by equation approach used by Litterman. This does, however, come with the disadvantage of imposing a Kronecker structure on the variance-covariance matrix of $\boldsymbol{\gamma}$.

Using the trick of adding and subtracting $\mathbf{Z}\widehat{\mathbf{\Gamma}}$ in the likelihood (7) and letting $\mathbf{S} = \left(\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{\Gamma}}\right)' \left(\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{\Gamma}}\right)$ be the error sum of squares we see that the likelihood

$$\begin{aligned} L\left(\mathbf{Y}|\mathbf{\Gamma}, \mathbf{\Psi}\right) \propto |\mathbf{\Psi}|^{-T/2} \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\mathbf{\Psi}^{-1}\mathbf{S}\right]\right\} \\ \times \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\mathbf{\Psi}^{-1}\left(\mathbf{\Gamma} - \widehat{\mathbf{\Gamma}}\right)'\mathbf{Z}'\mathbf{Z}\left(\mathbf{\Gamma} - \widehat{\mathbf{\Gamma}}\right)\right]\right\} \end{aligned}$$

has the form of a normal-Wishart distribution when considered as a function of $\mathbf{\Gamma}$ and $\mathbf{\Psi}$. Specifying the prior similarly,

$$\begin{aligned} \mathbf{\Gamma}|\mathbf{\Psi} &\sim MN_{km}\left(\underline{\mathbf{\Gamma}}, \mathbf{\Psi}, \underline{\mathbf{\Omega}},\right) \\ \mathbf{\Psi} &\sim iW\left(\underline{\mathbf{S}}, \underline{v}\right), \end{aligned} \tag{16}$$

we have the conjugate normal-Wishart prior with the corresponding posterior,

$$p\left(\mathbf{\Gamma},\mathbf{\Psi}|\mathbf{Y}_T\right)\propto|\mathbf{\Psi}|^{-T/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\mathbf{\Psi}^{-1}\left(\mathbf{\Gamma}-\widehat{\mathbf{\Gamma}}\right)'\mathbf{Z}'\mathbf{Z}\left(\mathbf{\Gamma}-\widehat{\mathbf{\Gamma}}\right)\right]\right\}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\mathbf{\Psi}^{-1}\mathbf{S}\right]\right\} \tag{17}$$

$$\times|\mathbf{\Psi}|^{-(v+m+k+1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\mathbf{\Psi}^{-1}\left(\mathbf{\Gamma}-\underline{\mathbf{\Gamma}}\right)'\underline{\mathbf{\Omega}}^{-1}\left(\mathbf{\Gamma}-\underline{\mathbf{\Gamma}}\right)\right]\right\}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\mathbf{\Psi}^{-1}\underline{\mathbf{S}}\right]\right\}$$

$$=|\mathbf{\Psi}|^{-(T+v+m+k+1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\mathbf{\Psi}^{-1}\left(\mathbf{\Gamma}-\overline{\mathbf{\Gamma}}\right)'\overline{\mathbf{\Omega}}^{-1}\left(\mathbf{\Gamma}-\overline{\mathbf{\Gamma}}\right)\right]\right\}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\mathbf{\Psi}^{-1}\overline{\mathbf{S}}\right]\right\},$$

where the last line is obtained by completing the square for $\mathbf{\Gamma}$. That is

$$\begin{aligned}\mathbf{\Gamma}|\mathbf{Y}_T,\mathbf{\Psi}&\sim MN_{km}\left(\overline{\mathbf{\Gamma}},\mathbf{\Psi},\overline{\mathbf{\Omega}}\right)\\ \overline{\mathbf{\Omega}}^{-1}&=\underline{\mathbf{\Omega}}^{-1}+\mathbf{Z}'\mathbf{Z},\\ \overline{\mathbf{\Gamma}}&=\overline{\mathbf{\Omega}}\left(\underline{\mathbf{\Omega}}^{-1}\underline{\mathbf{\Gamma}}+\mathbf{Z}'\mathbf{Z}\widehat{\mathbf{\Gamma}}\right)=\overline{\mathbf{\Omega}}\left(\underline{\mathbf{\Omega}}^{-1}\underline{\mathbf{\Gamma}}+\mathbf{Z}'\mathbf{Y}\right)\end{aligned} \tag{18}$$

and

$$\begin{aligned}\mathbf{\Psi}|\mathbf{Y}_T&\sim iW\left(\overline{\mathbf{S}},\overline{v}\right),\ \overline{v}=T+\underline{v}\\ \overline{\mathbf{S}}&=\underline{\mathbf{S}}+\mathbf{S}+\left(\underline{\mathbf{\Gamma}}-\widehat{\mathbf{\Gamma}}\right)'\left(\underline{\mathbf{\Omega}}+\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\right)^{-1}\left(\underline{\mathbf{\Gamma}}-\widehat{\mathbf{\Gamma}}\right)\end{aligned} \tag{19}$$

with $\widehat{\mathbf{\Gamma}}=\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y}$ and $\mathbf{S}=\left(\mathbf{Y}-\mathbf{Z}\widehat{\mathbf{\Gamma}}\right)'\left(\mathbf{Y}-\mathbf{Z}\widehat{\mathbf{\Gamma}}\right)$.

For the conjugate normal-Wishart prior the marginal likelihood is available in closed form. It can easily be derived by integrating out $\mathbf{\Gamma}$ and $\mathbf{\Psi}$ in (17) while keeping track of all the constants that have been left out in the product of the likelihood and the prior. Alternatively we rely on the properties of the matricvariate normal and inverse Wishart distributions given in Appendix C. From the likelihood we have the conditional distribution of $\mathbf{Y}$ as $\mathbf{Y}|\mathbf{\Gamma},\mathbf{\Psi}\sim MN_{Tm}\left(\mathbf{Z\Gamma},\mathbf{\Psi},\mathbf{I}_T\right)$, from the prior we deduce that $\mathbf{Z\Gamma}|\mathbf{\Psi}\sim MN_{Tm}\left(\mathbf{Z}\underline{\mathbf{\Gamma}},\mathbf{\Psi},\mathbf{Z}\underline{\mathbf{\Omega}}\mathbf{Z}'\right)$ and $\mathbf{Y}|\mathbf{\Psi}\sim MN_{Tm}\left(\mathbf{Z}\underline{\mathbf{\Gamma}},\mathbf{\Psi},\mathbf{I}_T+\mathbf{Z}\underline{\mathbf{\Omega}}\mathbf{Z}'\right)$. Finally, since the prior for $\mathbf{\Psi}$ is inverse Wishart this leads to a matricvariate-t marginal distribution for $\mathbf{Y}$,

$$\mathbf{Y}\sim Mt_{Tm}\left(\mathbf{Z}\underline{\mathbf{\Gamma}},\left(\mathbf{I}_T+\mathbf{Z}\underline{\mathbf{\Omega}}\mathbf{Z}'\right)^{-1},\underline{\mathbf{S}},\underline{v}\right). \tag{20}$$

Specifying the prior beliefs Specifying the prior means in the fashion of the Minnesota prior is straightforward while the prior variances involve some difficulties. First, recall that the marginal prior distribution of $\mathbf{\Gamma}$ is matricvariate t with variance-covariance matrix $V\left(\boldsymbol{\gamma}\right)=\frac{1}{\underline{v}-m-1}\underline{\mathbf{S}}\otimes\underline{\mathbf{\Omega}}$ and that $\boldsymbol{\gamma}$ has moments up to order $v-m$. The Kronecker structure of the variance-covariance matrix makes it apparent that it is not possible to specify the prior standard deviations or variances as in (14). The variance-covariance matrix of one equation must be proportional to the variance-covariance matrix of the other equations. With $V\left(\boldsymbol{\gamma}_j\right)=\frac{s_{jj}}{\underline{v}-m-1}\underline{\mathbf{\Omega}}$ we can set the diagonal elements of $\underline{\mathbf{\Omega}}$ to

$$\underline{\omega}_{ii}=\begin{cases}\pi_1^2/\left(l^{\pi_3}s_r\right)^2,\text{lag } l \text{ of variable } r, i=(l-1)\,m+r\\ \left(\pi_1\pi_4\right)^2, i=mp+1,\ldots,k\end{cases} \tag{21}$$

and let $s_{jj} = (\underline{v} - m - 1)\, s_j^2$ to achieve something which approximates the variances of the Minnesota prior. That is, the prior parameter matrix for the inverse Wishart is

$$\underline{\mathbf{S}} = (\underline{v} - m - 1)\, \text{diag}\left(s_1^2, \ldots s_m^2\right) \tag{22}$$

with prior expectation $E(\mathbf{\Psi}) = \text{diag}(s_1^2, \ldots s_m^2)$. We are implicitly setting $\pi_2 = 1$ in (14) and it is reasonable to use a smaller value of π_1 here to balance between the Minnesota type tight prior on lags of other variables and a looser prior on own lags.

It is, in general, advisable to set the prior variances for coefficients on deterministic variables to a large positive number as in (21) rather than the improper uniform prior in the original Minnesota prior. Noting that $\underline{\mathbf{\Omega}}$ enters the prior as the inverse and that $\overline{\mathbf{S}}$ can be rewritten as a function of $\underline{\mathbf{\Omega}}^{-1}$ it is, however, possible to work with $\underline{\mathbf{\Omega}}^{-1}$ and specify an improper prior by setting the corresponding diagonal elements of $\underline{\mathbf{\Omega}}^{-1}$ to zero.

The prior degrees of freedom of the inverse Wishart for $\mathbf{\Psi}$ might also require some care, we must have $\underline{v} \geq m + 2$ for the prior variance to exists and $\underline{v} \geq m + 2h - T$ for the variance of the predictive distribution at lead time h to exist.

Simulating from the posterior distribution With a normal-Wishart posterior we can proceed as in Algorithm 1 using the posterior distributions (18) and (19).

3.2.2 The normal-diffuse and independent normal-Wishart priors

The normal-diffuse prior takes a simple form with prior independence between $\mathbf{\Gamma}$ and $\mathbf{\Psi}$. A normal prior for $\boldsymbol{\gamma}$, $\boldsymbol{\gamma} \sim N\left(\underline{\boldsymbol{\gamma}}, \underline{\mathbf{\Sigma}}_\gamma\right)$ and a Jeffreys' prior for $\mathbf{\Psi}$,

$$p(\mathbf{\Psi}) \propto |\mathbf{\Psi}|^{-(m+1)/2}. \tag{23}$$

This prior lacks the computationally convenient Kronecker structure of the variance-covariance matrix of the normal-Wishart prior but it has the great advantage of not placing any restrictions on the prior variance-covariance $\underline{\mathbf{\Sigma}}$. The joint posterior distribution has the form

$$p(\mathbf{\Gamma}, \mathbf{\Psi}|\mathbf{Y}_T) \propto |\mathbf{\Psi}|^{-(T+m+1)/2} \exp\left[-\frac{1}{2}(\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}})' \left(\mathbf{\Psi}^{-1} \otimes \mathbf{Z}'\mathbf{Z}\right)(\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}})\right]$$
$$\times \exp\left[-\frac{1}{2}\left(\boldsymbol{\gamma} - \underline{\boldsymbol{\gamma}}\right)' \underline{\mathbf{\Sigma}}_\gamma^{-1}\left(\boldsymbol{\gamma} - \underline{\boldsymbol{\gamma}}\right)\right].$$

This prior was first considered by Zellner (1971) in the context of seemingly unrelated regression models. He showed that the marginal posterior for $\boldsymbol{\gamma}$ can be expressed as the product of the normal prior and the marginal matricvariate t-distribution (9). The marginal posterior is bimodal if there is a sufficiently large difference between the center of the prior information and the center of the data information. This can be troublesome for MCMC schemes which might get stuck at one of the modes.

The full conditional posteriors are easy to derive. Completing the square for $\boldsymbol{\gamma}$ we have

$$\begin{aligned} \boldsymbol{\gamma}|\mathbf{Y}_T, \mathbf{\Psi} &\sim N\left(\overline{\boldsymbol{\gamma}}, \overline{\mathbf{\Sigma}}_\gamma\right) \\ \overline{\mathbf{\Sigma}}_\gamma &= \left(\underline{\mathbf{\Sigma}}_\gamma^{-1} + \mathbf{\Psi}^{-1} \otimes \mathbf{Z}'\mathbf{Z}\right)^{-1}, \\ \overline{\boldsymbol{\gamma}} &= \overline{\mathbf{\Sigma}}_\gamma\left[\underline{\mathbf{\Sigma}}_\gamma^{-1}\underline{\boldsymbol{\gamma}} + \left(\mathbf{\Psi}^{-1} \otimes \mathbf{Z}'\mathbf{Z}\right)\widehat{\boldsymbol{\gamma}}\right] = \overline{\mathbf{\Sigma}}_\gamma\left[\underline{\mathbf{\Sigma}}_\gamma^{-1}\underline{\boldsymbol{\gamma}} + \text{vec}\left(\mathbf{Z}'\mathbf{Y}\mathbf{\Psi}^{-1}\right)\right] \end{aligned} \tag{24}$$

where we have used that $\widehat{\boldsymbol{\gamma}} = \left[\boldsymbol{\Psi}^{-1} \otimes \mathbf{Z}'\mathbf{Z}\right]^{-1} \left(\boldsymbol{\Psi}^{-1} \otimes \mathbf{Z}'\right) \mathbf{y}$ and $\left(\boldsymbol{\Psi}^{-1} \otimes \mathbf{Z}'\right) \mathbf{y} = \text{vec}\left(\mathbf{Z}'\mathbf{Y}\boldsymbol{\Psi}^{-1}\right)$ for the last line. Note that this involves the inversion of the $mk \times mk$ matrix $\underline{\boldsymbol{\Sigma}}_{\gamma}^{-1} + \boldsymbol{\Psi}^{-1} \otimes \mathbf{Z}'\mathbf{Z}$ which can be computationally demanding and numerically unstable for large models.[6] The conditional posterior for $\boldsymbol{\Psi}$ follows directly from the likelihood (7),

$$\begin{aligned} \boldsymbol{\Psi}|\mathbf{Y}_T, \boldsymbol{\Gamma} &\sim iW\left(\overline{\mathbf{S}}, \overline{v}\right), \ \overline{v} = T \\ \overline{\mathbf{S}} &= \left(\mathbf{Y} - \mathbf{Z}\boldsymbol{\Gamma}\right)'\left(\mathbf{Y} - \mathbf{Z}\boldsymbol{\Gamma}\right). \end{aligned} \tag{25}$$

The normal-diffuse prior is not a proper prior, which might be an issue in some cases, even if we are assured that the posterior is proper as long as $T > k$. A simple modification is to replace the improper Jeffreys' prior for $\boldsymbol{\Psi}$ with an inverse Wishart, $\boldsymbol{\Psi} \sim iW\left(\underline{\mathbf{S}}, \underline{v}\right)$. The use of the *independent normal-Wishart* prior leaves the conditional posterior for $\boldsymbol{\Gamma}$ unaffected and the conditional posterior for $\boldsymbol{\Psi}$ is still inverse Wishart but now with parameters

$$\overline{\mathbf{S}} = \underline{\mathbf{S}} + \left(\mathbf{Y} - \mathbf{Z}\boldsymbol{\Gamma}\right)'\left(\mathbf{Y} - \mathbf{Z}\boldsymbol{\Gamma}\right), \ \overline{v} = T + \underline{v}. \tag{26}$$

Specifying the prior beliefs With the $N\left(\underline{\boldsymbol{\gamma}}, \underline{\boldsymbol{\Sigma}}_{\gamma}\right)$ form of the prior for $\boldsymbol{\gamma}$ it is straightforward to implement a basic Minnesota prior that is informative about all regression parameters. Improper priors for the coefficients on deterministic variables can be implemented by working with the prior precision and setting the corresponding diagonal elements of $\underline{\boldsymbol{\Sigma}}_{\gamma}^{-1}$ to zero. Similarly, in order to implement the sum of coefficients prior or the initial observations prior it is most convenient to form the dummy observations $\mathbf{R}_j\boldsymbol{\gamma}_j = \mathbf{r}_j$ and add $\frac{1}{s_j^2}\mathbf{R}_j'\mathbf{R}_j$ to the corresponding diagonal block of $\underline{\boldsymbol{\Sigma}}_{\gamma}^{-1}$ and $\frac{1}{s_j^2}\mathbf{R}'\mathbf{r}$ to $\boldsymbol{\gamma}_j$.

Simulating from the posterior distribution With the full conditional posteriors in hand a straightforward Gibbs sampling scheme is available for sampling from the posterior and predictive distributions, see Algorithm 2. The experience of Kadiyala and Karlsson (1997) is that the Gibbs sampler convergences quickly to the posterior distribution and a few hundred draws may be sufficient as burn-in when the posterior is unimodal.

3.2.3 A hierarchical prior for the hyperparameters

The prior hyperparameters are in general chosen in three different ways, as default values similar to the ones used by Litterman, to minimize the forecast errors over a training sample or in an empirical Bayes fashion by maximizing the marginal likelihood with respect to the hyperparameters. As an alternative Giannone et al. (2012) suggests a more flexible approach where one more layer is added to the prior structure by placing a prior on the hyperparameters in a hierarchical fashion. Collecting the hyperparameters in the vector $\boldsymbol{\delta}$ and working with the normal-Wishart family of prior distributions the prior structure becomes

$$\pi\left(\boldsymbol{\Gamma}|\boldsymbol{\Psi},\boldsymbol{\delta}\right)\pi\left(\boldsymbol{\Psi}|\boldsymbol{\delta}\right)\pi\left(\boldsymbol{\delta}\right).$$

[6] This is exactly the computational advantage of the Normal-Wishart prior. By retaining the Kronecker structure of the variance-covariance matrix, $\boldsymbol{\Psi}^{-1} \otimes \overline{\mathbf{S}}$, in the conditional posterior for γ only inversion of $m \times m$ and $k \times k$ matrices is needed and it is only $\boldsymbol{\Psi}^{-1}$ (or it's Cholesky factor) that needs to be recomputed for each draw from the posterior.

Algorithm 2 Gibbs sampler for normal-diffuse and independent normal-Wishart priors

Select a starting value, $\boldsymbol{\gamma}^{(0)}$ for $\boldsymbol{\gamma}$. For $j = 1, \ldots, B + R$

1. Generate $\boldsymbol{\Psi}^{(j)}$ from the full conditional posterior (25) with $\overline{\mathbf{S}}$ evaluated at $\boldsymbol{\gamma}^{(j-1)}$ where the posterior parameters are given by (25) for the normal-diffuse prior and (26) for the independent normal-Wishart prior.

2. Generate $\boldsymbol{\gamma}^{(j)}$ from the full conditional posterior (24) with $\overline{\boldsymbol{\Sigma}}_{\gamma}$ evaluated at $\boldsymbol{\Psi}^{(j)}$.

3. For $j > B$, generate $\mathbf{u}_{T+1}^{(j)}, \ldots, \mathbf{u}_{T+H}^{(j)}$ from $\mathbf{u}_t \sim N\left(0, \boldsymbol{\Psi}^{(j)}\right)$ and calculate recursively

$$\widetilde{\mathbf{y}}_{T+h}^{(j)\prime} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}_{T+h-i}^{(j)\prime} \mathbf{A}_i^{(j)} + \sum_{i=h}^{p} \mathbf{y}_{T+h-i}^{\prime} \mathbf{A}_i^{(j)} + \mathbf{x}_{T+h}^{\prime} \mathbf{C}^{(j)} + \mathbf{u}_{T+h}^{(j)\prime}.$$

   The first B draws are discarded as burn-in. Discarding the parameters yields $\left\{\widetilde{\mathbf{y}}_{T+1}^{(j)}, \ldots \widetilde{\mathbf{y}}_{T+H}^{(j)}\right\}_{j=B+1}^{B+R}$ as a sample of independent draws from the joint predictive distribution.

Conditioning on $\boldsymbol{\delta}$ the analysis is as before and the results in section 3.2.1 holds when interpreted conditional on $\boldsymbol{\delta}$. In addition to putting a prior on the hyperparameters Giannone et al. (2012) relaxes some of the simplifying choices that are commonly made when setting up the prior. Instead of setting the diagonal elements of $\underline{\mathbf{S}}$ in the prior for $\boldsymbol{\Psi}$ based on the residual standard variance from OLS estimated VAR or univariate AR models Giannone et al. proposes treating them as parameters. That is, they set $\underline{\mathbf{S}} = \operatorname{diag}(\kappa_1, \ldots, \kappa_m)$ and endow κ_i with independent inverse Gamma priors, $\kappa_i \sim iG(a_\kappa, b_\kappa)$. The conditional prior for $\boldsymbol{\Psi}$ is thus $\boldsymbol{\Psi}|\boldsymbol{\delta} \sim iW(\underline{\mathbf{S}}, \underline{v})$. The prior variance specification for $\boldsymbol{\Gamma}$ can then be completed by setting

$$\underline{\omega}_{ii} = \begin{cases} \left(\pi_1 (\underline{v} - m - 1) / (l\kappa_r)\right)^2, \text{ lag } l \text{ of variable } r, i = (l-1)m + r \\ \pi_4^2 (\underline{v} - m - 1), \ i = mp + 1, \ldots, k \end{cases}$$

with $\underline{\boldsymbol{\Omega}} = \operatorname{diag}(\underline{\boldsymbol{\omega}})$ yielding the prior variances

$$V(\gamma_{ij}) = \begin{cases} (\pi_1^2 \kappa_j) / (l^2 \kappa_r), \text{ lag } l \text{ of variable } r, i = (l-1)m + r \\ \pi_4^2 \kappa_j, \ i = mp + 1, \ldots, k \end{cases}.$$

The prior means of $\boldsymbol{\Gamma}$ is, $\underline{\boldsymbol{\Gamma}}$, set to one for the first own lag and zero otherwise, the prior for $\boldsymbol{\Gamma}$ is thus $\boldsymbol{\Gamma}|\boldsymbol{\Psi},\boldsymbol{\delta} \sim MN_{km}(\underline{\boldsymbol{\Gamma}}, \boldsymbol{\Psi}, \underline{\boldsymbol{\Omega}})$

In addition to this Giannone et al. adds dummy observations for a sum of coefficients prior and a dummy initial observation prior. Let $\widetilde{\mathbf{Y}}$ and $\widetilde{\mathbf{Z}}$ be the dummy observations specified similar to section 3.1.1. Giannone et al. sets

$$\widetilde{\mathbf{Y}} = \begin{pmatrix} \frac{1}{\pi_6} \operatorname{diag}(\overline{\mathbf{y}}_0) \\ \frac{1}{\pi_7} \overline{\mathbf{y}}_0^{\prime} \end{pmatrix}, \ \widetilde{\mathbf{Z}} = \begin{pmatrix} \mathbf{j}_m^{\prime} \otimes \left(\frac{1}{\pi_6} \operatorname{diag}(\overline{\mathbf{y}}_0)\right) & \mathbf{0}_{m \times (k-mp)} \\ \mathbf{j}_m^{\prime} \otimes \left(\frac{1}{\pi_7} \overline{\mathbf{y}}_0^{\prime}\right) & \frac{1}{\pi_7} \mathbf{j}_{(k-mp)}^{\prime} \end{pmatrix}.$$

The dummy observations are then appended to the data matrices $\mathbf{Y}$ and $\mathbf{Z}$ and the posterior parameters calculated as usual.[7] Giannone et al. uses independent Gamma priors for the scale factors and we set $\pi_i \sim G(a_i, b_i)$, $i = 1, 4, 6, 7$. The collection of hyperparameters is thus $\boldsymbol{\delta} = (\kappa_1, \ldots, \kappa_m, \pi_1, \pi_4, \pi_6, \pi_7)'$.

Specifying the prior beliefs The priors for the hyperparameters $\pi_1, \pi_4, \pi_6, \pi_7$ can be centered on "standard" settings for these parameters with variance depending on how confident we are about the "standard" values. Giannone et al. (2012) sets the modes for π_1, π_6 and π_7 to 0.2, 1 and 1 with standard deviations 0.4, 1 and 1. For π_4 a large mode, say 50, with a large standard deviation seems reasonable. For the diagonal elements of $\underline{\mathbf{S}}$, κ_i, Giannone et al. implements the prior in terms of $\kappa_i / (\underline{v} - m - 1)$, i.e. the prior mean of $\underline{\mathbf{S}}$, and use a highly non-informative prior with $a_\kappa = b_\kappa = 0.02^2$.

Simulating from the posterior distribution The joint posterior of $\boldsymbol{\Gamma}$, $\boldsymbol{\Psi}$ and $\boldsymbol{\delta}$ is not available in closed form but Giannone et al. (2012) devices a Metropolis-Hastings sampler for the joint distribution, see Algorithm 3. The algorithm generates $\boldsymbol{\delta}$ from the marginal posterior with a Metropolis-Hastings update, after convergence of the $\boldsymbol{\delta}$ sampler $\boldsymbol{\Psi}$ and $\boldsymbol{\Gamma}$ can be drawn from their distributions conditional on $\boldsymbol{\delta}$. While Giannone et al. takes advantage of the availability of the marginal likelihood conditional on $\boldsymbol{\delta}$ to simplify the acceptance probability in the Metropolis-Hastings step and achieve a marginal sampler for $\boldsymbol{\delta}$ this is not a requirement. The acceptance probability can also be written in terms of the likelihood and the priors and a Metropolis within Gibbs sampler can be devised when the conditional marginal likelihood is not available in closed form.

Forecasting performance Giannone et al. (2012) conducts a forecasting experiment where they forecast the US GDP, GDP deflator and federal funds rate. This done using three different BVARs implemented using the hierarchical prior with 3, 7 and 22 variables with all variables in log-levels. In addition to the BVARs forecasts are also produced with VARs estimated with OLS, a random walk with drift and a dynamic factor model based on principal components from a data set with 149 macro variables. In terms of mean square error the BVARs improve with the size of the model (in contrast to the OLS estimated VARs) and the largest BVAR produces better one step ahead forecasts than the factor model for the GDP deflator and the federal funds rate and better four step ahead forecasts for the GDP deflator.

[7] The additional information in the dummy observations can of course also be incorporated through the priors. The implied prior parameters are $\underline{\boldsymbol{\Omega}}^* = \left(\underline{\boldsymbol{\Omega}}^{-1} + \widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}\right)^{-1}$, $\underline{\boldsymbol{\Gamma}}^* = \underline{\boldsymbol{\Omega}}^* \left(\underline{\boldsymbol{\Omega}}^{-1}\underline{\boldsymbol{\Gamma}} + \widetilde{\mathbf{Z}}'\widetilde{\mathbf{Y}}\right)$ and $\underline{\mathbf{S}}^* = \underline{\mathbf{S}} + \left(\mathbf{Y}^* - \mathbf{Z}^*\widehat{\boldsymbol{\Gamma}}^*\right)' \left(\mathbf{Y}^* - \mathbf{Z}^*\widehat{\boldsymbol{\Gamma}}^*\right) - \left(\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\Gamma}}\right)' \left(\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\Gamma}}\right)$ where $\mathbf{Y}^*$ and $\mathbf{Z}^*$ is the augmented data, $\widehat{\boldsymbol{\Gamma}}^*$ the OLS estimate on the augmented data and $\widehat{\boldsymbol{\Gamma}}$ the OLS estimate on the original data. The effect on $\underline{\boldsymbol{\Omega}}$ and $\underline{\boldsymbol{\Gamma}}$ is clear and intuitive whereas $\underline{\mathbf{S}}$ is inflated in a data dependent and non-obvious way. The mixed estimation technique underlying the device of adding prior information through dummy observations works well when the error variance is assumed known but is less transparent when it is unknown.

Algorithm 3 MCMC sampler for a VAR with hierarchical prior

For the VAR model with the hierarchical prior outlined in section 3.2.3 select starting values for the hyperparameters $\boldsymbol{\delta}^{(0)}$, Giannone et al. (2012) suggests using the posterior mode of $\boldsymbol{\delta}$ as starting values and setting the tuning constant c to achieve approximately 20% acceptance rate. Step 1 of the sampler samples from the marginal posterior for $\boldsymbol{\delta}$, steps 2 and 3 draws from the posterior for $\boldsymbol{\Psi}$ and $\boldsymbol{\Gamma}$ conditional on $\boldsymbol{\delta}$.
For $j = 1, \ldots, B + R$

1. Draw a proposal, $\boldsymbol{\delta}^*$, for the hyperparameters from the random walk proposal distribution, $\boldsymbol{\delta}^* \sim N\left(\boldsymbol{\delta}^{(j-1)}, c\mathbf{H}^{-1}\right)$ where $\mathbf{H}$ is the Hessian of the negative of the logposterior for $\boldsymbol{\delta}$. Set $\boldsymbol{\delta}^{(j)} = \boldsymbol{\delta}^*$ with probability α, otherwise set $\boldsymbol{\delta}^{(j)} = \boldsymbol{\delta}^{(j-1)}$ where
$$\alpha = \min\left(1, \frac{m\left(\mathbf{Y}|\boldsymbol{\delta}^*\right)\pi\left(\boldsymbol{\delta}^*\right)}{m\left(\mathbf{Y}|\boldsymbol{\delta}^{(j-1)}\right)\pi\left(\boldsymbol{\delta}^{(j-1)}\right)}\right)$$
and $m\left(\mathbf{Y}|\boldsymbol{\delta}\right)$ is given by (20).
Redo 1 if $j < B$ otherwise continue.

2. Draw $\boldsymbol{\Psi}^{(j)}$ from the full conditional posterior $\boldsymbol{\Psi}|\mathbf{Y}_T, \boldsymbol{\delta}^{(j)}$ in (19)

3. Draw $\Gamma^{(j)}$ from the full conditional posterior $\boldsymbol{\Gamma}|\mathbf{Y}_T, \boldsymbol{\Psi}^{(j)}, \boldsymbol{\delta}^{(j)}$ in (18).

4. Generate $\mathbf{u}_{T+1}^{(j)}, \ldots, \mathbf{u}_{T+H}^{(j)}$ from $\mathbf{u}_t \sim N\left(0, \boldsymbol{\Psi}^{(j)}\right)$ and calculate recursively
$$\widetilde{\mathbf{y}}_{T+h}^{(j)\prime} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}_{T+h-i}^{(j)\prime}\mathbf{A}_i^{(j)} + \sum_{i=h}^{p} \mathbf{y}_{T+h-i}^{\prime}\mathbf{A}_i^{(j)} + \mathbf{x}_{T+h}^{\prime}\mathbf{C}^{(j)} + \mathbf{u}_{T+h}^{(j)\prime}.$$
Discarding the parameters yields $\left\{\widetilde{\mathbf{y}}_{T+1}^{(j)}, \ldots \widetilde{\mathbf{y}}_{T+H}^{(j)}\right\}_{j=B+1}^{B+R}$ as a sample of independent draws from the joint predictive distribution.

3.3 The steady state VAR

Villani (2009) observed that it is often easier to formulate a prior opinion about the steady state (unconditional mean) of a stationary VAR than about the dynamics. At the same time this is one feature of the data that a Minnesota type prior is silent about with the uninformative prior on the coefficients on deterministic variables.[8] This is, however, not surprising as the unconditional expectation is a highly nonlinear function of the parameters when the VAR is written as a linear regression model which makes it difficult to express any prior beliefs about the steady state. Let $\mathbf{A}\left(L\right) = \mathbf{I} - \mathbf{A}_1^{\prime}L - \ldots - \mathbf{A}_p^{\prime}L^p$ we can then write the stationary VAR (6) as
$$\mathbf{A}\left(L\right)\mathbf{y}_t = \mathbf{C}^{\prime}\mathbf{x}_t + \mathbf{u}_t.$$

[8]The initial observations prior could be used to incorporate information about the steady state in the prior formulation by replacing $\overline{y}_0$ with the expected steady state.

The unconditional expectation is the $E(\mathbf{y}_t) = \mu_t = \mathbf{A}^{-1}(L)\mathbf{C}'\mathbf{x}_t = \mathbf{\Lambda}\mathbf{x}_t$.[9] Given information about likely values for μ_t it is straightforward to formulate an informative prior for $\mathbf{\Lambda}$ but the implied prior for $\mathbf{C}$ is highly complicated. Instead Villani (2009) suggested writing the model in mean deviation form,

$$\mathbf{A}(L)(\mathbf{y}_t - \mathbf{\Lambda}\mathbf{x}_t) = \mathbf{u}_t. \tag{27}$$

This makes the model non-linear in parameters which complicates estimation but makes it easy to formulate a prior for all the parameters.

Let $\mathbf{\Gamma}_d' = \left(\mathbf{A}_1', \ldots \mathbf{A}_p'\right)$ represent the dynamics. Villani (2009) argued that there is no obvious connection between the steady state and the parameters governing the dynamics and suggested the prior

$$\pi(\mathbf{\Gamma}_d, \mathbf{\Lambda}, \mathbf{\Psi}) = \pi(\mathbf{\Gamma}_d)\pi(\mathbf{\Lambda})\pi(\mathbf{\Psi})$$

with $\pi(\mathbf{\Gamma}_d)$ and $\pi(\mathbf{\Lambda})$ normal,

$$\begin{aligned} \boldsymbol{\gamma}_d &\sim N\left(\underline{\gamma}_d, \underline{\mathbf{\Sigma}}_d\right), \\ \boldsymbol{\lambda} = \mathrm{vec}(\mathbf{\Lambda}) &\sim N(\underline{\boldsymbol{\lambda}}, \underline{\mathbf{\Sigma}}_\lambda) \end{aligned} \tag{28}$$

and a Jeffreys' prior (23) for $\mathbf{\Psi}$. Alternatively a proper inverse Wishart, $\mathbf{\Psi} \sim iW(\underline{\mathbf{S}}, \underline{v})$,for $\mathbf{\Psi}$ can be used. $\pi(\mathbf{\Gamma}_d)$ can be based on the prior beliefs in the Minnesota prior, variances as in (14) with prior means for the first own lag, $\underline{\gamma}_{jj}$ less than 1 indicating stationarity and existence of the steady state.

The joint posterior is, due to the nonlinearities, not a known distribution but Villani derived the full conditional posteriors for Λ, $\mathbf{\Gamma}_d$ and $\mathbf{\Psi}$ which can serve as the basis for a Gibbs sampler and MCMC based inference. To this end rewrite (27) as a nonlinear regression

$$\begin{aligned} \mathbf{y}_t' &= \mathbf{x}_t'\mathbf{\Lambda}' + \left[\mathbf{w}_t' - \mathbf{q}_t'(\mathbf{I}_p \otimes \mathbf{\Lambda}')\right]\mathbf{\Gamma}_d + \mathbf{u}_t' \\ \mathbf{Y} &= \mathbf{X}\mathbf{\Lambda}' + \left[\mathbf{W} - \mathbf{Q}(\mathbf{I}_p \otimes \mathbf{\Lambda}')\right]\mathbf{\Gamma}_d + \mathbf{U} \end{aligned}$$

with $\mathbf{w}_t' = \left(\mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-p}'\right)$, $\mathbf{q}_t' = \left(\mathbf{x}_{t-1}', \ldots, \mathbf{x}_{t-p}'\right)$. The full conditional posterior for $\mathbf{\Psi}$ is easy to derive and analogous to the normal-diffuse prior, form $\mathbf{U} = \mathbf{Y} - \mathbf{X}\Lambda' - \left[\mathbf{W} - \mathbf{Q}(\mathbf{I}_p \otimes \mathbf{\Lambda}')\right]\mathbf{\Gamma}_d$ and $\mathbf{S} = \mathbf{U}'\mathbf{U}$, the error sum of squares matrix conditional on $\mathbf{\Lambda}$ and $\mathbf{\Gamma}_d$. The conditional posterior for $\mathbf{\Psi}$ is then inverse Wishart

$$\mathbf{\Psi}|\mathbf{Y}_T, \mathbf{\Gamma}_d, \mathbf{\Lambda} \sim iW\left(\overline{\mathbf{S}}, \overline{v}\right) \tag{29}$$

with $\overline{\mathbf{S}} = \mathbf{S}$ and $\overline{v} = T$ for Jeffreys' prior and $\overline{\mathbf{S}} = \mathbf{S} + \underline{\mathbf{S}}$ and $\overline{v} = T + \underline{v}$ for the inverse Wishart prior.

For the full conditional posterior for $\mathbf{\Gamma}_d$ we can treat $\mathbf{\Lambda}$ as known and thus calculate $\mathbf{Y}_\Lambda = \mathbf{Y} - \mathbf{X}\mathbf{\Lambda}'$ and $\mathbf{W}_\Lambda = \left[\mathbf{W} - \mathbf{Q}(\mathbf{I}_p \otimes \mathbf{\Lambda}')\right]$. With these in hand we can write the model as $\mathbf{Y}_\Lambda = \mathbf{W}_\Lambda\mathbf{\Gamma}_d + \mathbf{U}$, a standard multivariate regression conditional on $\mathbf{\Lambda}$ and

[9] For simplicity we assume that $\mathbf{x}_t$ only consists of simple deterministic variables such as a constant, time trend and seasonal dummies.

$\boldsymbol{\Psi}$. This is analogous to the normal-diffuse prior (section 3.2.2) and the full conditional posterior for $\boldsymbol{\gamma}_d$ is normal

$$\begin{aligned}\boldsymbol{\gamma}_d|\mathbf{Y}_T, \boldsymbol{\Lambda}, \boldsymbol{\Psi} &\sim N\left(\overline{\boldsymbol{\gamma}}_d, \overline{\boldsymbol{\Sigma}}_d\right) \\ \overline{\boldsymbol{\Sigma}}_d &= \left(\underline{\boldsymbol{\Sigma}}_d^{-1} + \boldsymbol{\Psi}^{-1} \otimes \mathbf{W}_\Lambda' \mathbf{W}_\Lambda\right)^{-1} \\ \overline{\boldsymbol{\gamma}}_d &= \overline{\boldsymbol{\Sigma}}_d \left(\underline{\boldsymbol{\Sigma}}_d^{-1} \underline{\boldsymbol{\gamma}}_d + \text{vec}\left(\mathbf{W}_\Lambda' \mathbf{Y}_\Lambda \boldsymbol{\Psi}^{-1}\right)\right).\end{aligned} \tag{30}$$

The full conditional posterior for $\boldsymbol{\Lambda}$ is more complicated to derive and requires some matrix manipulations. Let $\mathbf{Y}_{\boldsymbol{\Gamma}} = \mathbf{Y} - \mathbf{W}\boldsymbol{\Gamma}_d$, $\mathbf{B} = (\mathbf{X}, -\mathbf{Q})$ and $\boldsymbol{\Theta}' = [\boldsymbol{\Lambda}, \boldsymbol{\Gamma}_d' (\mathbf{I}_p \otimes \boldsymbol{\Lambda})] = \left[\boldsymbol{\Lambda}, \mathbf{A}_1'\boldsymbol{\Lambda}, \ldots, \mathbf{A}_p'\boldsymbol{\Lambda}\right]$ the regression can then be written as

$$\begin{aligned}\mathbf{Y}_{\boldsymbol{\Gamma}} &= \mathbf{B}\boldsymbol{\Theta} + \mathbf{U} \\ \text{vec}\left(\mathbf{Y}_{\boldsymbol{\Gamma}}'\right) &= \text{vec}\left(\boldsymbol{\Theta}'\mathbf{B}'\right) + \text{vec}\left(\mathbf{U}'\right) \\ &= \left(\mathbf{B} \otimes \mathbf{I}\right) \text{vec}\left(\boldsymbol{\Theta}'\right) + \text{vec}\left(\mathbf{U}'\right) \\ &= \left(\mathbf{B} \otimes \mathbf{I}\right) \mathbf{F}\, \text{vec}\left(\boldsymbol{\Lambda}\right) + \text{vec}\left(\mathbf{U}'\right)\end{aligned}$$

a standard univariate regression with regression parameters $\boldsymbol{\lambda}$ for $\mathbf{F}' = [\mathbf{I}, \mathbf{I} \otimes \mathbf{A}_1, \mathbf{I} \otimes \mathbf{A}_p]$ and $\text{vec}(\mathbf{U}') \sim N\left(0, \mathbf{I}_T \otimes \boldsymbol{\Psi}\right)$. The usual Bayesian calculations yields a normal posterior for $\boldsymbol{\lambda}$ conditional on $\boldsymbol{\Gamma}_d$ and $\boldsymbol{\Psi}$,

$$\begin{aligned}\boldsymbol{\lambda}|\mathbf{Y}_T, \boldsymbol{\Gamma}_d, \boldsymbol{\Psi} &\sim N\left(\overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\Sigma}}_\lambda\right) \\ \overline{\boldsymbol{\Sigma}}_\lambda &= \left(\underline{\boldsymbol{\Sigma}}_\lambda^{-1} + \mathbf{F}'\left(\mathbf{B}'\mathbf{B} \otimes \boldsymbol{\Psi}^{-1}\right)\mathbf{F}\right)^{-1} \\ \overline{\boldsymbol{\lambda}} &= \overline{\boldsymbol{\Sigma}}_\lambda \left[\underline{\boldsymbol{\Sigma}}_\lambda^{-1} \underline{\boldsymbol{\lambda}} + \mathbf{F}'\left(\mathbf{B}'\mathbf{B} \otimes \boldsymbol{\Psi}^{-1}\right)\mathbf{F}\widehat{\boldsymbol{\lambda}}\right] \\ &= \overline{\boldsymbol{\Sigma}}_\lambda \left[\underline{\boldsymbol{\Sigma}}_\lambda^{-1} \underline{\boldsymbol{\lambda}} + \mathbf{F}' \text{vec}\left(\boldsymbol{\Psi}^{-1}\mathbf{Y}_{\boldsymbol{\Gamma}}'\mathbf{B}\right)\right]\end{aligned} \tag{31}$$

for $\widehat{\boldsymbol{\lambda}} = \left[\mathbf{F}'\left(\mathbf{B}'\mathbf{B} \otimes \boldsymbol{\Psi}^{-1}\right)\mathbf{F}\right]^{-1} \mathbf{F}'\left(\mathbf{B}' \otimes \boldsymbol{\Psi}^{-1}\right) \text{vec}\left(\mathbf{Y}_{\boldsymbol{\Gamma}}'\right)$ the GLS estimate.

Forecasting performance Villani (2009) conducts a small forecasting exercise where he compares the forecast performance of the steady-state prior to a standard BVAR with the Litterman prior and a standard VAR estimated with maximum likelihood. The focus is on modelling the Swedish economy and with Swedish GDP growth, inflation and interest rate, the corresponding foreign (world) variables and the exchange rate in trade weighted form included in the VAR models. The estimation period includes the Swedish financial crisis at the beginning of the 90-ties and the subsequent shift in monetary policy to inflation targeting. To accommodate this $\mathbf{x}_t$ includes a constant term and a dummy for the pre-crisis period. The prior on the constant terms in the steady-state VAR are thus centered on the perceived post-crisis steady state and the prior on the dummy variable coefficients reflects the higher pre-crisis inflation and interest rates and the belief that the crisis had no effect on long run GDP growth. For the dynamics, the prior on $\boldsymbol{\Gamma}_d$, Villani follows the Litterman prior with the addition of treating the foreign variables as exogenous, i.e. applying more aggressive shrinkage towards zero in the prior, and sets the prior mean of the first own lag to 0.9. The forecast performance is evaluated over the period 1999 to 2005. The steady-state VAR performs considerably better for the Swedish variables, confirming the intuition that it is useful to be informative about (changes in) the steady state.

Adolfson, Andersson, Linde, Villani and Vredin (2007) evaluates the forecast performance of two forecasting models in use at Sveriges Riksbank (the central bank of Sweden), a steady-state BVAR with the same variables as Villani (2009) similar prior set up, and the open economy DSGE model of Adolfson, Lasen, Lind and Villani (2008). The BVAR provides better forecasts of Swedish inflation up to 5 quarters ahead while the DSGE model has lower RMSE when forecasting 7 and 8 quarters ahead and both models improve on the official Riksbank forecast. The BVAR outperforms the DSGE model at all lead times when forecasting the interest rate and the forecast performance for GDP growth is almost identical but worse than the official Riksbank forecast except for lead times 6 through 8.

Österholm (2008*a*) forecasts the Swedish inflation and interest rate using a bivariate steady state BVAR and a univariate variant of the steady state BVAR, i.e. $\phi(L)(y_t - \alpha - \theta d_t)$, allowing for a shift at the time of the change in monetary policy regime. The forecasts are compared to forecasts from standard BVAR and Bayesian AR models with the dummy variable but without prior information about steady state. For inflation there is very little difference between the models whereas the steady state models do significantly better for the interest rate.

Beechey and Österholm (2010) forecasts the inflation rate for five inflation targeting countries, Australia, Canada, New Zealand, Sweden, the UK and the US using a univariate variant of the steady state VAR as in Österholm (2008*a*). The prior for θ is informative and centered on the target inflation rate with a diffuse prior for α and a Minnesota type lag decay on the autoregressive parameters in $\phi(L)$. As a comparison a standard AR model with the dummy d_t is also estimated using Bayesian and frequentist techniques, thus allowing for a shift in average inflation level but without adding information about the inflation target through a prior. The steady state AR improves on the forecasts of the other two models by a large amount for Australia, New Zealand and Sweden, less so for Canada and offer no improvement for the UK. The US is a special case with no officially announced inflation target, if a shift in the (unofficial) target is assumed in 1993 there is no improvement from the steady state model whereas there are substantial gains if the target is assumed constant.

Wright (2010) propose to anchor the steady state at the long run expectation of the variables as measured by survey responses. Specifically at each time point the prior mean of the steady state is set to the latest estimate from the Blue Chip survey. This is a convenient way of bringing in expectational data and Wright refers to this as a "democratic prior". Using VARs with monthly data on 10 variables Wright forecasts the US real GDP growth, GDP deflator, CPI inflation, industrial production growth three month yields and the unemployment rate at horizons 0 - 13. The VAR variants include one estimated by OLS, a normal-diffuse prior with Minnesota type prior beliefs and the democratic steady state prior with three different ways of specifying the prior mean on the first own lag, 0 for all variables, 0 for real variables and 0.85 for nominal variables and estimated from the survey data. The BVARs improve on the OLS estimated VAR and the democratic priors do better than the Minnesota prior with little difference between the alternative specification of the prior means. Wright also comparing the VAR forecasts with additional forecast devices for a subset of the variables. When the comparison is with survey estimates of short term expectations the differences are small with a few cases where a BVAR improves significantly on the survey estimates. Comparing the VAR

Algorithm 4 Gibbs sampler for the steady state prior

With the steady state prior 28 a Gibbs sampling algorithm follows immediately from the full conditional posteriors. Select starting values, $\boldsymbol{\gamma}_d^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$. For $j = 1, \ldots, B + R$

1. Generate $\boldsymbol{\Psi}^{(j)}$ from the full conditional posterior in (29) with $\overline{\mathbf{S}}$ evaluated at $\boldsymbol{\gamma}_d^{(j-1)}$ and $\boldsymbol{\lambda}^{(j-1)}$. Note that $\overline{\mathbf{S}}$ and v depends on the choice of prior for $\boldsymbol{\Psi}$.

2. Generate $\boldsymbol{\gamma}_d^{(j)}$ from the full conditional posterior in (30) with $\overline{\boldsymbol{\gamma}}_d$ and $\overline{\boldsymbol{\Sigma}}_d$ evaluated at $\boldsymbol{\Psi}^{(j)}$ and $\boldsymbol{\lambda}^{(j-1)}$.

3. Generate $\boldsymbol{\lambda}^{(j)}$ from the full conditional posterior in (31) with $\overline{\boldsymbol{\lambda}}$ and $\overline{\boldsymbol{\Sigma}}_\lambda$ evaluated at $\boldsymbol{\Psi}^{(j)}$ and $\boldsymbol{\gamma}^{(j)}$.

4. If $j > B$, generate $\mathbf{u}_{T+1}^{(j)}, \ldots, \mathbf{u}_{T+H}^{(j)}$ from $\mathbf{u}_t \sim N\left(0, \boldsymbol{\Psi}^{(j)}\right)$ and calculate recursively

$$\widetilde{\mathbf{y}}_{T+h}^{(j)\prime} = \mathbf{x}_{T+h}^{\prime}\boldsymbol{\Lambda}^{\prime(j)} + \sum_{+i=1}^{h-1}\left(\widetilde{\mathbf{y}}_{T+h-i}^{(j)\prime} - \mathbf{x}_{T+h-1}^{\prime}\boldsymbol{\Lambda}^{\prime(j)}\right)\mathbf{A}_i^{(j)} + \sum_{i=h}^{p}\left(\mathbf{y}_{T+h-i}^{\prime} - \mathbf{x}_{T+h-1}^{\prime}\boldsymbol{\Lambda}^{\prime(j)}\right)\mathbf{A}_i^{(j)} + \mathbf{u}_{T+h}^{(j)\prime}.$$

   The first B draws are discarded as burn-in. Discarding the parameters yields $\left\{\widetilde{\mathbf{y}}_{T+1}^{(j)}, \ldots \widetilde{\mathbf{y}}_{T+H}^{(j)}\right\}_{j=B+1}^{B+R}$ as a sample of independent draws from the joint predictive distribution.

forecasts with two time varying parameter models, an unobserved components stochastic volatility (UCSV) model and the TVP-VAR with stochastic volatility of Primiceri (2005) discussed in section 7, the VARs do better than the UCSV and the performance is similar to the TVP-VAR with a slight edge for the TVP-VAR.

Simulating from the posterior distribution With the full conditional posteriors in hand a straightforward Gibbs sampling scheme is available for sampling from the posterior and predictive distributions, see Algorithm 4. Villani reports that the Gibbs sampler convergences quickly to the posterior distribution but also notes that there is a possible issue of local nonidentification of $\boldsymbol{\Lambda}$ when there are unit roots or explosive roots in the autoregressive polynomial. This is only an issue for the convergence of the Gibbs sampler if the prior for $\boldsymbol{\Lambda}$ is uninformative and the posterior for $\boldsymbol{\Gamma}_d$ has non-negligible probability mass in the nonstationary region.

3.4 Model specification and choice of prior

Carriero, Clark and Marcellino (2011) conducts an impressive study of the many specification choices needed when formulating a BVAR for forecasting purposes. Their main application use monthly real time data on 18 US macroeconomic and financial variables. The baseline model is a BVAR with all 18 variables and 1 lag using the normal-Wishart

prior with "standard" choices for the hyperparameters. $\underline{\mathbf{\Omega}}$ and $\underline{\mathbf{S}}$ are specified as in (21) and (22) with $\underline{v} = m + 2$, $\pi_1 = \sqrt{0.1}$, $\pi_3 = 1$ and a diffuse prior on the constant term. The prior mean of the first own lag is set to 1 except when a variable is differenced in which case it is set to zero. The forecasts are constructed using the recursion

$$\widetilde{\mathbf{y}}'_{T+h} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}'_{T+h-i}\mathbf{A}_i + \sum_{i=h}^{p} \mathbf{y}'_{T+h-i}\mathbf{A}_i + \mathbf{x}'_{T+h}\mathbf{C} \tag{32}$$

with the parameters set at the posterior means.

Choice of hyperparameters and lag length: Alternatives considered are setting the hyperparameters by maximizing the marginal likelihood and using lag lengths 1-12. Increasing the lag length improves forecast performance for most of the variables but not for all. Choosing the lag length by maximizing the marginal likelihood leads to modest improvements for a majority of the variables with small losses for the other variables compared to the baseline. Choosing both hyperparameters and lag length by maximizing the marginal likelihood offers greater improvements than just maximizing with respect to one of them. The gains are on the whole relatively small and Carriero, Clark and Marcellino conclude that a lag length of 12 with $\pi_1 = \sqrt{0.1}$ is a simple and effective choice.

Multi-step forecasting: The forecast function (32) is non-linear in the parameters and using the posterior means of the parameters does not produce the means of the predictive distribution when $h > 1$. Alternatives considered are 1) simulating from the posterior distribution of the parameters and averaging over the forecasts and 2) using directs forecasts based on estimating models that are specific to each horizon

$$\mathbf{y}'_{t+h} = \sum_{i=1}^{p} \mathbf{y}'_{t-i}\mathbf{A}_i + \mathbf{x}'_t\mathbf{C} + \mathbf{e}'_{t+h}.$$

The gains from simulating the parameters is found to be negligible. Overall the differences between the iterated and direct forecasts are small but there are large gains from the direct forecast for some of the variables. This is presumably because the direct forecast is more robust to misspecification.

Cross-variable shrinkage and treatment of the error variance: The normal-Wishart prior forces a symmetric treatment of the variables whereas the original Litterman prior shrinks the parameters on "other" variables harder towards zero. On the other hand the normal-Wishart prior relaxes the assumption of a fixed and diagonal error variance matrix. Forecasting using the prior of Litterman as implemented in section 3.1, equation by equation estimation and two choices of π_2, $\sqrt{0.2}$ and $\sqrt{0.5}$ makes little difference except for the federal funds rate where the improvement is dramatic for for the shorter forecast horizons. The independent normal Wishart prior offers both the possibility to impose cross-variable shrinkage and an unrestricted error variance matrix. When comparing the forecast performance for the independent normal Wishart and Litterman priors the differences are very small with a slight edge for the Litterman prior.

Size of model: When comparing the forecast performance of the 18 variable VAR to a reduced model with 7 variables the larger model is found to forecast better. The gain from using the larger model is smaller with direct forecasts than iterated forecasts, again presumably due to the greater robustness against misspecification.

Levels or differences: A specification in levels can make use of any cointegration between the variables which should improve forecasts, on the other hand a specification in differences offers some robustness in the presence of structural breaks. The specification in differences improves on the levels specification, the root mean square error is on average 11% larger with the levels specification and the specification in differences has the lowest RMSE in 74% of the considered cases.

Carriero, Clark and Marcellino (2011) also conducts a robustness check using data from Canada, France and the UK using a reduced set of variables. Overall the conclusions from the US data is confirmed when using data from these three countries.

Summarizing their findings Carriero, Clark and Marcellino notes that "simple works" and recommends transforming variables to stationarity, using a relatively long lag length (12 with monthly data), the normal-Wishart prior and forecasts based on the posterior means of the parameters.

4 Structural VARs

The reduced form VAR is designed to capture the time series properties of the data and can, when coupled with suitable prior information, be an excellent forecasting device. The reduced form nature makes it difficult to incorporate economic insights into the prior. Take, for example, the "exogenous" variables prior in section 3.1.1. While it is tempting to think about this as implying exogeneity it is actually a statement about Granger causality. That, in a small open economy model, we do not expect the domestic variables to be useful for forecasting the variables representing the rest of the world. Restrictions on the variance-covariance matrix $\mathbf{\Psi}$ are needed in order to make claims about exogeneity. This brings us to structural or identified VAR-models that, by allowing limited structural interpretations of the parameters in the model, makes it possible to incorporate more economic insights in the model formulation. If done well this has the potential to improve the forecast performance of the model.

The basic structural VAR has the form

$$\begin{aligned} \mathbf{y}_t'\mathbf{\Lambda} &= \sum_{i=1}^{p} \mathbf{y}_{t-i}'\mathbf{B}_i + \mathbf{x}_t'\mathbf{D} + \mathbf{e}_t' \\ \mathbf{y}_t'\mathbf{\Lambda} &= \mathbf{z}_t'\mathbf{\Theta} + \mathbf{e}_t' \end{aligned} \tag{33}$$

where $\mathbf{\Lambda}$ is full rank, $\mathbf{\Theta}' = \left(\mathbf{B}_1', \ldots, \mathbf{B}_p', \mathbf{D}'\right)$ and $\mathbf{e}_t$ has a diagonal variance-covariance matrix. The relation with the reduced form (6) is straightforward, $\mathbf{\Gamma} = \mathbf{\Theta}\mathbf{\Lambda}^{-1}$, $\mathbf{A}_i = \mathbf{B}_i\mathbf{\Lambda}^{-1}$, $\mathbf{C} = \mathbf{D}\mathbf{\Lambda}^{-1}$, $\mathbf{u}_t = \mathbf{\Lambda}^{-\mathsf{T}}\mathbf{e}_t$ and $\mathbf{\Psi} = \mathbf{\Lambda}^{-\mathsf{T}} V(\mathbf{e}_t)\mathbf{\Lambda}^{-1}$.[10] The structural VAR (33) imposes restrictions on the form of the reduced form variance-covariance matrix $\mathbf{\Psi}$ but leaves the reduced form regression parameters $\mathbf{\Gamma}$ unrestricted since $\mathbf{\Lambda}$ is a full rank matrix unless there are additional restrictions on $\mathbf{\Theta}$. For simplicity we take $V(\mathbf{e}_t) = \mathbf{I}$, with $m(m+1)/2$ free parameters in the symmetric matrix $\mathbf{\Psi}$ this implies a simple order condition that $m(m-1)/2$ restrictions on $\mathbf{\Lambda}$ are needed for identification.[11] The simplest such scheme

[10] The SVAR can also be written as $\mathbf{y}_t' = \sum_{i=1}^{p} \mathbf{y}_{t-i}'\mathbf{A}_i + \mathbf{x}_t'\mathbf{C} + \mathbf{e}_t'\mathbf{L}$ with $\mathbf{L} = \mathbf{\Lambda}^{-\mathsf{T}}$ where the structure of $\mathbf{L}$ indicates which of the "identified" innovations e_{ti} has an immediate impact on y_{tj}.

[11] This is only a neccessary and not a sufficient condition for identification. Identification is discussed in more detail in section 4.3.

is to let $\mathbf{L} = \mathbf{\Lambda}^{-\mathsf{T}}$ be the (lower) triangular Cholesky decomposition of $\mathbf{\Psi} = \mathbf{L}\mathbf{L}'$. Subject to a normalization that the diagonal elements of $\mathbf{L}$ and $\mathbf{\Lambda}$ are positive this is a one-to-one mapping between $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ and yields exact identification without, in fact, imposing any restrictions on the reduced form. In the following we will frequently work with $\boldsymbol{\lambda} = \text{vec}(\mathbf{\Lambda})$ and $\boldsymbol{\lambda}_j$, column j of Λ, it is then important to keep in mind that these are subject to restrictions and that not all elements can vary freely.

The normalization is needed because the reduced form coefficients are left unchanged by reversing the sign of column j of $\mathbf{\Lambda}$ and $\mathbf{\Theta}$. The choice of normalization is, in general, not innocuous. Waggoner and Zha (2003*b*) demonstrate how an unfortunate choice of normalization can lead to misleading inference about $\mathbf{\Lambda}$ and impulse responses and give a rule for finding a good normalization. As our focus is on forecasting where the predictive distribution depends on the reduced form parameters we will largely ignore these issues.

4.1 "Unrestricted" triangular structural form

The structural form likelihood has the form

$$\begin{aligned}
L(\mathbf{Y}|\mathbf{\Theta}, \mathbf{\Lambda}) &\propto |\det \mathbf{\Lambda}|^T \exp\left\{-\frac{1}{2}\,\text{tr}\left[(\mathbf{Y}\mathbf{\Lambda} - \mathbf{Z}\mathbf{\Theta})(\mathbf{Y}\mathbf{\Lambda} - \mathbf{Z}\mathbf{\Theta})'\right]\right\} \\
&= |\det \mathbf{\Lambda}|^T \exp\left\{-\frac{1}{2}\left[\text{vec}(\mathbf{Y}\mathbf{\Lambda}) - (\mathbf{I}_m \otimes \mathbf{Z})\boldsymbol{\theta}\right]'\left[\text{vec}(\mathbf{Y}\mathbf{\Lambda}) - (\mathbf{I}_m \otimes \mathbf{Z})\boldsymbol{\theta}\right]\right\} \\
&= |\det \mathbf{\Lambda}|^T \exp\left\{-\frac{1}{2}\left[\text{vec}(\mathbf{Y}\mathbf{\Lambda}) - (\mathbf{I}_m \otimes \mathbf{Z})\widehat{\boldsymbol{\theta}}\right]'\left[\text{vec}(\mathbf{Y}\mathbf{\Lambda}) - (\mathbf{I}_m \otimes \mathbf{Z})\widehat{\boldsymbol{\theta}}\right]\right\} \\
&\times \exp\left[-\frac{1}{2}\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)'(\mathbf{I}_m \otimes \mathbf{Z}'\mathbf{Z})\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right)\right]
\end{aligned}$$

of a normal distribution for $\boldsymbol{\theta} = \text{vec}(\mathbf{\Theta})$ conditional on $\mathbf{\Lambda}$ with $\widehat{\boldsymbol{\theta}} = \text{vec}\left(\widehat{\mathbf{\Theta}}\right) = \text{vec}\left[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}\mathbf{\Lambda}\right]$. Sims and Zha (1998) suggested matching this by specifying a normal prior for $\boldsymbol{\theta}$ conditional on $\mathbf{\Lambda}$, $\boldsymbol{\theta}|\mathbf{\Lambda} \sim N(\text{vec}(\underline{\mathbf{M}}\mathbf{\Lambda}), \underline{\mathbf{\Sigma}}_\theta)$ with $\underline{\mathbf{M}}' = (\mathbf{I}_m, \mathbf{0})$ together with a marginal prior, $\pi(\mathbf{\Lambda})$ for $\mathbf{\Lambda}$. The choice of $\underline{\mathbf{M}}$ implies a prior mean for the reduced form parameters $\mathbf{\Gamma}$ that coincides with the univariate random walk of the Minnesota prior. The conditional posterior for $\boldsymbol{\theta}$ is then normal,

$$\begin{aligned}
\boldsymbol{\theta}|\mathbf{Y}_T, \mathbf{\Lambda} &\sim N\left(\overline{\boldsymbol{\theta}}, \overline{\mathbf{\Sigma}}_\theta\right) \\
\overline{\mathbf{\Sigma}}_\theta &= \left(\underline{\mathbf{\Sigma}}_\theta^{-1} + \mathbf{I}_m \otimes \mathbf{Z}'\mathbf{Z}\right)^{-1} \\
\overline{\boldsymbol{\theta}} &= \overline{\mathbf{\Sigma}}_\theta\left(\underline{\mathbf{\Sigma}}_\theta^{-1}\,\text{vec}(\underline{\mathbf{M}}\mathbf{\Lambda}) + \text{vec}(\mathbf{Z}'\mathbf{Y}\mathbf{\Lambda})\right).
\end{aligned}$$

Similar to the normal-diffuse prior this involves the inversion of the $mk \times mk$ matrix $\overline{\mathbf{\Sigma}}_\theta$ which can be computationally demanding. As noted by Sims and Zha (1998) this can be simplified considerably if $\underline{\mathbf{\Sigma}}_\theta$ is block diagonal with diagonal blocks $\underline{\mathbf{\Sigma}}_{\theta,j}$ corresponding to the equations. That is, there is independence between the priors for the different equations conditional on $\mathbf{\Lambda}$,

$$\boldsymbol{\theta}_j|\mathbf{\Lambda} \sim N\left(\underline{\mathbf{M}}\boldsymbol{\lambda}_j, \underline{\mathbf{\Sigma}}_{\theta,j}\right). \tag{34}$$

The inversion of a $mk \times mk$ matrix is then replaced by m inversions of $k \times k$ matrices as we solve for the posterior parameters equation by equation,

$$\begin{aligned}\boldsymbol{\theta}_j|\mathbf{Y}_T, \boldsymbol{\Lambda} &\sim N\left(\overline{\boldsymbol{\theta}}_j, \overline{\boldsymbol{\Sigma}}_{\theta,j}\right)\\ \overline{\boldsymbol{\Sigma}}_{\theta,j} &= \left(\underline{\boldsymbol{\Sigma}}_{\theta,j} + \mathbf{Z}'\mathbf{Z}\right)^{-1}\\ \overline{\boldsymbol{\theta}}_j &= \overline{\boldsymbol{\Sigma}}_{\theta,j}\left(\underline{\boldsymbol{\Sigma}}_{\theta,j}^{-1}\underline{\mathbf{M}}\boldsymbol{\lambda}_j + \mathbf{Z}'\mathbf{Y}\boldsymbol{\lambda}_j\right) = \overline{\mathbf{M}}_j\boldsymbol{\lambda}_j,\end{aligned}$$

and this brings us close to the computational convenience of the normal-Wishart prior.

A further simplification is available if the prior variance is the same for all equations, $\underline{\boldsymbol{\Sigma}}_{\theta,j} = \underset{\sim}{\boldsymbol{\Sigma}}_\theta$ and $\underline{\boldsymbol{\Sigma}}_\theta = \mathbf{I}_m \otimes \underset{\sim}{\boldsymbol{\Sigma}}_\theta$. The conditional posteriors for $\boldsymbol{\theta}_j$ then only differs in the conditional mean with

$$\begin{aligned}\boldsymbol{\theta}_j|\mathbf{Y}_T, \boldsymbol{\Lambda} &\sim N\left(\widetilde{\mathbf{M}}\boldsymbol{\lambda}_j, \widetilde{\boldsymbol{\Sigma}}_\theta\right)\\ \widetilde{\boldsymbol{\Sigma}}_\theta &= \left(\underset{\sim}{\boldsymbol{\Sigma}}_\theta^{-1} + \mathbf{Z}'\mathbf{Z}\right)^{-1}\\ \widetilde{\mathbf{M}} &= \widetilde{\boldsymbol{\Sigma}}_\theta\left(\underset{\sim}{\boldsymbol{\Sigma}}_\theta^{-1}\underline{\mathbf{M}} + \mathbf{Z}'\mathbf{Y}\right)\end{aligned} \tag{35}$$

which puts the computational requirements on par with the normal-Wishart prior for a reduced form VAR.

The posterior for $\boldsymbol{\Lambda}$ is more complicated. Integrating out $\boldsymbol{\Theta}$ from the joint posterior and keeping track of the extra terms from completing the square for $\boldsymbol{\theta}$ yields the marginal posterior

$$\begin{aligned}&\pi\left(\boldsymbol{\Lambda}|\mathbf{Y}_T\right) \propto \pi\left(\boldsymbol{\Lambda}\right)\left|\det \boldsymbol{\Lambda}\right|^T\\ &\times\exp\left\{-\frac{1}{2}\boldsymbol{\lambda}'\begin{pmatrix}\mathbf{I}_m \otimes \mathbf{Y}'\mathbf{Y} + \left(\mathbf{I}_m \otimes \underline{\mathbf{M}}\right)'\underline{\boldsymbol{\Sigma}}_\theta^{-1}\left(\mathbf{I}_m \otimes \underline{\mathbf{M}}\right)\\ -\left[\underline{\boldsymbol{\Sigma}}_\theta^{-1}\left(\mathbf{I}_m \otimes \underline{\mathbf{M}}\right) + \mathbf{I}_m \otimes \mathbf{Z}'\mathbf{Y}\right]'\overline{\boldsymbol{\Sigma}}_\theta^{-1}\left[\underline{\boldsymbol{\Sigma}}_\theta^{-1}\left(\mathbf{I}_m \otimes \underline{\mathbf{M}}\right) + \mathbf{I}_m \otimes \mathbf{Z}'\mathbf{Y}\right]\end{pmatrix}\boldsymbol{\lambda}\right\}.\end{aligned}$$

This is not a known distribution except in special cases. One such case arises under the prior $\boldsymbol{\theta}_j|\boldsymbol{\Lambda} \sim N\left(\underline{\mathbf{M}}\boldsymbol{\lambda}_j, \underset{\sim}{\boldsymbol{\Sigma}}_\theta\right)$ on $\boldsymbol{\theta}$ discussed above. The Kronecker structure of the prior variance-covariance matrix $\underline{\boldsymbol{\Sigma}}_\theta$ is inherited by the posterior variance-covariance and there is also a Kronecker structure in the posterior mean. The exponent in the posterior for $\boldsymbol{\Lambda}$ simplifies

$$\begin{aligned}\pi\left(\boldsymbol{\Lambda}|\mathbf{Y}_T\right) &\propto \pi\left(\boldsymbol{\Lambda}\right)\left|\det \boldsymbol{\Lambda}\right|^T \exp\left\{-\frac{1}{2}\boldsymbol{\lambda}'\left(\mathbf{I}_m \otimes \left[\mathbf{Y}'\mathbf{Y} + \underline{\mathbf{M}}'\underset{\sim}{\boldsymbol{\Sigma}}_\theta\underline{\mathbf{M}} - \widetilde{\mathbf{M}}'\widetilde{\boldsymbol{\Sigma}}_\theta^{-1}\widetilde{\mathbf{M}}\right]\right)\boldsymbol{\lambda}\right\}\\ &= \pi\left(\boldsymbol{\Lambda}\right)\left|\det \boldsymbol{\Lambda}\right|^T \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\left[\mathbf{Y}'\mathbf{Y} + \underline{\mathbf{M}}'\underset{\sim}{\boldsymbol{\Sigma}}_\theta\underline{\mathbf{M}} - \widetilde{\mathbf{M}}'\widetilde{\boldsymbol{\Sigma}}_\theta^{-1}\widetilde{\mathbf{M}}\right]\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\right)\right\}.\end{aligned}$$

Ignoring the prior, this is similar to a Wishart distribution for $\boldsymbol{\Psi}^{-1} = \Lambda\Lambda'$. It is, however, only a Wishart if the structure of $\boldsymbol{\Lambda}$ imposes no restrictions on $\boldsymbol{\Psi}$, e.g. if $\boldsymbol{\Lambda}$ is upper triangular with no other restrictions except the normalization. In this, special case, it is reasonable to specify an uninformative prior for $\boldsymbol{\Lambda}$, $\pi\left(\boldsymbol{\Lambda}\right) \propto 1$ and the implied posterior for $\boldsymbol{\Psi}^{-1}$ is Wishart,

$$\begin{aligned}\boldsymbol{\Psi}^{-1}|\mathbf{Y}_T &\sim W_m\left(\mathbf{S}^{-1}, T + m + 1\right)\\ \mathbf{S} &= \mathbf{Y}'\mathbf{Y} + \underline{\mathbf{M}}'\underset{\sim}{\boldsymbol{\Sigma}}_\theta\underline{\mathbf{M}} - \widetilde{\mathbf{M}}'\widetilde{\boldsymbol{\Sigma}}_\theta^{-1}\widetilde{\mathbf{M}}.\end{aligned} \tag{36}$$

A draw from the posterior of $\mathbf{\Lambda}$ can then be obtained by generating $\mathbf{\Psi}^{-1}$ from the Wishart distribution and solving for $\mathbf{\Lambda}$. In fact, if $\mathbf{\Lambda}$ is triangular it can be generated directly as the Bartlett decomposition of a Wishart distributed matrix. Sims and Zha (1998) and Zha (1999) suggests $\pi(\mathbf{\Lambda}) \propto |\det \mathbf{\Lambda}|^k$ as an uninformative improper prior. This is, however, in a slightly different context and working with the prior and posterior for $(\mathbf{\Lambda}, \mathbf{\Gamma})$ rather than $(\mathbf{\Lambda}, \mathbf{\Theta})$ and the factor $|\det \mathbf{\Lambda}|^k$ corresponds to the Jacobian when transforming from the $(\mathbf{\Lambda}, \mathbf{\Theta})$ parameterization to the $(\mathbf{\Lambda}, \mathbf{\Gamma})$ parameterization. Inference in the two parameterizations is thus equivalent with these two priors on $\mathbf{\Lambda}$ provided that the priors on $\mathbf{\Gamma}$ and $\mathbf{\Theta}$ are equivalent.

Specifying the prior The triangular SVAR is just a reparameterization of the reduced form VAR of section 3 and it is tempting to base the prior specification on the Minnesota prior. It should, however, be clear that it is not possible to mimic the Minnesota prior completely without losing the computational convenience as the transformation $\mathbf{\Theta} = \mathbf{\Gamma\Lambda}$ implies prior dependence between the columns of $\mathbf{\Theta}$. Sims and Zha (1998) proposed setting the prior standard deviations to

$$sd(\theta_{ij}) = \begin{cases} (\pi_1\pi_2) / (l^{\pi_3} s_r), \text{ lag } l \text{ of variable } r, \ i = (l-1)m + r \\ \pi_1\pi_4, \text{ deterministic variables}, \ i = mp+1, \ldots, k \end{cases} . \tag{37}$$

This is close to the Minnesota prior but differs in two aspects, there is no distinction between own lags and "other" variables since the choice of dependent variable in a simultaneous equation system is arbitrary and the scale factor s_j drops out since the error variances are normalized to 1. This leads to a common prior variance, $\underset{\sim}{\mathbf{\Sigma}}_\theta$, in (34) and the simplified posterior (35) in the spirit of the Minnesota prior.

The symmetric treatment of the structural form equations does, however, not imply a symmetric treatment of the reduced form equations. With $\mathbf{\Lambda}$ upper triangular we have $\boldsymbol{\gamma}_j = \sum_{i=1}^{j} \boldsymbol{\theta}_i \lambda^{ij}$ for λ^{ij} element i, j of $\mathbf{\Lambda}^{-1}$ and the ordering of the equations clearly matters for the implied prior on the reduced form. The unconditional prior expectation of $\mathbf{\Gamma}$ is $\underline{\mathbf{M}}$ and the random walk type prior with $\underline{m}_{ii} = 1$ can easily be modified to accommodate variables that are believed to be stationary by setting $\underline{m}_{ii}$ less than 1.

With $\mathbf{\Lambda}$ triangular a truly structural interpretation of the parameters is difficult and an uninformative prior, $\pi(\mathbf{\Lambda}) \propto 1$ seems appropriate.

Sampling from the posterior distribution With $\mathbf{\Lambda}$ triangular, the prior $\pi(\mathbf{\Lambda}) \propto 1$, $\boldsymbol{\theta}_j | \mathbf{\Lambda} \sim N\left(\underline{\mathbf{M}}\boldsymbol{\lambda}_j, \underset{\sim}{\mathbf{\Sigma}}_\theta\right)$ simulating from the posterior and predictive distributions using algorithm 5 is straightforward.

4.2 Homogenous restrictions on the structural form parameters

When the structure of $\mathbf{\Lambda}$ implies restrictions on $\mathbf{\Psi}$ the posterior becomes quite complicated irrespective of the choice of prior for $\mathbf{\Lambda}$. Sims and Zha (1998) proposes to use importance sampling, generating $\mathbf{\Lambda}$ from an approximation to the marginal posterior and $\mathbf{\Theta}$ or $\mathbf{\Gamma}$ conditional on $\mathbf{\Lambda}$ and Zha (1999) devices a scheme where blocks of equations can be treated independently and importance sampling can be used for each block. The block scheme should be more efficient as one high dimensional problem is replaced by several

Algorithm 5 Simulating the posterior and predictive distributions for a triangular SVAR

For the SVAR with $\mathbf{\Lambda}$ triangular and the prior (34) with $\underline{\mathbf{\Sigma}}_{\theta,j} = \underline{\mathbf{\Sigma}}_{\theta}$ for $\mathbf{\Theta}$ and an uninformative prior for $\mathbf{\Lambda}$, $\pi(\mathbf{\Lambda}) \propto 1$ draws from the posterior and predictive distributions can be obtained as follows

For $j = 1, \ldots R$

1. Generate $\mathbf{\Lambda}^{(j)}$ directly as the Bartlett decomposition of a draw form the marginal posterior (36).

2. For $i = 1, \ldots, m$ generate $\boldsymbol{\theta}_i^{(j)}$ from the conditional posterior (35).

3. Calculate the reduced form parameters $\mathbf{\Gamma}^{(j)} = \mathbf{\Theta}^{(j)}\mathbf{\Lambda}^{(j)}$.

4. Generate $\mathbf{u}_{T+1}^{(j)}, \ldots, \mathbf{u}_{T+H}^{(j)}$ from $\mathbf{u}_t \sim N\left(0, \mathbf{\Psi}^{(j)}\right)$ with $\mathbf{\Psi}^{(j)} = \left(\mathbf{\Lambda}^{(j)}\mathbf{\Lambda}^{(j)\prime}\right)^{-1}$ and calculate recursively

$$\widetilde{\mathbf{y}}_{T+h}^{(j)\prime} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}_{T+h-i}^{(j)\prime}\mathbf{A}_i^{(j)} + \sum_{i=h}^{p} \mathbf{y}_{T+h-i}^{\prime}\mathbf{A}_i^{(j)} + \mathbf{x}_{T+h}^{\prime}\mathbf{C}^{(j)} + \mathbf{u}_{T+h}^{(j)\prime}.$$

   Discarding the parameters yields $\left\{\widetilde{\mathbf{y}}_{T+1}^{(j)}, \ldots \widetilde{\mathbf{y}}_{T+H}^{(j)}\right\}_{j=1}^{R}$ as a sample of independent draws from the joint predictive distribution.

---

problems of smaller dimension. Nevertheless importance sampling has proven to be quite inefficient as it is difficult to find a good approximation to the marginal posterior of $\mathbf{\Lambda}$.

Waggoner and Zha (2003*a*) develops a Gibbs sampler for the marginal posterior of $\mathbf{\Lambda}$ in a setting allowing for exact restrictions and informative priors on both $\mathbf{\Lambda}$ and $\mathbf{\Theta}$. They consider homogenous restrictions on the parameters of one equation (column in $\mathbf{\Lambda}$ and $\mathbf{\Theta}$) of the form

$$\begin{aligned} \mathbf{Q}_j\boldsymbol{\lambda}_j &= \mathbf{0} \\ \mathbf{R}_j\boldsymbol{\theta}_j &= \mathbf{0} \end{aligned} \tag{38}$$

for $\mathbf{Q}_j$ a $(m - q_j) \times m$ matrix of rank $m - q_j$ and $\mathbf{R}_j$ a $(k - r_j) \times k$ matrix of rank $k - r_j$, i.e. there are $m - q_j$ restrictions on $\boldsymbol{\lambda}_j$ resulting in q_j free parameters and $m - r_j$ restrictions on $\boldsymbol{\theta}_j$ resulting in r_j free parameters, together with the normal prior suggested by Sims and Zha (1998) for the unrestricted parameters,

$$\begin{aligned} \boldsymbol{\lambda}_j &\sim N\left(\mathbf{0}, \underline{\mathbf{\Sigma}}_{\lambda,j}\right) \\ \boldsymbol{\theta}_j|\boldsymbol{\lambda}_j &\sim N\left(\underline{\mathbf{M}}_j\boldsymbol{\lambda}_j, \underline{\mathbf{\Sigma}}_{\theta,j}\right). \end{aligned} \tag{39}$$

To form a prior incorporating the restrictions Waggoner and Zha conditioned on the restrictions (38) in the prior (39). To this end let $\mathbf{U}_j$ and $\mathbf{V}_j$ be $m \times q_j$ and $k \times r_j$ orthonormal matrices satisfying $\mathbf{Q}_j\mathbf{U}_j = \mathbf{0}$ and $\mathbf{R}_j\mathbf{V}_j = \mathbf{0}$[12] if the restrictions hold there

[12] $\mathbf{U}_j$ and $\mathbf{V}_j$ form basis for the null spaces of $\mathbf{Q}_j$ and $\mathbf{R}_j$ and can be obtained from the QR decompositions of $\mathbf{Q}_j^{\prime}$ and $\mathbf{R}_j^{\prime}$ as follows. For $\mathbf{A}$ $m \times n$ $(m > n)$ of rank n we have $\mathbf{A} = \mathbf{QR}$ with $\mathbf{Q}$ a

must then be vectors $\mathbf{d}_j$ and $\mathbf{t}_j$ that satisfy $\boldsymbol{\lambda}_j = \mathbf{U}_j\mathbf{d}_j$ and $\boldsymbol{\theta}_j = \mathbf{V}_j\mathbf{t}_j$. $\mathbf{d}_j$ and $\mathbf{t}_j$ represents the free parameters and it is more convenient to work directly with them. The implied prior for $\mathbf{d}_j$ and $\mathbf{t}_j$ is obtained by Waggoner and Zha as

$$\mathbf{d}_j \sim N\left(\mathbf{0}, \underline{\boldsymbol{\Sigma}}_{d,j}\right), \ \mathbf{t}_j|\mathbf{d}_j \sim N\left(\underset{\sim}{\underline{\mathbf{M}}}_j\mathbf{d}_j, \underline{\boldsymbol{\Sigma}}_{t,j}\right) \tag{40}$$

with

$$\begin{aligned}
\underline{\boldsymbol{\Sigma}}_{t,j} &= \left(\mathbf{V}_j'\underline{\boldsymbol{\Sigma}}_{\theta,j}^{-1}\mathbf{V}_j\right)^{-1} \\
\underset{\sim}{\underline{\mathbf{M}}}_j &= \underline{\boldsymbol{\Sigma}}_{t,j}\mathbf{V}_j'\underline{\boldsymbol{\Sigma}}_{\theta,j}^{-1}\underline{\mathbf{M}}_j\mathbf{U}_j \\
\underline{\boldsymbol{\Sigma}}_{d,j} &= \left(\mathbf{U}_j'\underline{\boldsymbol{\Sigma}}_{\lambda,j}^{-1}\mathbf{U}_j + \mathbf{U}_j'\underline{\mathbf{M}}_j'\underline{\boldsymbol{\Sigma}}_{\theta,j}^{-1}\underline{\mathbf{M}}_j\mathbf{U}_j - \underset{\sim}{\underline{\mathbf{M}}}_j'\underline{\boldsymbol{\Sigma}}_{t,j}^{-1}\underset{\sim}{\underline{\mathbf{M}}}_j\right)^{-1}.
\end{aligned}$$

In the case that there are no restrictions on $\boldsymbol{\theta}_j$ we can take $\mathbf{V}_j = \mathbf{I}_k$ and the expressions simplify to $\underline{\boldsymbol{\Sigma}}_{t,j} = \underline{\boldsymbol{\Sigma}}_{\theta,j}$, $\underset{\sim}{\underline{\mathbf{M}}}_j = \underline{\mathbf{M}}_j\mathbf{U}_j$ and $\underline{\boldsymbol{\Sigma}}_{d,j} = \left(\mathbf{U}_j'\underline{\boldsymbol{\Sigma}}_{\lambda,j}^{-1}\mathbf{U}_j\right)^{-1}$.

Let $\mathbf{H} = (\mathbf{U}_1\mathbf{d}_1, \ldots, \mathbf{U}_m\mathbf{d}_m)$, the likelihood for $\mathbf{d}_j$ and $\mathbf{t}_j$, $j = 1, \ldots, m$ is then

$$\begin{aligned}
L\left(\mathbf{Y}|\mathbf{d}, \mathbf{t}\right) &\propto \left|\det \mathbf{H}\right|^T \exp\left\{-\frac{1}{2}\sum_{j=1}^{m}\left(\mathbf{Y}\mathbf{U}_j\mathbf{d}_j - \mathbf{Z}\mathbf{V}_j\mathbf{t}_j\right)'\left(\mathbf{Y}\mathbf{U}_j\mathbf{d}_j - \mathbf{Z}\mathbf{V}_j\mathbf{t}_j\right)\right\} \\
&= \left|\det \mathbf{H}\right|^T \exp\left\{-\frac{1}{2}\sum \mathbf{d}_j'\left(\mathbf{Y}\mathbf{U}_j - \mathbf{Z}\mathbf{V}_j\widehat{\mathbf{M}}_j\right)'\left(\mathbf{Y}\mathbf{U}_j - \mathbf{Z}\mathbf{V}_j\widehat{\mathbf{M}}_j\right)\mathbf{d}_j\right\} \\
&\times \exp\left[-\frac{1}{2}\sum\left(\mathbf{t}_j - \widehat{\mathbf{M}}_j\mathbf{d}_j\right)'\mathbf{V}_j'\mathbf{Z}'\mathbf{Z}\mathbf{V}_j\left(\mathbf{t}_j - \widehat{\mathbf{M}}_j\mathbf{d}_j\right)\right]
\end{aligned}$$

for $\widehat{\mathbf{M}}_j = \left(\mathbf{V}_j'\mathbf{Z}'\mathbf{Z}\mathbf{V}_j\right)^{-1}\mathbf{V}_j'\mathbf{Z}'\mathbf{Y}\mathbf{U}_j$. Multiplying with the prior (40), completing the square for $\mathbf{t}_j$ and collecting terms yields the joint posterior

$$p\left(\mathbf{d}, \mathbf{t}|\mathbf{Y}_T\right) \propto \left|\det \mathbf{H}\right|^T \exp\left\{-\frac{1}{2}\sum \mathbf{d}_j'\left[\mathbf{U}_j'\mathbf{Y}'\mathbf{Y}\mathbf{U}_j + \underset{\sim}{\underline{\mathbf{M}}}_j'\underline{\boldsymbol{\Sigma}}_{t,j}^{-1}\underset{\sim}{\underline{\mathbf{M}}}_j - \overline{\mathbf{M}}_j'\overline{\boldsymbol{\Sigma}}_{t,j}^{-1}\overline{\mathbf{M}}_j + \underline{\boldsymbol{\Sigma}}_{d,j}^{-1}\right]\mathbf{d}_j\right\} \tag{41}$$

$$\times \exp\left\{-\frac{1}{2}\sum\left(\mathbf{t}_j - \overline{\mathbf{M}}_j\mathbf{d}_j\right)'\overline{\boldsymbol{\Sigma}}_{t,j}^{-1}\left(\mathbf{t}_j - \overline{\mathbf{M}}_j\mathbf{d}_j\right)\right\}$$

with $\overline{\boldsymbol{\Sigma}}_{t,j} = \left(\underline{\boldsymbol{\Sigma}}_{t,j}^{-1} + \mathbf{V}_j'\mathbf{Z}'\mathbf{Z}\mathbf{V}_j\right)^{-1}$ and $\widetilde{\mathbf{M}}_j = \overline{\boldsymbol{\Sigma}}_{t,j}\left(\underline{\boldsymbol{\Sigma}}_{t,j}^{-1}\underset{\sim}{\underline{\mathbf{M}}}_j + \mathbf{V}_j'\mathbf{Z}'\mathbf{Y}\mathbf{U}_j\right)$. The conditional posterior for $\mathbf{t}_j$ is thus normal,

$$\mathbf{t}_j|\mathbf{Y}_T, \mathbf{d}_j \sim N\left(\widetilde{\mathbf{M}}_j\mathbf{d}_j, \overline{\boldsymbol{\Sigma}}_{t,j}\right),$$

and the conditional posteriors for $\mathbf{t}_j$ are independent conditional on $\mathbf{d}_1, \ldots, \mathbf{d}_m$. The marginal posterior for $\mathbf{d}_1, \ldots, \mathbf{d}_m$ is given by the first line of (41) where we must take account of $\mathbf{H}$ being a function of $\mathbf{d}_1, \ldots, \mathbf{d}_m$. Clearly this is not a known distribution even though it in part looks like a normal distribution with mean zero for $\mathbf{d}_j$.

$m \times m$ orthonormal matrix and $\mathbf{R} = \left(\mathbf{R}_1', \mathbf{0}_{n\times(m-n)}\right)'$ $m \times n$ with $\mathbf{R}_1$ upper triangular. We then have $\mathbf{A}'\mathbf{Q} = \mathbf{R}'\mathbf{Q}'\mathbf{Q} = \mathbf{R}'$. Partioning $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2)$ with $\mathbf{Q}_2$ $m \times (m-n)$ we see that $\mathbf{A}'\mathbf{Q}_2 = \mathbf{0}$ and $\mathbf{Q}_2$ is a basis for the null space of $\mathbf{A}'$. We can thus take $\mathbf{U}_j$ as the last q_j columns of the $\mathbf{Q}$ matrix of the QR decomposition of $\mathbf{Q}_j'$ and $\mathbf{V}_j$ as the last r_j columns of the $\mathbf{Q}$ matrix of the QR decomposition of $\mathbf{R}_j$.

Waggoner and Zha (2003*a*) develops a Gibbs sampling algorithm for the marginal posterior of $\mathbf{d}_1, \ldots, \mathbf{d}_m$ that operates on the posterior distributions for $\mathbf{d}_j$ conditional on $\mathbf{d}_i$, $i \neq j$, the set of full conditional posteriors. To this end let

$$\mathbf{S}_j^{-1} = \frac{1}{T}\left[\mathbf{U}_j'\mathbf{Y}'\mathbf{Y}\mathbf{U}_j + \underset{\sim}{\mathbf{M}}_j'\underline{\boldsymbol{\Sigma}}_{t,j}^{-1}\underset{\sim}{\mathbf{M}}_j - \overline{\mathbf{M}}_j'\overline{\boldsymbol{\Sigma}}_{t,j}^{-1}\overline{\mathbf{M}}_j + \underline{\boldsymbol{\Sigma}}_{d,j}^{-1}\right],$$

$\mathbf{T}_j\mathbf{T}_j' = \mathbf{S}_j$ and write $\mathbf{d}_j = \mathbf{T}_j\sum_{i=1}^{q_j}\beta_i\mathbf{w}_i = \mathbf{T}_j\mathbf{W}\boldsymbol{\beta}$ where $\mathbf{W}$ is a $q_j \times q_j$ orthonormal matrix with columns $\mathbf{w}_i$. The Jacobian for the change of variables from $\mathbf{d}_j$ to $\beta_1, \ldots, \beta_{q_j}$ is unity and the trick is to choose $\mathbf{W}$ in a clever way where $\mathbf{W}$ does not depend on $\mathbf{d}_j$. Let $\mathbf{w}$ be a $m \times 1$ vector that is orthogonal to each of $\mathbf{U}_i\mathbf{d}_i$, $i \neq j$,[13] set $\mathbf{w}_1 = \mathbf{T}_j'\mathbf{U}_j'\mathbf{w}/\sqrt{\mathbf{w}'\mathbf{U}\mathbf{T}\mathbf{T}_j'\mathbf{U}_j'\mathbf{w}}$ and $\mathbf{w}_i' = (w_{11}w_{i1}, \ldots, w_{i-1,1}w_{i1}, -c_{i-1}, 0, \ldots, 0)/\sqrt{c_{i-1}c_i}$ for $i = 2, \ldots, q_j$ where w_{i1} is element i of $\mathbf{w}_1$ and $c_i = \sum_{k=1}^{i} w_{k1}^2$. By construction $\mathbf{U}_j\mathbf{T}_j\mathbf{w}_1$ is linearly independent of $\mathbf{U}_i\mathbf{d}_i$, $i \neq j$, while $\mathbf{U}_j\mathbf{T}_j\mathbf{w}_i$, $i > 1$, are contained in the column space. Consequently

$$\begin{aligned}
\det\mathbf{H} &= \det\left(\mathbf{U}_1\mathbf{d}_1, \ldots, \mathbf{U}_{j-1}\mathbf{d}_{j-1}, \mathbf{U}_j\mathbf{T}_j\sum_{i=1}^{q_j}\beta_i\mathbf{w}_i, \mathbf{U}_{j+1}\mathbf{d}_{j+1}, \ldots, \mathbf{U}_m\mathbf{d}_m\right) \\
&= \sum_{i=1}^{q_j}\beta_i\det\left(\mathbf{U}_1\mathbf{d}_1, \ldots, \mathbf{U}_{j-1}\mathbf{d}_{j-1}, \mathbf{U}_j\mathbf{T}_j\mathbf{w}_i, \mathbf{U}_{j+1}\mathbf{d}_{j+1}, \ldots, \mathbf{U}_m\mathbf{d}_m\right) \\
&= \beta_1\det\left(\mathbf{U}_1\mathbf{d}_1, \ldots, \mathbf{U}_{j-1}\mathbf{d}_{j-1}, \mathbf{U}_j\mathbf{T}_j\mathbf{w}_1, \mathbf{U}_{j+1}\mathbf{d}_{j+1}, \ldots, \mathbf{U}_m\mathbf{d}_m\right) \propto \beta_1,
\end{aligned}$$

$\sum_{i=1}^{m}\mathbf{d}_i'\mathbf{S}_i^{-1}\mathbf{d}_i = \sum_{k=1}^{q_j}\beta_k^2 + \sum_{i\neq j}\mathbf{d}_i'\mathbf{S}_i^{-1}\mathbf{d}_i$ and the conditional posterior simplifies to

$$\begin{aligned}
p\left(\beta_1, \ldots, \beta_q|\mathbf{Y}_T, \mathbf{d}_{i\neq j}\right) &\propto |\beta_1|^T\exp\left[-\frac{T}{2}\sum_{k=1}^{q_j}\beta_k^2\right] \\
&= |\beta_1|^T\exp\left(-\frac{T\beta_1^2}{2}\right)\exp\left[-\frac{T}{2}\sum_{k=2}^{q_j}\beta_k^2\right]
\end{aligned}$$

the product of a Gamma distribution for $r = \beta_1^2$ and $q_j - 1$ independent $N(0, 1/T)$ distributions.[14]

Specifying the prior Waggoner and Zha (2003*a*) starts by specifying a prior (39) for the unrestricted structural form parameters, $\boldsymbol{\lambda}_j$ and $\boldsymbol{\theta}_j$, and conditions on the restrictions (38) in order to derive the prior (40) for the free parameters $\mathbf{d}_j$ and $\mathbf{t}_j$ in each equation. As a default, the conditional prior for $\boldsymbol{\theta}_j$ can be specified as in the unrestricted SVAR, e.g. prior variances in accordance with (37) and a choice of $\underline{\mathbf{M}}$ indicating if variables are believed to be non-stationary or not. Unlike the unrestricted SVAR there are no

[13] $\mathbf{w}$ can be obtained by solving the equation system $\mathbf{w}'\mathbf{U}_i\mathbf{d}_i = 0$, $i \neq j$. A practical method is to form the $m \times (m-1)$ matrix $\mathbf{A} = (\mathbf{U}_1\mathbf{d}_1, \ldots, \mathbf{U}_{j-1}\mathbf{d}_{j-1}, \mathbf{U}_{j+1}\mathbf{d}_{j+1}, \ldots, \mathbf{U}_m\mathbf{d}_m)$ and calculate the QR decomposition $\mathbf{A} = \mathbf{Q}\mathbf{R}$ and set $\mathbf{w} = \mathbf{q}_m$, the last column of $\mathbf{Q}$. Since the last row of $\mathbf{R}$ is zero and $\mathbf{Q}$ is orthonormal we have $\mathbf{w}'\mathbf{A} = \mathbf{w}'\mathbf{Q}\mathbf{R} = \mathbf{0}$.

[14] The distribution of $r = \beta_1^2$ is $f(r) \propto r^{T/2}\exp\left(-\frac{Tr}{2}\right)\left|\frac{\partial\beta_1}{\partial r}\right| = r^{(T+1)/2-1}\exp\left(-\frac{Tr}{2}\right)/2$ a $Gamma\left((T+1)/2, T/2\right)$ distribution.

computational gains from treating the equations symmetrically and the hard restrictions in (38) can easily be combined with "soft" restrictions on specific parameters.

It might be difficult to formulate economically meaningful priors on $\boldsymbol{\lambda}_j$ with the prior means fixed at zero as in (39) but one can, at least, be informative about the relative magnitude of coefficients by working with the prior variances. Imposing the restrictions (38) can have unexpected consequences on the prior if there is prior correlation between coefficients and the implied prior for $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ should be checked in this case.

Sampling form the posterior The sampler developed by Waggoner and Zha (2003*a*) is straightforward to implement and outlined in algorithm 6.

4.3 Identification under general restrictions

Following Rothenberg (1971) we say that a parameter point $(\mathbf{\Lambda}, \mathbf{\Theta})$ is identified if there is no other parameter point that is observationally equivalent, i.e. that they imply the same likelihood and hence the same reduced form parameters. Since the reduced form parameters are given by $\mathbf{\Gamma} = \mathbf{\Theta}\mathbf{\Lambda}^{-1}$, $\mathbf{\Psi} = (\mathbf{\Lambda}\mathbf{\Lambda}')^{-1}$ it is clear that $(\mathbf{\Lambda}, \mathbf{\Theta})$ and $\left(\widetilde{\mathbf{\Lambda}}, \widetilde{\mathbf{\Theta}}\right)$ are observationally equivalent if and only if there exists an orthonormal matrix $\mathbf{P}$ such that $\widetilde{\mathbf{\Lambda}} = \Lambda\mathbf{P}$ and $\widetilde{\mathbf{\Theta}} = \mathbf{\Theta}\mathbf{P}$. A SVAR is thus (globally) identified at $(\mathbf{\Lambda}, \mathbf{\Theta})$, subject to a set of restrictions, if the only orthonormal matrix for which both $(\mathbf{\Lambda}, \mathbf{\Theta})$ and $(\mathbf{\Lambda}\mathbf{P}, \mathbf{\Theta}\mathbf{P})$ satisfies the restrictions is the identity matrix.

Rubio-Ramirez, Waggoner and Zha (2010) considers general restrictions on the structural form parameters and obtains necessary and sufficient conditions for identification of SVARs. Let $f(\mathbf{\Lambda}, \mathbf{\Theta})$ be a $n \times m$ matrix valued function of the structural form parameters and $\mathbf{R}_j$ a $r_j \times n$ matrix of linear restrictions on column j of $f(\mathbf{\Lambda}, \mathbf{\Theta})$, i.e.

$$\mathbf{R}_j f(\mathbf{\Lambda}, \mathbf{\Theta})\, \mathbf{e}_j = \mathbf{0} \tag{42}$$

for $\mathbf{e}_j$ column j of the identity matrix $\mathbf{I}_m$ where the structural form parameters are subject to a normalization rule as in Waggoner and Zha (2003*b*). The order of the columns (equations) in $f()$ is arbitrary, as a convention the columns are ordered so that $r_1 \geq r_2 \geq \ldots \geq r_m$. Some regularity conditions on $f()$ are needed in order to state the identification results:

- Admissible: the restrictions are said to be admissible if $f(\mathbf{\Lambda}\mathbf{P}, \mathbf{\Theta}\mathbf{P}) = f(\mathbf{\Lambda}, \mathbf{\Theta})\mathbf{P}$ for $\mathbf{P}$ any orthonormal matrix.
- Regular: the restrictions are said to be regular if the domain U of $f()$ is an open set and f is continuously differentiable with f' of rank nm for all $(\mathbf{\Lambda}, \mathbf{\Theta}) \in U$.
- Strongly regular: the restrictions are said to be strongly regular if f is regular and $f(U)$ is dense in the set of $n \times m$ matrices.

Examples of admissible and strongly regular functions include the identity function $f(\mathbf{\Lambda}, \mathbf{\Theta}) = (\mathbf{\Lambda}', \mathbf{\Theta}')'$ for linear restrictions on the parameters, the short run impulse responses $f(\mathbf{\Lambda}, \mathbf{\Theta}) = \mathbf{\Lambda}^{-1}$, long run impulse responses $f(\mathbf{\Lambda}, \mathbf{\Theta}) = (\mathbf{\Lambda}' - \sum_{i=1}^{p} \mathbf{B}_i')^{-1}$ as well as intermediate impulse responses and combinations of these.

Algorithm 6 Gibbs sampler for restricted Structural form VARs

The sampler is based on the Gibbs sampler for $\mathbf{d}_1, \ldots, \mathbf{d}_m$. Once convergence is achieved draws of $\mathbf{t}_1, \ldots, \mathbf{t}_m$ can be obtained from the conditional posterior and the structural form parameters $\mathbf{\Lambda}$ and $\mathbf{\Theta}$ calculated. Start by precomputing the matrices, $\mathbf{U}_k$, $\mathbf{V}_k$, $\mathbf{S}_k$, and $\mathbf{T}_k$ for $k = 1, \ldots, m$ and select starting values $\mathbf{d}_2^{(0)}, \ldots, \mathbf{d}_m^{(0)}$.
For $j = 1, \ldots, B + R$

1. For $k = 1, \ldots, m$

   (a) Construct a vector $\mathbf{w}$ that is orthogonal to $\mathbf{U}_1\mathbf{d}_1^{(j)}, \ldots, \mathbf{U}_{k-1}\mathbf{d}_{k-1}^{(j)}, \mathbf{U}_{k+1}\mathbf{d}_{k+1}^{(j-1)}, \ldots, \mathbf{U}_m\mathbf{d}_m^{(j-1)}$ and calculate the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{q_k}$.

   (b) Generate r from a $G\left((T+1)/2, T/2\right)$ and u from a uniform $(0,1)$ distribution. Let $\beta_1 = -\sqrt{r}$ if $u \leq 0.5$ and set $\beta_1 = \sqrt{r}$ otherwise.

   (c) Generate $\beta_2, \ldots, \beta_{q_k}$ as independent $N\left(0, 1/T\right)$ random numbers

   (d) Calculate $\mathbf{d}_k^{(j)} = \mathbf{T}_k \sum_{i=1}^{q_k} \beta_i \mathbf{w}_i$

2. If $j > B$,

   (a) For $k = 1, \ldots, m$ generate $\mathbf{t}_k^{(j)}$ from the conditional posterior $\mathbf{t}_k | \mathbf{Y}_T, \mathbf{d}_k^{(j)} \sim N\left(\widetilde{\mathbf{M}}_k \mathbf{d}_k^{(j)}, \overline{\mathbf{\Sigma}}_{t,k}\right)$

   (b) Calculate the structural form parameters $\boldsymbol{\lambda}_k^{(j)} = \mathbf{U}_k \mathbf{d}_k^{(j)}$ and $\boldsymbol{\theta}_k^{(j)} = \mathbf{V}_k \mathbf{t}_k^{(j)}$ for $k = 1, \ldots, m$ and form the matrices $\Lambda^{(j)}$ and $\mathbf{\Theta}^{(j)}$.
   A normalization as in Waggoner and Zha (2003*b*) should be applied if the purpose is inference on the structural form parameters or impulse responses.

   (c) Calculate the reduced form parameters $\mathbf{\Gamma}^{(j)} = \mathbf{\Theta}^{(j)} \mathbf{\Lambda}^{(j)}$.

   (d) Generate $\mathbf{u}_{T+1}^{(j)}, \ldots, \mathbf{u}_{T+H}^{(j)}$ from $\mathbf{u}_t \sim N\left(0, \mathbf{\Psi}^{(j)}\right)$ with $\mathbf{\Psi}^{(j)} = \left(\mathbf{\Lambda}^{(j)} \mathbf{\Lambda}^{(j)\prime}\right)^{-1}$ and calculate recursively

$$\widetilde{\mathbf{y}}_{T+h}^{(j)\prime} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}_{T+h-i}^{(j)\prime} \mathbf{A}_i^{(j)} + \sum_{i=h}^{p} \mathbf{y}_{T+h-i}^{\prime} \mathbf{A}_i^{(j)} + \mathbf{x}_{T+h}^{\prime} \mathbf{C}^{(j)} + \mathbf{u}_{T+h}^{(j)\prime}.$$

   Discarding the parameters yields $\left\{\widetilde{\mathbf{y}}_{T+1}^{(j)}, \ldots \widetilde{\mathbf{y}}_{T+H}^{(j)}\right\}_{j=B}^{B+R}$ as a sample of independent draws from the joint predictive distribution.

Theorems 1 and 3 of Rubio-Ramirez et al. (2010) establishes that an SVAR with admissible and regular restrictions (42) is globally identified almost everywhere[15] if and only if the matrices $\mathbf{M}_j\left(f\left(\mathbf{\Lambda},\mathbf{\Theta}\right)\right), j=1,\ldots,m$, has rank m for some $(\mathbf{\Lambda},\mathbf{\Theta})$ that satisfies the restrictions. The $(r_j+j)\times m$ matrix $\mathbf{M}_j\left(f\left(\mathbf{\Lambda},\mathbf{\Theta}\right)\right)$ is given by

$$\mathbf{M}_j\left(f\left(\mathbf{\Lambda},\mathbf{\Theta}\right)\right)=\begin{pmatrix}\mathbf{R}_j f\left(\mathbf{\Lambda},\mathbf{\Theta}\right)\\ \mathbf{I}_j \quad \mathbf{0}_{j\times(m-j)}\end{pmatrix}$$

with $\mathbf{M}_j\left(f\left(\mathbf{\Lambda},\mathbf{\Theta}\right)\right)=\left(\mathbf{I}_j,\mathbf{0}\right)$ if there are no restrictions on column j of $f\left(\right)$.

Rubio-Ramirez et al. (2010, theorem 7) also develops a simple necessary and sufficient condition for exact identification.[16] A SVAR with admissible and strongly regular restrictions (42) is exactly identified if and only if $r_j=m-j$, $j=1,\ldots,m$. The restrictions must thus follow a pattern, a simple special case is when $\mathbf{\Lambda}$ is a triangular matrix with no other restrictions on the structural form parameters as in section 4.1.

To illustrate consider the following structure for the contemporaneous parameters in $\mathbf{\Lambda}$ (see Rubio-Ramirez et al. (2010, section 5.2) for motivation and additional details, note that our definition of $\mathbf{R}_j$ and consequently $\mathbf{M}_j$ differs in that we leave out redundant rows of these matrices)

$$\mathbf{\Lambda}=\begin{pmatrix}\lambda_{11} & \lambda_{12} & 0 & \lambda_{14} & \lambda_{15}\\ 0 & \lambda_{22} & 0 & \lambda_{24} & \lambda_{25}\\ 0 & 0 & \lambda_{33} & \lambda_{34} & \lambda_{35}\\ 0 & 0 & \lambda_{43} & \lambda_{44} & \lambda_{45}\\ 0 & 0 & 0 & 0 & \lambda_{55}\end{pmatrix}.$$

With no other restrictions, $f\left(\right)$ is just $f\left(\mathbf{\Lambda},\mathbf{\Theta}\right)=\mathbf{\Lambda}$ and the corresponding restriction matrices are

$$\mathbf{R}_1=\left(\mathbf{0}_{4\times1},\mathbf{I}_4\right),\ \mathbf{R}_2=\left(\mathbf{0}_{3\times2},\mathbf{I}_3\right),\ \mathbf{R}_3=\begin{pmatrix}\mathbf{e}_1'\\ \mathbf{e}_2'\\ \mathbf{e}_5'\end{pmatrix},\ \mathbf{R}_4=\mathbf{e}_5'.$$

We can immediately see that the SVAR would be exactly identified if there was one less zero restriction on the third column of $\mathbf{\Lambda}$, or – after reordering the equations – one less restriction on the second equation. As is, we need to verify that there exists a parameter point that satisfies the restrictions and for which all the $\mathbf{M}_j$ matrices has full rank in order to establish global identification. Multiplying $f\left(\mathbf{\Lambda},\mathbf{\Theta}\right)$ with $\mathbf{R}_j$ and filling out the

[15] "Globally identified almost everywhere" implies that we can check the rank condition at an arbitrary parameter point satisfying the restrictions.

[16] The SVAR is exactly identified if for all, except for a set of measure zero, reduced form parameters $(\mathbf{\Gamma},\mathbf{\Psi})$ there is a unique structural parameter point $(\mathbf{\Lambda},\mathbf{\Theta})$ satisfying $\mathbf{\Gamma}=\mathbf{\Theta}\mathbf{\Lambda}^{-1}$, $\mathbf{\Psi}=\left(\mathbf{\Lambda}\mathbf{\Lambda}'\right)^{-1}$.

bottom rows we have

$$\mathbf{M}_1 = \begin{pmatrix} 0 & \lambda_{22} & 0 & \lambda_{24} & \lambda_{25} \\ 0 & 0 & \lambda_{33} & \lambda_{34} & \lambda_{35} \\ 0 & 0 & \lambda_{43} & \lambda_{44} & \lambda_{45} \\ 0 & 0 & 0 & 0 & \lambda_{55} \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}, \; \mathbf{M}_2 = \begin{pmatrix} 0 & 0 & \lambda_{33} & \lambda_{34} & \lambda_{35} \\ 0 & 0 & \lambda_{43} & \lambda_{44} & \lambda_{45} \\ 0 & 0 & 0 & 0 & \lambda_{55} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{M}_3 = \begin{pmatrix} \lambda_{11} & \lambda_{12} & 0 & \lambda_{14} & \lambda_{15} \\ 0 & \lambda_{22} & 0 & \lambda_{24} & \lambda_{25} \\ 0 & 0 & 0 & 0 & \lambda_{55} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \; \mathbf{M}_4 = \begin{pmatrix} 0 & 0 & 0 & 0 & \lambda_{55} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \; \mathbf{M}_5 = \mathbf{I}_5$$

$\mathbf{M}_5$ is trivially full rank and $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{M}_4$ have, for example, full rank if λ_{22}, λ_{33}, λ_{44} and λ_{55} are non-zero with $\lambda_{24} = \lambda_{25} = \lambda_{34} = \lambda_{35} = \lambda_{43} = \lambda_{45} = 0$. $\mathbf{M}_3$ will have full rank if, in addition, λ_{14} is non-zero with $\lambda_{11} = \lambda_{12} = \lambda_{15} = 0$. The SVAR is thus identified.

Rubio-Ramirez et al. (2010, theorem 5) also gives an alternative condition for exact identification that is useful for posterior simulation. A SVAR is exactly identified if for almost every structural parameter point $(\mathbf{\Lambda}, \mathbf{\Theta}) \in U$ there is a unique orthonormal matrix $\mathbf{P}$ such that $(\mathbf{\Lambda P}, \mathbf{\Theta P})$ satisfies the restrictions. That is, we can simulate from the unrestricted SVAR with $\mathbf{\Lambda}$ triangular and transform the draws into parameter points that satisfy a set of exactly identifying restrictions provided that we can find the matrix $\mathbf{P}$. Normally a transformation with an orthonormal matrix will not affect the posterior distribution since the Jacobian is unity. In this case $\mathbf{P}$ is a function of $(\mathbf{\Lambda}, \mathbf{\Theta})$ and some care is needed to ensure that $(\mathbf{\Lambda P}, \mathbf{\Theta P})$ and $(\mathbf{\Lambda}, \mathbf{\Theta})$ has the same prior distribution. Rubio-Ramirez et al. (2010, theorem 5) verifies that the prior (39) with common variances for the equations, i.e., $\boldsymbol{\lambda} \sim N(\mathbf{0}, \mathbf{I}_m \otimes \underline{\mathbf{\Sigma}}_\lambda)$ and $\boldsymbol{\theta}|\boldsymbol{\lambda} \sim N(\mathrm{vec}(\underline{\mathbf{M}}\mathbf{\Lambda}), \mathbf{I}_m \otimes \underline{\mathbf{\Sigma}}_\theta)$, is – due to the Kronecker structure of the variances and the zero mean for $\boldsymbol{\lambda}$ – unaffected by a transformation with $\mathbf{P}$. It is also easy to see that this holds if the proper prior on $\boldsymbol{\lambda}$ is replaced by the improper prior $p(\boldsymbol{\lambda}) \propto |\det \mathbf{\Lambda}|^v$ since $\det(\mathbf{\Lambda P}) = \pm \det \mathbf{\Lambda}$. In addition, since an orthonormal transformation is observationally equivalent, it is possible to work with any prior on the reduced form parameters, sample from the posterior distribution of $(\mathbf{\Gamma}, \mathbf{\Psi})$, Cholesky decompose $\mathbf{\Psi} = \mathbf{L}\mathbf{L}'$ and transform to a triangular SVAR with $\mathbf{\Lambda} = \mathbf{L}^{-\mathrm{T}}$ and $\mathbf{\Theta} = \mathbf{\Gamma L}'$.

An important implication of the alternative condition for exact identification is that, modulo the effects of the prior specification, the predictive distribution from a simple triangular SVAR or reduced form VAR is identical to the predictive distribution from *any* exactly identified SVAR. That this is the case is readily seen by noting that the orthonormal transformation $(\mathbf{\Lambda P}, \mathbf{\Theta P})$ has no effect on reduced form parameters. For forecasting purposes it is thus, depending on the choice of prior specification, sufficient to work with the reduced form model or a triangular SVAR as long as the set of restrictions considered identify the SVAR exactly. Note that the triangular SVAR can be as in section 4.1 with $p(\boldsymbol{\lambda}) \propto |\det \mathbf{\Lambda}|^v$ or as in section 4.2 with $\boldsymbol{\lambda} \sim N(\mathbf{0}, \mathbf{I}_m \otimes \underline{\mathbf{\Sigma}}_\lambda)$ and restrictions $\mathbf{Q}_i \boldsymbol{\lambda}_j = \mathbf{0}$ yielding a triangular $\mathbf{\Lambda}$. For completeness the algorithm for finding the orthonormal transformation matrix $\mathbf{P}$ devised by Rubio-Ramirez et al. (2010) and generating random numbers from the posterior distribution of an exactly identified SVAR

Algorithm 7 Sampler for exactly identified SVARs

Depending on the choice of prior specification, generate reduced form parameters $\left(\mathbf{\Gamma}^{(j)}, \mathbf{\Lambda}^{(j)}\right)$ using one of the algorithms in section 3 or the structural form parameters $\left(\mathbf{\Lambda}^{(j)}, \mathbf{\Theta}^{(j)}\right)$ for a triangular SVAR using algorithm 5 with $p(\boldsymbol{\lambda}) \propto |\det \mathbf{\Lambda}|^{v}$ or algorithm 6 with $\boldsymbol{\lambda} \sim N(\mathbf{0}, \mathbf{I}_m \otimes \underline{\mathbf{\Sigma}}_{\lambda})$. In the former case calculate the Cholesky decomposition $\mathbf{\Psi}^{(j)} = \mathbf{L}\mathbf{L}'$ and set $\mathbf{\Lambda}^{(j)} = \mathbf{L}^{-\mathrm{T}}$, $\mathbf{\Theta}^{(j)} = \mathbf{\Gamma}^{(j)}\mathbf{L}'$. Discard any burn-in as needed.
For each draw from the original sampler

1. For $k = 1, \ldots, m$

   (a) Set $\widetilde{\mathbf{R}}_k = \begin{pmatrix} \mathbf{R}_k f\left(\mathbf{\Lambda}^{(j)}, \mathbf{\Theta}^{(j)}\right) \\ \mathbf{p}_1' \\ \ldots \\ \mathbf{p}_{j-1}' \end{pmatrix}$ $(\widetilde{\mathbf{R}}_1 = \mathbf{R}_1 f\left(\mathbf{\Lambda}^{(j)}, \mathbf{\Theta}^{(j)}\right))$

   (b) Solve for $\widetilde{\mathbf{R}}_j \mathbf{p}_j = 0$, for example by calculating the QR decomposition of $\widetilde{\mathbf{R}}_j' = \mathbf{Q}\mathbf{R}$ and setting $\mathbf{p}_j = \mathbf{q}_m$, the last column of $\mathbf{Q}$.

2. Form $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_m)$ and calculate structural form parameters $\widetilde{\mathbf{\Lambda}}^{(j)} = \mathbf{\Lambda}^{(j)}\mathbf{P}$ and $\widetilde{\mathbf{\Theta}}^{(j)} = \mathbf{\Theta}^{(j)}\mathbf{P}$ that satisfies the restrictions.

is given as algorithm 7.

Forecasting performance Österholm (2008*b*) use a structural BVAR to construct fan charts for Sweden and provides a limited forecast evaluation. The model contains nine variables, the foreign trade weighted GDP growth, inflation and interest rate, the Swedish unemployment rate, GDP growth, growth rate in wages, inflation, interest rate and the trade weighted exchange rate. The SVAR puts restriction on the $\mathbf{\Lambda}$ matrix which has a basic lower triangular structure with the additional restrictions $\lambda_{2,1} = \lambda_{3,1} = \lambda_{4,1} = \lambda_{4,2} = \lambda_{4,3} = \lambda_{5,4} = \lambda_{6,3} = \lambda_{7,4} = \lambda_{8,1} = \lambda_{8,5} = 0$ and allows $\lambda_{4,5}$, $\lambda_{5,7}$, $\lambda_{5,8}$, $\lambda_{6,7}$ and $\lambda_{8,9}$ to be non-zero. In the forecast evaluation a steady-state version of the SVAR and a naive random walk is also included. The steady state SVAR produces the best forecasts for Swedish inflation and forecast performance of the SVAR is somewhat better than the random walk. For GDP growth the steady state SVAR is again best followed by the random walk and the SVAR. The random walk provides the best forecasts for the inflation rate followed by the steady state SVAR and the SVAR.

5 Cointegration

Cointegration, that two or more non-stationary (integrated) variables can form a stationary linear combination and thus are tied together in the long run, is a powerful concept that is appealing both from an economic and forecasting standpoint. Economically this can be interpreted as a statement about long run equilibria and the information that the variables tend to move together in the long run should be useful for forecasting.

In order to explicitly model the cointegrating properties of the data we write the VAR (6) in error correction form

$$\Delta \mathbf{y}_t = \mathbf{\Pi} \mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \mathbf{B}_i \Delta \mathbf{y}_{t-i} + \mathbf{C}' \mathbf{x}_t + \mathbf{u}_t \tag{43}$$

where $\mathbf{\Pi} = -\left(\mathbf{I}_m - \sum_{i=1}^{p} \mathbf{A}_i'\right)$ and $\mathbf{B}_i = -\sum_{j=i+1}^{p} \mathbf{A}_j'$. If the m time series in $\mathbf{y}_t$ are stationary $\mathbf{\Pi}$ is a full rank matrix and if they all are non-stationary, integrated of order 1 or $I(1)$, but there is no cointegration $\mathbf{\Pi}$ will be a zero matrix. Here the focus is on the intermediate case where $\mathbf{\Pi}$ is of reduced rank $r < m$ and can be decomposed into two $m \times r$ matrices $\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ with $\boldsymbol{\beta}$ forming r cointegrating relations, $\boldsymbol{\beta}'\mathbf{y}_t$, or stationary linear combinations of the $I(1)$ variables in $\mathbf{y}_t$. The analysis of the cointegrated VECM (43) is complicated by the non-linear parameterization and, more fundamentally, by two issues of identification. Firstly, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are globally non-identified since any transformation with a full rank matrix $\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}\mathbf{P}$, $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}\mathbf{P}^{-\intercal}$ leaves $\mathbf{\Pi}$ unchanged. This is commonly solved by imposing a normalization $\boldsymbol{\beta}' = (\mathbf{I}_r, \boldsymbol{\beta}_*')$ but this can, as we shall see later, be problematic. Secondly, as noted by Kleibergen and van Dijk (1994), $\boldsymbol{\beta}$ is locally non-identified when $\boldsymbol{\alpha}$ has reduced rank, e.g. when $\boldsymbol{\alpha} = \mathbf{0}$. See Koop, Strachan, van Dijk and Villani (2006) for a more comprehensive review of Bayesian approaches to cointegration.

5.1 Priors on the cointegrating vectors

It is relatively straightforward to form prior opinions about the cointegrating vectors, for example in the form of specific relations between the variables that are suggested by economic theory. It is thus quite natural to formulate a prior on the cointegrating vectors, $\boldsymbol{\beta}$, and proceed with the analysis based on this prior. This leads to a relatively straightforward procedure for posterior inference but it is not without problems as it overlooks some of the fundamental issues in the analysis of the cointegrated VAR-model.

For a given number of cointegrating relations, r, the VECM can be rewritten in matrix form as

$$\begin{aligned} \mathbf{Y}_\Delta &= \mathbf{Y}_{-1}\boldsymbol{\beta}\boldsymbol{\alpha}' + \mathbf{X}\mathbf{\Theta} + \mathbf{U} \\ &= \mathbf{Z}_\beta \mathbf{\Gamma} + \mathbf{U} \end{aligned} \tag{44}$$

where $\mathbf{Y}_\Delta$ has rows $\Delta\mathbf{y}_t$, $\mathbf{Y}_{-1}$ rows $\mathbf{y}_{t-1}$, $\mathbf{X}$ rows $\left(\Delta\mathbf{y}_{t-1}', \ldots, \Delta\mathbf{y}_{t-p+1}', \mathbf{x}_t'\right)$, $\mathbf{Z}_\beta = (\mathbf{Y}_{-1}\boldsymbol{\beta}, \mathbf{X})$, and $\mathbf{\Theta}' = (\mathbf{B}_1, \ldots, \mathbf{B}_{p-1}, \mathbf{C}')$ and $\mathbf{\Gamma}' = (\boldsymbol{\alpha}, \mathbf{\Theta}')$ $k \times m$ and $(k+r) \times m$ parameter matrices. With $\mathbf{u}_t \sim N(\mathbf{0}, \mathbf{\Psi})$, and conditioning on $\boldsymbol{\beta}$, (44) is just a standard multivariate regression model and can be analyzed using one of the prior families for $(\mathbf{\Gamma}, \mathbf{\Psi})$ discussed in section 3.2 if there is prior independence between $\boldsymbol{\beta}$ and $(\mathbf{\Gamma}, \mathbf{\Psi})$. In particular, Geweke (1996*a*) specified an independent normal-Wishart type prior (section 3.2.2) for the parameters in (44) with $\mathbf{\Psi} \sim iW(\underline{\mathbf{S}}, \underline{v})$ and independent normal priors for $\mathrm{vec}(\boldsymbol{\alpha})$, $\mathrm{vec}(\boldsymbol{\beta})$ and $\mathrm{vec}(\mathbf{\Theta})$ with mean zero and variance-covariance matrix $\tau^{-2}\mathbf{I}$. Based on this he derived the full conditional posteriors and proposed a Gibbs sampling algorithm for exploring the joint posterior. Here we will consider a slightly more general prior specification,

$$\mathrm{vec}(\boldsymbol{\alpha}') \sim N\left(\mathrm{vec}(\underline{\boldsymbol{\alpha}}'), \underline{\mathbf{\Sigma}}_\alpha\right), \ \boldsymbol{\theta} = \mathrm{vec}(\mathbf{\Theta}) \sim N\left(\underline{\boldsymbol{\theta}}, \underline{\mathbf{\Sigma}}_\theta\right), \ \mathbf{\Psi} \sim iW(\underline{\mathbf{S}}, \underline{v}) \tag{45}$$

and an independent normal prior for the free elements of $\boldsymbol{\beta}$ to be specified later. Note that the prior for $\boldsymbol{\alpha}$ is specified in terms of the transpose of $\boldsymbol{\alpha}$. The full conditionals for $\boldsymbol{\Psi}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are obtained using standard results. We have

$$\boldsymbol{\Psi}|\mathbf{Y}_T, \boldsymbol{\beta}, \boldsymbol{\Gamma} \sim iW\left(\overline{\mathbf{S}}, \overline{v}\right), \ \overline{\mathbf{S}} = \underline{\mathbf{S}} + \left(\mathbf{Y} - \mathbf{Z}_\beta \boldsymbol{\Gamma}\right)'\left(\mathbf{Y} - \mathbf{Z}_\beta \boldsymbol{\Gamma}\right), \ \overline{v} = \underline{v} + T. \tag{46}$$

Combine the priors for $\boldsymbol{\alpha}$ and $\boldsymbol{\Theta}$ into a joint prior for $\boldsymbol{\Gamma} = \left(\boldsymbol{\alpha}, \boldsymbol{\Theta}'\right)'$, $\boldsymbol{\gamma} = \mathrm{vec}\left(\boldsymbol{\Gamma}\right) \sim N\left(\underline{\boldsymbol{\gamma}}, \underline{\boldsymbol{\Sigma}}_\gamma\right)$, we then have the full conditional posterior as

$$\boldsymbol{\gamma}|\mathbf{Y}_T, \boldsymbol{\beta}, \boldsymbol{\Psi} \sim N\left(\overline{\boldsymbol{\gamma}}, \overline{\boldsymbol{\Sigma}}_\gamma\right) \tag{47}$$

where $\overline{\boldsymbol{\gamma}}$ and $\overline{\boldsymbol{\Sigma}}_\gamma$ are given by (24) with $\mathbf{Z}$ and $\mathbf{Y}$ replaced by $\mathbf{Z}_\beta$ and $\mathbf{Y}_\Delta$.

The full conditional posterior for $\boldsymbol{\beta}$ is more complicated due to the nonlinear nature of the model and the need for at least r^2 identifying restrictions. A common identifying scheme is to set $\boldsymbol{\beta}' = \left(\mathbf{I}_r, \boldsymbol{\beta}_*'\right)$, more generally we can consider restrictions of the form $\mathbf{R}_i\boldsymbol{\beta}_i = \mathbf{r}_i$ on the individual cointegrating vectors (columns of $\boldsymbol{\beta}$). These restrictions are conveniently expressed as $\boldsymbol{\beta}_i = \mathbf{h}_i + \mathbf{H}_i\boldsymbol{\xi}_i$ where $\boldsymbol{\xi}_i$ corresponds to the free parameters in $\boldsymbol{\beta}_i$.[17] To derive the full conditional posterior for $\boldsymbol{\xi}$, we follow Villani (2001) and vectorize the model $\mathbf{Y}_\theta = \mathbf{Y}_\Delta - \mathbf{X}\boldsymbol{\Theta} = \mathbf{Y}_{-1}\boldsymbol{\beta}\boldsymbol{\alpha}' + \mathbf{U}$ to obtain

$$\begin{aligned}\mathbf{y}_\theta &= \left(\boldsymbol{\alpha} \otimes \mathbf{Y}_{-1}\right)\mathrm{vec}\left(\boldsymbol{\beta}\right) + \mathbf{u} = \mathbf{Y}_{-1,\alpha}\left(\mathbf{h} + \mathbf{H}\boldsymbol{\xi}\right) + \mathbf{u} \\ \mathbf{y}_{\theta,\alpha} &= \mathbf{y}_\theta - \mathbf{Y}_{-1,\alpha}\mathbf{h} = \mathbf{Y}_{-1,\alpha}\mathbf{H}\boldsymbol{\xi} + \mathbf{u}\end{aligned}$$

where $\mathbf{h} = \left(\mathbf{h}_1', \ldots \mathbf{h}_r'\right)'$, $\mathbf{H} = \mathrm{diag}\left(\mathbf{H}_i\right)$ and $\boldsymbol{\xi} = \left(\boldsymbol{\xi}_1', \ldots, \boldsymbol{\xi}_r'\right)$. With a normal prior on $\boldsymbol{\xi}$,

$$\boldsymbol{\xi} \sim N\left(\underline{\boldsymbol{\xi}}, \underline{\boldsymbol{\Sigma}}_\xi\right), \tag{48}$$

i.e. $\mathrm{vec}\left(\boldsymbol{\beta}\right) \sim N\left(\mathbf{h} + \mathbf{H}\underline{\boldsymbol{\xi}}, \mathbf{H}\underline{\boldsymbol{\Sigma}}_\xi\mathbf{H}'\right)$ which is a degenerate distribution due to the restrictions on $\boldsymbol{\beta}$, standard results yields the full conditional posterior as

$$\begin{aligned}\boldsymbol{\xi}|\mathbf{Y}_T, \boldsymbol{\Gamma}, \boldsymbol{\Psi} &\sim N\left(\overline{\boldsymbol{\xi}}, \overline{\boldsymbol{\Sigma}}_\xi\right) \\ \overline{\boldsymbol{\Sigma}}_\xi &= \left(\underline{\boldsymbol{\Sigma}}_\xi^{-1} + \mathbf{H}'\left(\boldsymbol{\alpha}'\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha} \otimes \mathbf{Y}_{-1}'\mathbf{Y}_{-1}\right)\mathbf{H}\right)^{-1} \\ \overline{\boldsymbol{\xi}} &= \overline{\boldsymbol{\Sigma}}_\xi\left[\underline{\boldsymbol{\Sigma}}_\xi^{-1}\underline{\boldsymbol{\xi}} + \mathbf{H}'\left(\boldsymbol{\alpha}'\boldsymbol{\Psi}^{-1} \otimes \mathbf{Y}_{-1}'\right)\mathbf{y}_{\theta,\alpha}\right].\end{aligned} \tag{49}$$

A Gibbs sampler can thus easily be constructed by sampling from the full conditional posteriors for $\boldsymbol{\xi}$ (and forming $\boldsymbol{\beta}$), $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$.

It is, as noted by among others Kleibergen and van Dijk (1994) and Geweke (1996*a*), crucial that a proper prior are used for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Without this the local nonidentification, as well as the possibility that the true cointegrating is less than r, will lead to an improper posterior.

It is also possible to work with a normal-Wishart type prior as in section 3.2.1, this is close to a conjugate prior and leads to some simplifications. Bauwens and Lubrano (1996) achieve similar simplifications with an uninformative Jeffreys type prior $\pi\left(\boldsymbol{\alpha}, \boldsymbol{\Theta}, \boldsymbol{\Psi}\right) \propto \left|\boldsymbol{\Psi}\right|^{-(m+1)/2}$ together with an independent prior on $\boldsymbol{\beta}$. Similar to Sugita (2002) we specify a normal-Wishart type prior,

$$\boldsymbol{\alpha}'|\boldsymbol{\Psi} \sim MN_{rm}\left(\underline{\boldsymbol{\alpha}}', \boldsymbol{\Psi}, \underline{\boldsymbol{\Omega}}_\alpha\right), \ \boldsymbol{\Theta}|\boldsymbol{\Psi} \sim MN_{km}\left(\underline{\boldsymbol{\Theta}}, \boldsymbol{\Psi}, \underline{\boldsymbol{\Omega}}_\theta\right) \text{ and } \boldsymbol{\Psi} \sim iW\left(\underline{\mathbf{S}}, \underline{v}\right) \tag{50}$$

[17] Set $\mathbf{h}_i = \mathbf{e}_i$, column i in the identity matrix $\mathbf{I}_m$, and $\mathbf{H}_i = \left(\mathbf{0}_{(m-r)\times r}, \mathbf{I}_{m-r}\right)'$ to obtain the "default" normalisation $\boldsymbol{\beta}' = \left(\mathbf{I}_r, \boldsymbol{\beta}_*'\right)$.

together with the independent normal prior (48) for the free elements $\boldsymbol{\xi}$ in $\boldsymbol{\beta}$.[18] It is convenient to combine the priors on $\boldsymbol{\alpha}$ and $\boldsymbol{\Theta}$ in a prior on $\boldsymbol{\Gamma}$,

$$\boldsymbol{\Gamma}|\boldsymbol{\Psi} \sim MN_{(r+k),m}\left(\underline{\boldsymbol{\Gamma}}, \boldsymbol{\Psi}, \underline{\boldsymbol{\Omega}}_{\gamma}\right), \ \underline{\boldsymbol{\Gamma}}' = \left(\underline{\boldsymbol{\alpha}}, \underline{\boldsymbol{\Theta}}'\right), \ \underline{\boldsymbol{\Omega}}_{\gamma} = \operatorname{diag}\left(\underline{\boldsymbol{\Omega}}_{\alpha}, \underline{\boldsymbol{\Omega}}_{\theta}\right). \tag{51}$$

With the prior for $\boldsymbol{\beta}$ independent of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$ it is clear that the posterior for $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$ conditional on $\boldsymbol{\beta}$ is of the normal-Wishart form,

$$\begin{aligned} \boldsymbol{\Gamma}|\mathbf{Y}_T, \boldsymbol{\beta}, \boldsymbol{\Psi} &\sim MN_{(r+k),m}\left(\overline{\boldsymbol{\Gamma}}, \boldsymbol{\Psi}, \overline{\boldsymbol{\Omega}}_{\gamma}\right), \\ \overline{\boldsymbol{\Omega}}_{\gamma}^{-1} &= \underline{\boldsymbol{\Omega}}_{\gamma}^{-1} + \mathbf{Z}_{\beta}'\mathbf{Z}_{\beta}, \ \overline{\boldsymbol{\Gamma}} = \overline{\boldsymbol{\Omega}}_{\gamma}\left(\underline{\boldsymbol{\Omega}}_{\gamma}^{-1}\underline{\boldsymbol{\Gamma}} + \mathbf{Z}_{\beta}'\mathbf{Y}_{\Delta}\right), \end{aligned} \tag{52}$$

$$\begin{aligned} \boldsymbol{\Psi}|\mathbf{Y}_T, \boldsymbol{\beta} &\sim iW\left(\overline{\mathbf{S}}, \overline{v}\right), \\ \overline{\mathbf{S}} &= \underline{\mathbf{S}} + \mathbf{S} + \left(\underline{\boldsymbol{\Gamma}} - \widehat{\boldsymbol{\Gamma}}\right)' \left(\underline{\boldsymbol{\Omega}}_{\gamma} + \left(\mathbf{Z}_{\beta}'\mathbf{Z}_{\beta}\right)^{-1}\right)^{-1} \left(\underline{\boldsymbol{\Gamma}} - \widehat{\boldsymbol{\Gamma}}\right), \ \overline{v} = T + \underline{v}. \end{aligned} \tag{53}$$

Peters et al. (2010) propose using a Metropolis within Gibbs MCMC scheme for sampling from the joint posterior distribution of $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$ with a Metropolis-Hastings step for $\boldsymbol{\beta}$. Peters et al. (2010) considered two random walk type proposals, a mixture proposal with one component designed to produce local moves and one component producing global moves and an adaptive proposal where variance-covariance matrix is continuously updated based on the previous output of the Markov chain.

The MCMC scheme of Peters et al. (2010) has the advantage that it does not rely on a specific form for the prior on $\boldsymbol{\beta}$. On the other hand, if we specify a normal prior for $\boldsymbol{\beta}$ (or $\boldsymbol{\xi}$) the derivations leading to the full conditional posterior (49) with the normal-Wishart type prior on $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$ and a standard Gibbs sampler is available in this case.

Specifying the prior beliefs The Minnesota prior is a useful starting point when thinking about the prior for $\boldsymbol{\Theta}$. Considering that $\boldsymbol{\Theta}$ contains $\mathbf{B}_1, \ldots, \mathbf{B}_{p-1}$ which are autoregressive coefficient matrices on the stationary first differences a reasonable choice is to set the prior means to zero and prior variances as in 14 with the modifications discussed in section 3.2.1 for the normal-Wishart type prior. Alternatively one can start with a Minnesota prior for the autoregressive parameters $\mathbf{A}_i$ in the reduced form VAR (6) and derive the prior mean and variance for $\mathbf{B}_j$ from the relation $\mathbf{B}_j = -\sum_{i=j+1}^{p} \mathbf{A}_i'$.

The priors for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (or $\boldsymbol{\xi}$) is a more delicate matter. Economic theory can in many cases suggest plausible cointegrating vectors and restrictions $\mathbf{R}_i\boldsymbol{\beta}_i = \mathbf{r}_i$. Care is however needed in the specification of the restrictions, at least one element of $\mathbf{r}_i$ must be nonzero otherwise the i^{th} cointegrating vector will only be identified up to an arbitrary scale factor. This in turn has implication for which variables are, in fact, cointegrated and fixing the coefficient of a variable that is not cointegrated to a non-zero value will clearly result in misleading inference. It is harder to form prior beliefs about the adjustment coefficients $\boldsymbol{\alpha}$ and a relatively uninformative prior with zero mean might be suitable. Note, however, that under prior independence we have $E(\boldsymbol{\Pi}) = E(\boldsymbol{\alpha})E(\boldsymbol{\beta}')$ and a prior mean of zero for $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ implies that $E(\boldsymbol{\Pi}) = \mathbf{0}$ which is at odds with the assumption that $\boldsymbol{\Pi}$ has rank $r > 0$.

[18] Sugita (2002) and later Peters, Kannan, Lassock and Mellen (2010) used a matric-variate normal for $\boldsymbol{\beta}_*$ in the normalization $\boldsymbol{\beta}' = (\mathbf{I}_r, \boldsymbol{\beta}_*')$ but there seem to be no particular advantage to the more restrictive Kronecker structure of the prior variance-covariance.

Algorithm 8 Gibbs sampler for VECM with a prior on $\boldsymbol{\beta}$

With the normal-Wishart type prior (50), (48) select starting values $\boldsymbol{\beta}^{(0)}$
For $j = 1, \ldots, B + R$

1. Generate $\boldsymbol{\Psi}^{(j)}$ from the conditional posterior $\boldsymbol{\Psi}|\mathbf{Y}_T, \boldsymbol{\beta}^{(j-1)} \sim iW\left(\overline{\mathbf{S}}, \overline{v}\right)$ in (53)

2. Generate $\boldsymbol{\Gamma}^{(j)}$ from the full conditional posterior $\boldsymbol{\Gamma}|\mathbf{Y}_T, \boldsymbol{\beta}^{(j-1)}, \boldsymbol{\Psi}^{(j)} \sim MN_{(r+k),m}\left(\overline{\boldsymbol{\Gamma}}, \boldsymbol{\Psi}^{(j)}, \overline{\boldsymbol{\Omega}}_{\gamma}\right)$ in (52)

3. Generate $\boldsymbol{\xi}^{(j)}$ from the full conditional posterior $\boldsymbol{\xi}|\mathbf{Y}_T, \boldsymbol{\Gamma}^{(j)}, \boldsymbol{\Psi}^{(j)} \sim N\left(\overline{\boldsymbol{\xi}}, \overline{\boldsymbol{\Sigma}}_{\xi}\right)$ in (49) and form $\boldsymbol{\beta}^{(j)}$

4. Generate $\mathbf{u}_{T+1}^{(j)}, \ldots, \mathbf{u}_{T+H}^{(j)}$ from $\mathbf{u}_t \sim N\left(0, \boldsymbol{\Psi}^{(j)}\right)$ and calculate recursively

$$\widetilde{\mathbf{y}}_{T+h}^{(j)\prime} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}_{T+h-i}^{(j)\prime} \mathbf{A}_i^{(j)} + \sum_{i=h}^{p} \mathbf{y}_{T+h-i}^{\prime} \mathbf{A}_i^{(j)} + \mathbf{x}_{T+h}^{\prime} \mathbf{C}^{(j)} + \mathbf{u}_{T+h}^{(j)\prime}$$

   using $\mathbf{A}_1 = \mathbf{I} + \mathbf{B}_1^{\prime} + \boldsymbol{\beta}\boldsymbol{\alpha}^{\prime}$, $\mathbf{A}_i = \mathbf{B}_i^{\prime} - \mathbf{B}_{i-1}^{\prime}$, $i = 2, \ldots, p-1$ and $\mathbf{A}_p = -\mathbf{B}_{p-1}^{\prime}$. Discarding the parameters yields $\left\{\widetilde{\mathbf{y}}_{T+1}^{(j)}, \ldots \widetilde{\mathbf{y}}_{T+H}^{(j)}\right\}_{j=B+1}^{B+R}$ as a sample from the joint predictive distribution.

Simulating from the posterior distribution The adaptive MCMC scheme is well described in Peters et al. (2010) and the Gibbs sampler is outlined in algorithm 8.

5.2 Priors on the cointegrating space

As alluded to above, the approach of working directly with the cointegrating vectors in $\boldsymbol{\beta}$ can be problematic. The issues are most easily discussed in relation to the linear normalization $\boldsymbol{\beta} = (\mathbf{I}_r, \boldsymbol{\beta}_*^{\prime})^{\prime}$ frequently used to identify the model. Partitioning $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\prime}, \boldsymbol{\beta}_2^{\prime})$ with $\boldsymbol{\beta}_1$ a $r \times r$ matrix, the normalization sets $\boldsymbol{\beta}_* = \boldsymbol{\beta}_2 \boldsymbol{\beta}_1^{-1}$. This has two implications, firstly the variables must be ordered in such a way that $\boldsymbol{\beta}_1$ is a full rank matrix, secondly $\boldsymbol{\beta}_*$ will have a fat tailed distribution, possibly with no posterior moments[19], unless a sufficiently informative prior on $\boldsymbol{\beta}_*$ is used.

The fundamental issue underlying this is the lack of identification of $\boldsymbol{\beta}$ and that only the space spanned by $\boldsymbol{\beta}$ (the *cointegrating space*, $\mathfrak{p} = \mathrm{sp}(\boldsymbol{\beta})$) is identified by the model. As argued by Villani (2000) we should then consider the prior for $\boldsymbol{\beta}$ in terms of this space. The columns of the rank r, $m \times r$ matrix $\boldsymbol{\beta}$ defines an r-dimensional hyperplane in $\mathbb{R}^m$, the space spanned by the columns of $\boldsymbol{\beta}$. Formally, the set of all such hyperplanes is known as the *Grassman manifold,* $\mathbb{G}_{r,m-r}$, and Villani (2005) shows that a uniform prior on $\mathbb{G}_{r,m-r}$ implies a matricvariate t distribution with r degrees of freedom on $\boldsymbol{\beta}_*$, $\boldsymbol{\beta}_* \sim$

[19]For the case of one cointegrating vector, $r = 1$, and an impoper prior on $\boldsymbol{\beta}$ Bauwens and Lubrano (1996) shows that the marginal posterior for $\boldsymbol{\beta}$ has finite moments up to the order of the number of overidentifying restrictions. That is, for $r = 1$ two restrictions in addition to normalizing the first element of $\boldsymbol{\beta}$ to 1 is needed for the posterior variance to exist.

$Mt_{m-r,r}(\mathbf{0},\mathbf{I},\mathbf{I},r)$ or $p(\boldsymbol{\beta}_*) \propto |\mathbf{I}_r + \boldsymbol{\beta}_*'\boldsymbol{\beta}_*|^{-m/2}$, when the linear normalization is used. This is quite different from using a uniform prior on $\boldsymbol{\beta}_*$. While the linear normalization implies a strong prior belief that the first r variables are included in cointegrating relations and that $\boldsymbol{\beta}_1$ is a full rank matrix, a uniform prior on $\boldsymbol{\beta}_*$ will, as pointed out by Strachan and Inder (2004), in fact put most of the prior mass on regions where $\boldsymbol{\beta}_1$ is (close to) non-invertible.

Departing from the uniform prior on the cointegrating spaces we consider a prior that is similar to the reference prior of Villani (2005),

$$\begin{aligned} \boldsymbol{\beta}_* &\sim Mt_{m-r,r}(\mathbf{0},\mathbf{I},\mathbf{I},r) \\ \boldsymbol{\alpha}'|\boldsymbol{\beta},\boldsymbol{\Psi} &\sim MN_{r,m}\left(\mathbf{0},\boldsymbol{\Psi},c^{-1}(\boldsymbol{\beta}'\boldsymbol{\beta})^{-1}\right) \\ \boldsymbol{\Theta}|\boldsymbol{\Psi} &\sim MN_{km}(\underline{\boldsymbol{\Theta}},\boldsymbol{\Psi},\underline{\boldsymbol{\Omega}}_\theta) \\ \boldsymbol{\Psi} &\sim iW(\underline{\mathbf{S}},\underline{v}). \end{aligned} \tag{54}$$

The main difference compared to Villani is that we use a normal prior for $\boldsymbol{\Theta}$ instead of the improper prior $p(\boldsymbol{\Theta}) \propto 1$. The results for the flat prior on $\boldsymbol{\Theta}$ can be obtained by setting $\underline{\boldsymbol{\Omega}}_\theta^{-1}$ to zero below.

The prior distribution for $\boldsymbol{\alpha}$ can be motivated by considering the prior for $\boldsymbol{\alpha}$ when $\boldsymbol{\beta}$ is orthonormal, i.e. $\boldsymbol{\beta}'\boldsymbol{\beta} = \mathbf{I}_r$.[20] Postmultiplying $\boldsymbol{\beta}$ with $(\boldsymbol{\beta}'\boldsymbol{\beta})^{-1/2}$ results in a set of orthogonalized cointegrating vectors $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}(\boldsymbol{\beta}'\boldsymbol{\beta})^{-1/2}$ and to keep $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ unchanged we need to adjust $\boldsymbol{\alpha}$ accordingly, $\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}(\boldsymbol{\beta}'\boldsymbol{\beta})^{1/2}$. It follows that the conditional distribution of $\widetilde{\boldsymbol{\alpha}}$ is $\widetilde{\boldsymbol{\alpha}}'|\widetilde{\boldsymbol{\beta}},\boldsymbol{\Psi} \sim MN_{r,m}(\mathbf{0},\boldsymbol{\Psi},c^{-1}\mathbf{I})$ or $\widetilde{\boldsymbol{\alpha}}_i|\boldsymbol{\Psi} \sim N(0,c^{-1}\boldsymbol{\Psi})$, $i = 1,\ldots,r$. Note that within the class of matricvariate normal priors $\widetilde{\boldsymbol{\alpha}}'|\widetilde{\boldsymbol{\beta}},\boldsymbol{\Psi} \sim MN_{r,m}(\boldsymbol{\mu},\boldsymbol{\Omega}_1,\boldsymbol{\Omega}_2)$ the only ones which are invariant to orthogonal rotations of $\widetilde{\boldsymbol{\beta}}$ are those with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Omega}_2 = c^{-1}\mathbf{I}$. The marginal prior for $\widetilde{\boldsymbol{\alpha}}'$ is matricvariate t and the prior variance-covariance is $V(\operatorname{vec}\widetilde{\boldsymbol{\alpha}}) = \frac{1}{c(\underline{v}-m-1)}\mathbf{I}_r \otimes \underline{\mathbf{S}}$ which clarifies the role of the scale factor c^{-1} in tuning the prior variance.

Writing the VECM as $\mathbf{Y}_{\alpha\beta} = \mathbf{Y}_\Delta - \mathbf{Y}_{-1}\boldsymbol{\beta}\boldsymbol{\alpha}' = \mathbf{X}\boldsymbol{\Theta} + \mathbf{U}$ we can derive the posterior distributions for $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ conditional on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. This requires that we keep track of the contribution from the joint prior for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ conditional on $\boldsymbol{\Psi}$,

$$\begin{aligned} p(\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{\Psi}) &\propto |\boldsymbol{\beta}'\boldsymbol{\beta}|^{m/2}|\boldsymbol{\Psi}|^{-r/2}\exp\left[-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha}(c\boldsymbol{\beta}'\boldsymbol{\beta})\boldsymbol{\alpha}'\right)\right] \times |\boldsymbol{\beta}'\boldsymbol{\beta}|^{-m/2} \\ &= |\boldsymbol{\Psi}|^{-r/2}\exp\left[-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha}(c\boldsymbol{\beta}'\boldsymbol{\beta})\boldsymbol{\alpha}'\right)\right] \end{aligned} \tag{55}$$

where we have used that $\boldsymbol{\beta}'\boldsymbol{\beta} = \mathbf{I}_r + \boldsymbol{\beta}_*'\boldsymbol{\beta}_*$ in the prior for $\boldsymbol{\beta}$.

Using standard results we obtain the conditional posteriors as

$$\begin{aligned} \boldsymbol{\Theta}|\mathbf{Y}_T,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\Psi} &\sim MN_{k,m}\left(\overline{\boldsymbol{\Theta}},\boldsymbol{\Psi},\overline{\boldsymbol{\Omega}}_\theta\right) \\ \overline{\boldsymbol{\Omega}}_\theta &= \left(\underline{\boldsymbol{\Omega}}_\theta^{-1} + \mathbf{X}'\mathbf{X}\right)^{-1},\ \overline{\boldsymbol{\Theta}} = \overline{\boldsymbol{\Omega}}_\theta\left(\underline{\boldsymbol{\Omega}}_\theta^{-1}\underline{\boldsymbol{\Theta}} + \mathbf{X}'\mathbf{Y}_{\alpha\beta}\right), \end{aligned} \tag{56}$$

[20] The restriction $\boldsymbol{\beta}'\boldsymbol{\beta} = \mathbf{I}_r$ is not sufficient to identify $\boldsymbol{\beta}$ since it can always be rotated to a new orthogonal matrix by postmultiplying with an $r \times r$ orthogonal matrix.

and

$$\begin{aligned}\boldsymbol{\Psi}|\mathbf{Y}_T,\boldsymbol{\alpha},\boldsymbol{\beta} &\sim iW\left(\overline{\mathbf{S}},\overline{v}\right),\ \overline{v}=T+\underline{v}+r \\ \overline{\mathbf{S}} &= \mathbf{S}+\underline{\mathbf{S}}+c\boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{\beta}\boldsymbol{\alpha}'+\underline{\boldsymbol{\Theta}}'\underline{\boldsymbol{\Omega}}_\theta^{-1}\underline{\boldsymbol{\Theta}}+\widehat{\boldsymbol{\Theta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\Theta}}-\overline{\boldsymbol{\Theta}}'\overline{\boldsymbol{\Omega}}_\theta^{-1}\overline{\boldsymbol{\Theta}}\end{aligned} \tag{57}$$

for $\mathbf{S}=\left(\mathbf{Y}_{\alpha\beta}-\mathbf{X}\widehat{\boldsymbol{\Theta}}\right)'\left(\mathbf{Y}_{\alpha\beta}-\mathbf{X}\widehat{\boldsymbol{\Theta}}\right)$ and $\widehat{\boldsymbol{\Theta}}=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}_{\alpha\beta}$.

To derive the full conditional posterior for $\boldsymbol{\alpha}$ write the VECM as $\mathbf{Y}_\theta=\mathbf{Y}_\Delta-\mathbf{X}\boldsymbol{\Theta}=\mathbf{Y}_{-1}\boldsymbol{\beta}\boldsymbol{\alpha}'+\mathbf{U}$ and apply the results for the normal-Wishart prior in section 3.2.1 to obtain the conditional posterior

$$\begin{aligned}\boldsymbol{\alpha}'|\boldsymbol{\beta},\boldsymbol{\Theta},\boldsymbol{\Psi} &\sim MN_{r,m}\left(\overline{\boldsymbol{\alpha}}',\boldsymbol{\Psi},\overline{\boldsymbol{\Omega}}_\alpha\right) \\ \overline{\boldsymbol{\Omega}}_\alpha &= \left[\boldsymbol{\beta}'\left(c\mathbf{I}_m+\mathbf{Y}_{-1}'\mathbf{Y}_{-1}\right)\boldsymbol{\beta}\right]^{-1} \\ \overline{\boldsymbol{\alpha}}' &= \overline{\boldsymbol{\Omega}}_\alpha\boldsymbol{\beta}'\mathbf{Y}_{-1}'\mathbf{Y}_\theta.\end{aligned} \tag{58}$$

For the full conditional posterior for $\boldsymbol{\beta}$ we note that the contribution from the prior can be rewritten as $\operatorname{tr}\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha}\left(c\boldsymbol{\beta}'\boldsymbol{\beta}\right)\boldsymbol{\alpha}'=\operatorname{tr}\boldsymbol{\beta}\left(c\boldsymbol{\alpha}'\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha}\right)\boldsymbol{\beta}'=\operatorname{tr}c\boldsymbol{\alpha}'\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha}+\operatorname{tr}\boldsymbol{\beta}_*\left(c\boldsymbol{\alpha}'\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha}\right)\boldsymbol{\beta}_*'$ with $\operatorname{tr}\boldsymbol{\beta}_*\left(c\boldsymbol{\alpha}'\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha}\right)\boldsymbol{\beta}_*'=\operatorname{vec}\left(\boldsymbol{\beta}_*\right)'\left(\boldsymbol{\alpha}'\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha}\otimes c\mathbf{I}_{m-r}\right)\operatorname{vec}\left(\boldsymbol{\beta}_*\right)$. That is, the prior for $\boldsymbol{\beta}_*$ conditional on $\boldsymbol{\alpha}$ and $\boldsymbol{\Psi}$ is matricvariate normal, $MN_{m-r,m}\left(\mathbf{0},\left(\boldsymbol{\alpha}'\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha}\right)^{-1},c^{-1}\mathbf{I}_{m-r}\right)$. Next rewrite $\mathbf{Y}_{-1}\boldsymbol{\beta}\boldsymbol{\alpha}'=\left(\mathbf{Y}_{-1,1}+\mathbf{Y}_{-1,2}\boldsymbol{\beta}_*\right)\boldsymbol{\alpha}'$ and vectorize the regression $\mathbf{Y}_{\theta\alpha}=\mathbf{Y}_\Delta-\mathbf{X}\boldsymbol{\Theta}-\mathbf{Y}_{-1,1}\boldsymbol{\alpha}'=\mathbf{Y}_{-1,2}\boldsymbol{\beta}_*\boldsymbol{\alpha}'+\mathbf{U}$,

$$\mathbf{y}_{\theta\alpha}=\left(\boldsymbol{\alpha}\otimes\mathbf{Y}_{-1,2}\right)\operatorname{vec}\left(\boldsymbol{\beta}_*\right)+\mathbf{u}.$$

The full conditional posterior for $\boldsymbol{\beta}_*$ is then obtained as

$$\begin{aligned}\boldsymbol{\beta}_*|\mathbf{Y}_T,\boldsymbol{\alpha},\boldsymbol{\Theta},\boldsymbol{\Psi} &\sim MN_{m-r,m}\left(\overline{\boldsymbol{\beta}}_*,\left(\boldsymbol{\alpha}'\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha}\right)^{-1},\overline{\boldsymbol{\Omega}}_\beta\right) \\ \overline{\boldsymbol{\Omega}}_\beta &= \left(c\mathbf{I}_{m-r}+\mathbf{Y}_{-1,2}'\mathbf{Y}_{-1,2}\right)^{-1} \\ \overline{\boldsymbol{\beta}}_* &= \overline{\boldsymbol{\Omega}}_\beta\mathbf{Y}_{-1,2}'\mathbf{Y}_{\theta\alpha}\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha}\left(\boldsymbol{\alpha}'\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha}\right)^{-1}.\end{aligned} \tag{59}$$

Using the improper prior $p\left(\boldsymbol{\Theta}\right)\propto 1$ instead of a normal prior as here Villani (2005) derived the conditional posteriors $p\left(\boldsymbol{\alpha}|\mathbf{Y}_T,\boldsymbol{\beta}\right)$ and $p\left(\boldsymbol{\beta}|\mathbf{Y}_T,\boldsymbol{\alpha}\right)$ as well as the marginal posterior for $\boldsymbol{\beta}$. Villani also shows that the posterior distribution of $\boldsymbol{\beta}_*$ has no finite moments as can be expected with the linear normalization $\boldsymbol{\beta}=\left(\mathbf{I}_r,\boldsymbol{\beta}_*'\right)'$.

The choice of normalization or identifying restrictions on $\boldsymbol{\beta}$ is thus crucial. Strachan (2003) proposes a data based normalization which restricts the length of the cointegrating vectors and ensures that the posterior for $\boldsymbol{\beta}$ is proper with finite moments but also implies that the prior for $\boldsymbol{\beta}$ or $\operatorname{sp}\left(\boldsymbol{\beta}\right)$ is data based. Strachan and Inder (2004) instead propose working with the normalization $\boldsymbol{\beta}'\boldsymbol{\beta}=\mathbf{I}_r$. While this is not sufficient to identify $\boldsymbol{\beta}$ it does restrict $\boldsymbol{\beta}$ to the set of semi-orthonormal $m\times r$ matrices, the Stiefel manifold $\mathbb{V}_{r,m}$.

There is often prior information about likely cointegrating vectors and Strachan and Inder (2004) proposes a convenient method for specifying an informative prior on the cointegrating space. First specify an $m\times r$ matrix with likely cointegrating vectors, e.g.

$$\mathbf{H}_g=\begin{pmatrix}1 & 0\\ -1 & 1\\ 0 & -1\end{pmatrix}$$

for $m = 3$ and $r = 2$. Since $\text{sp}(\mathbf{H}_g) = \text{sp}(\mathbf{H}_g\mathbf{P})$ for any full rank $r \times r$ matrix we can map $\mathbf{H}_g$ into $\mathbb{V}_{2,3}$ by the transformation $\mathbf{H} = \mathbf{H}_g \left(\mathbf{H}_g'\mathbf{H}_g\right)^{-1/2}$ and calculate the orthogonal complement $\mathbf{H}_\perp$, i.e. $\mathbf{H}_\perp \subseteq \mathbb{V}_{m-r,m}$ and $\mathbf{H}'\mathbf{H}_\perp = \mathbf{0}$.[21] That is,

$$\mathbf{H} = \begin{pmatrix} \sqrt{1/12} + \frac{1}{2} & \frac{1}{2} - \sqrt{1/12} \\ -\sqrt{1/3} & \sqrt{1/3} \\ \sqrt{1/12} - \frac{1}{2} & -\sqrt{1/12} - \frac{1}{2} \end{pmatrix}, \ \mathbf{H}_\perp = \begin{pmatrix} \sqrt{1/3} \\ \sqrt{1/3} \\ \sqrt{1/3} \end{pmatrix}.$$

Next consider the space spanned by the matrix $\mathbf{P}_\tau = \mathbf{H}\mathbf{H}' + \tau\mathbf{H}_\perp\mathbf{H}_\perp'$, for $\tau = 0$ this is $\text{sp}(\mathbf{H})$ and for $\tau = 1$ we have $\mathbf{P}_\tau = \mathbf{I}_m$ and $\text{sp}(\mathbf{P}_\tau) = \mathbb{R}^m$. Specifying the prior for $\boldsymbol{\beta}$ as a matrix angular central Gaussian distribution with parameter $\mathbf{P}_\tau$, $MACG(\mathbf{P}_\tau)$,

$$p(\boldsymbol{\beta}) \propto |\mathbf{P}_\tau|^{-r/2} \left|\boldsymbol{\beta}'\mathbf{P}_\tau^{-1}\boldsymbol{\beta}\right|^{-m/2}, \tag{60}$$

centers the distribution of $\mathfrak{p} = \text{sp}(\boldsymbol{\beta})$ on $\text{sp}(\mathbf{H})$ with the dispersion controlled by τ. For $\tau = 0$ we have a dogmatic prior that $\mathfrak{p} = \text{sp}(\mathbf{H})$ and for $\tau = 1$ a uniform prior on the Stiefel manifold which is equivalent to the uniform prior used by Villani (2005). By varying $\tau \in [0, 1]$ we can thus make the prior more or less informative.

Strachan and Inder (2004) propose using a Metropolis-Hastings sampler to evaluate the posterior distribution of the parameters under the prior (60) on $\boldsymbol{\beta}$ and a prior similar to (54) on the remaining parameters. Koop, León-González and Strachan (2010) propose a convenient Gibbs sampling scheme that depends on reparameterizing and in turn sample from a parameterization where $\boldsymbol{\alpha}$ is semiorthogonal and $\boldsymbol{\beta}$ unrestricted and a parameterization where $\boldsymbol{\beta}$ is semiorthogonal and $\boldsymbol{\alpha}$ is unrestricted. This solves the main computational difficulty with the semiorthogonal normalization where it is difficult generate $\boldsymbol{\beta}$ subject to the restriction $\boldsymbol{\beta}'\boldsymbol{\beta} = \mathbf{I}_r$.

Koop et al. (2010) develops the Gibbs sampling algorithm in a VECM without lags of $\Delta\mathbf{y}_t$ or deterministic variables. We will consider the more general model (44) and will thus need a prior for $\boldsymbol{\Theta}$ in addition to the prior on $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\Psi}$ specified by Koop et al. (2010), in addition to (60) we have

$$\begin{aligned} \boldsymbol{\alpha}'|\boldsymbol{\beta}, \boldsymbol{\Psi} &\sim MN_{r,m}\left(\mathbf{0}, \boldsymbol{\Psi}, c^{-1}\left(\boldsymbol{\beta}'\mathbf{P}_{1/\tau}\boldsymbol{\beta}\right)^{-1}\right) \\ \boldsymbol{\Theta}|\boldsymbol{\Psi} &\sim MN_{km}\left(\underline{\boldsymbol{\Theta}}, \boldsymbol{\Psi}, \underline{\boldsymbol{\Omega}}_\theta\right) \\ \boldsymbol{\Psi} &\sim iW\left(\underline{\mathbf{S}}, \underline{v}\right), \end{aligned} \tag{61}$$

where $\mathbf{P}_{1/\tau} = \mathbf{H}\mathbf{H}' + \tau^{-1}\mathbf{H}_\perp\mathbf{H}_\perp' = \mathbf{P}_\tau^{-1}$ a choice which facilitates the development of the Gibbs sampler. Koop et al. (2010) also considers the improper prior $p(\boldsymbol{\Psi}) \propto |\boldsymbol{\Psi}|^{-m/2}$, the results for this prior can be obtained by setting $\underline{\mathbf{S}} = \mathbf{0}$ and $\underline{v} = 0$ below.

The key to the Gibbs sampler of Koop et al. (2010) is the reparameterization

$$\boldsymbol{\alpha}\boldsymbol{\beta}' = \left(\boldsymbol{\alpha}\kappa^{-1}\right)(\boldsymbol{\beta}\kappa)' = \left[\boldsymbol{\alpha}\left(\boldsymbol{\alpha}'\boldsymbol{\alpha}\right)^{-1/2}\right]\left[\left(\boldsymbol{\alpha}'\boldsymbol{\alpha}\right)^{1/2}\boldsymbol{\beta}\right]' = \mathbf{A}\mathbf{B}'$$

[21] The square root matrix of a positive definite and symmetric matrix, such as $\mathbf{C} = \mathbf{H}_g'\mathbf{H}_g$, is unique and can be obtained from the spectral decomposition $\mathbf{C} = \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}'$ where $\mathbf{X}$ is the matrix of orthonormal eigenvectors and $\boldsymbol{\Lambda}$ has the eigenvalues, λ_i, on the diagonal. Consequently $\mathbf{C}^{1/2} = \mathbf{X}\boldsymbol{\Lambda}^{1/2}\mathbf{X}'$ with $\lambda_i^{1/2}$ as the diagonal elements of $\boldsymbol{\Lambda}^{1/2}$ and $\mathbf{H}'\mathbf{H} = \left(\mathbf{H}_g'\mathbf{H}_g\right)^{-1/2}\mathbf{H}_g'\mathbf{H}_g\left(\mathbf{H}_g'\mathbf{H}_g\right)^{-1/2} = \mathbf{I}$.

where $\mathbf{A}$ is semiorthogonal and $\mathbf{B}$ is unrestricted. For further reference note that the transformations from $\boldsymbol{\alpha}$ to $(\mathbf{A}, \kappa)$ and from $\mathbf{B}$ to $(\boldsymbol{\beta}, \kappa)$ where $\kappa = (\boldsymbol{\alpha}'\boldsymbol{\alpha})^{1/2} = (\mathbf{B}'\mathbf{B})^{1/2}$ is symmetric and positive definite are one-to-one, in addition $\boldsymbol{\beta} = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1/2}$ and $\boldsymbol{\alpha} = \mathbf{A}(\mathbf{B}'\mathbf{B})^{1/2}$. The implied priors for $\mathbf{A}$ and $\mathbf{B}$ can be obtained as (see Koop et al. (2010) for details)

$$\begin{aligned} \mathbf{B}|\mathbf{A}, \boldsymbol{\Psi} &\sim MN_{m,r}\left(\mathbf{0}, \left(\mathbf{A}'\boldsymbol{\Psi}^{-1}\mathbf{A}\right)^{-1}, c^{-1}\mathbf{P}_{\tau}\right) \\ \mathbf{A}|\boldsymbol{\Psi} &\sim MACG(\boldsymbol{\Psi}). \end{aligned} \tag{62}$$

The derivation of the full conditional posteriors proceed as above. The full conditional posterior for $\boldsymbol{\Theta}$ is matricvariate normal and given by (56) and the full conditional posterior for $\boldsymbol{\Psi}$ is inverse Wishart,

$$\begin{aligned} \boldsymbol{\Psi}|\mathbf{Y}_T, \boldsymbol{\alpha}, \boldsymbol{\beta} &\sim iW\left(\overline{\mathbf{S}}, \overline{v}\right), \ \overline{v} = T + \underline{v} + r \\ \overline{\mathbf{S}} &= \mathbf{S} + \underline{\mathbf{S}} + c\boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{P}_{1/\tau}\boldsymbol{\beta}\boldsymbol{\alpha}' + \underline{\boldsymbol{\Theta}}'\underline{\boldsymbol{\Omega}}_{\theta}^{-1}\underline{\boldsymbol{\Theta}} + \widehat{\boldsymbol{\Theta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\Theta}} - \overline{\boldsymbol{\Theta}}'\overline{\boldsymbol{\Omega}}_{\theta}^{-1}\overline{\boldsymbol{\Theta}} \end{aligned} \tag{63}$$

with $\mathbf{S}$ and $\widehat{\boldsymbol{\Theta}}$ as in the conditional posterior (57). The full conditional posterior for $\boldsymbol{\alpha}$ is a straightforward modification of the conditional posterior (58),

$$\begin{aligned} \boldsymbol{\alpha}'|\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\Psi} &\sim MN_{r,m}\left(\overline{\boldsymbol{\alpha}}', \boldsymbol{\Psi}, \overline{\boldsymbol{\Omega}}_{\alpha}\right) \\ \overline{\boldsymbol{\Omega}}_{\alpha} &= \left[\boldsymbol{\beta}'\left(c\mathbf{P}_{1/\tau} + \mathbf{Y}_{-1}'\mathbf{Y}_{-1}\right)\boldsymbol{\beta}\right]^{-1} \\ \overline{\boldsymbol{\alpha}}' &= \overline{\boldsymbol{\Omega}}_{\alpha}\boldsymbol{\beta}'\mathbf{Y}_{-1}'\mathbf{Y}_{\theta}. \end{aligned} \tag{64}$$

The full conditional posterior for $\boldsymbol{\beta}$ is complicated by the semiorthogonal normalization, instead the Gibbs sampling scheme of Koop et al. (2010) make use the full conditional posterior for the unrestricted parameter $\mathbf{B}$,

$$\begin{aligned} \mathbf{B}|\mathbf{Y}_T, \mathbf{A}, \boldsymbol{\Theta}, \boldsymbol{\Psi} &\sim MN_{m,r}\left(\overline{\mathbf{B}}, \left(\mathbf{A}'\boldsymbol{\Psi}^{-1}\mathbf{A}\right)^{-1}, \overline{\boldsymbol{\Omega}}_B\right) \\ \overline{\boldsymbol{\Omega}}_B &= \left(c\mathbf{P}_{\tau}^{-1} + \mathbf{Y}_{-1}'\mathbf{Y}_{-1}\right)^{-1} \\ \overline{\mathbf{B}} &= \overline{\boldsymbol{\Omega}}_B\mathbf{Y}_{-1}'\left(\mathbf{Y}_{\Delta} - \mathbf{X}\boldsymbol{\Theta}\right)\boldsymbol{\Psi}^{-1}\mathbf{A}\left(\mathbf{A}'\boldsymbol{\Psi}^{-1}\mathbf{A}\right)^{-1}. \end{aligned} \tag{65}$$

The idea behind the Gibbs sampler of Koop et al. (2010) is based on the fact that a draw $\boldsymbol{\alpha}^{(*)}$ from (64) is also a draw $\left(\mathbf{A}^{(*)}, \kappa^{(*)}\right)$ from $p(\mathbf{A}, \kappa|\mathbf{Y}_T, \boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\Psi})$, second drawing $\mathbf{B}^{(j)}$ from (65) yields a draw of $\left(\boldsymbol{\beta}^{(j)}, \kappa^{(j)}\right)$ from $p\left(\boldsymbol{\beta}, \kappa|\mathbf{Y}_T, \mathbf{A}^{(*)}, \boldsymbol{\Theta}, \boldsymbol{\Psi}\right)$ and we can map $\mathbf{A}^{(*)}$ and $\mathbf{B}^{(j)}$ into $\boldsymbol{\alpha}^{(j)} = \mathbf{A}^{(*)}\left(\mathbf{B}^{(j)\prime}\mathbf{B}^{(j)}\right)^{1/2}$, $\boldsymbol{\beta}^{(j)} = \mathbf{B}^{(j)}\left(\mathbf{B}^{(j)\prime}\mathbf{B}^{(j)}\right)^{-1/2}$ and the draws $\kappa^{(*)}$ and $\kappa^{(j)}$ are simply discarded.

In addition to the just identified case discussed here, Koop et al. (2010), also studies the case with overidentifying restrictions of the form $\boldsymbol{\beta}_i = \mathbf{H}_i\boldsymbol{\xi}_i$ considered in section 5.1 and provides a Gibbs sampling algorithm.

Specifying the prior beliefs The same considerations for the prior on $\boldsymbol{\Theta}$ holds here as in section 5.1. The informative prior (60) for $\boldsymbol{\beta}$ requires that we specify the tuning constant τ. Keeping in mind that $\tau = 0$ corresponds to a dogmatic prior and $\tau = 1$ corresponds to

a uniform prior on sp $(\boldsymbol{\beta})$, setting $\tau < 1/2$ seems appropriate. It is, however, difficult to develop intuition for τ and some sensitivity analysis is advisable. The choice of central location $\mathbf{H}$ (or $\mathbf{H}_g$) should obviously be based on economic intuition and theory or other subject specific information.

The prior distribution for $\boldsymbol{\alpha}$ requires a choice of the scale factor c, the prior is centered on zero with variance $V(\operatorname{vec}\boldsymbol{\alpha}) = \frac{1}{c(\underline{v}-m-1)}\left(\boldsymbol{\beta}'\mathbf{P}_{1/\tau}\boldsymbol{\beta}\right)^{-1} \otimes \underline{\mathbf{S}}$ conditional on $\boldsymbol{\beta}$. Evaluating this at the central location $\boldsymbol{\beta} = \mathbf{H}$ of the informative prior (60) or $\mathbf{P}_{1/\tau} = \mathbf{I}$ for the uniform prior yields $V(\operatorname{vec}\boldsymbol{\alpha}) = \frac{1}{c(\underline{v}-m-1)}\mathbf{I}_r \otimes \underline{\mathbf{S}}$ which can serve as a guide when choosing c. Alternatively, as suggested by Koop et al. (2010), a hierarchical prior structure can be used with inverse Gamma priors on c (and v) if the researcher prefers to treat them as unknown parameters.

Sampling from the posterior distribution The essential difference between the posterior distributions discussed here is the type of normalization imposed on $\boldsymbol{\beta}$. For the linear normalization $\boldsymbol{\beta} = (\mathbf{I}_r, \boldsymbol{\beta}_*')'$ and a flat prior on sp $(\boldsymbol{\beta})$ an adaption of the Gibbs sampler of Villani (2005) is given as Algorithm 9. For the orthogonal normalization $\boldsymbol{\beta}'\boldsymbol{\beta} = \mathbf{I}_r$ and a possibly informative prior on the cointegrating space an adaption of the Gibbs sampler of Koop et al. (2010) is given in algorithm 10. Note that the orthogonal normalization does not identify $\boldsymbol{\beta}$ and additional normalizations may be needed to obtain easily interpretable cointegrating vectors.

5.3 Determining the cointegrating rank

A simple approach to inference on the cointegrating rank, r, used in early Bayesian work, e.g. DeJong (1992) and Dorfman (1995), is to work with the reduced form VAR using one of the priors in section 3. In this context the posterior distribution of the cointegrating rank can be obtained from the posterior distribution of the roots of the autoregressive polynomial or the rank of the impact matrix $\boldsymbol{\Pi} = -\left(\mathbf{I}_m - \sum_{i=1}^{p}\mathbf{A}_i'\right)$. Sampling from the posterior distribution of the parameters it is straightforward to estimate the posterior distribution of the cointegrating rank by counting the number of roots of the AR-polynomial that are greater than, say, 0.99 or using the QR or SVD decompositions to find the rank of $\boldsymbol{\Pi}$ for each draw from the posterior.

While the unrestricted reduced form approach is straightforward it does not take account of the reduced rank restrictions on $\boldsymbol{\Pi}$ for $r < m$. Proper Bayesian model selection and model averaging account for this by basing the analysis on marginal likelihoods for models with different cointegrating rank r and calculating posterior probabilities $p(r|\mathbf{Y}_T)$ as in (5). This does, however, require some care to ensure that the marginal likelihood is well defined. As a minimum proper priors for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are needed as these change dimension with r and as a general rule at least mildly informative priors should be used for all parameters with the possible exception of $\boldsymbol{\Psi}$.

Using a prior on the cointegrating vectors as in section 5.1 with partially prespecified cointegrating vectors Villani (2001) approximates the log marginal likelihood with the Bayesian Information Criteria of Schwarz (1978). Sugita (2002) shows how to use the generalized Savage-Dickey density ratio of Verdinelli and Wasserman (1995) to compute the Bayes factors $BF_{i,0}$ comparing the model with $r = i$ against the model with $r = 0$ and the posterior probabilities $p(r|\mathbf{Y}_T)$ with the prior setup (50) together with a

Algorithm 9 Gibbs sampler for VECM with a prior on sp $(\boldsymbol{\beta})$ and linear normalization

With the prior (54), an uninformative prior on sp $(\boldsymbol{\beta})$ coupled with the linear normalization $\boldsymbol{\beta} = (\mathbf{I}_r, \boldsymbol{\beta}_*')'$, select starting values $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$.
For $j = 1, \ldots, B + R$

1. Generate $\boldsymbol{\Psi}^{(j)}$ from the full conditional posterior $\boldsymbol{\Psi}|\mathbf{Y}_T, \boldsymbol{\alpha}^{(j-1)}, \boldsymbol{\beta}^{(j-1)} \sim iW\left(\overline{\mathbf{S}}, \overline{v}\right)$ in (57)

2. Generate $\boldsymbol{\Theta}^{(j)}$ from the full conditional posterior $\boldsymbol{\Theta}|\mathbf{Y}_T, \boldsymbol{\alpha}^{(j-1)}, \boldsymbol{\beta}^{(j-1)}, \boldsymbol{\Psi}^{(j)} \sim MN_{k,m}\left(\overline{\boldsymbol{\Theta}}, \boldsymbol{\Psi}, \overline{\boldsymbol{\Omega}}_\theta\right)$ in (56)

3. Generate $\boldsymbol{\alpha}^{(j)}$ from the full conditional posterior $\boldsymbol{\alpha}'|\boldsymbol{\beta}^{(j-1)}, \boldsymbol{\Theta}^{(j)}, \boldsymbol{\Psi}^{(j)} \sim MN_{r,m}\left(\overline{\boldsymbol{\alpha}}', \boldsymbol{\Psi}^{(j)}, \overline{\boldsymbol{\Omega}}_\alpha\right)$ in (58)

4. Generate $\boldsymbol{\beta}^{(j)}$ from the full conditional posterior $\boldsymbol{\beta}_*|\mathbf{Y}_T, \boldsymbol{\alpha}^{(j)}, \boldsymbol{\Theta}^{(j)}, \boldsymbol{\Psi}^{(j)} \sim MN_{m-r,m}\left(\overline{\boldsymbol{\beta}}_*, \left(\boldsymbol{\alpha}^{(j)\prime}\left(\boldsymbol{\Psi}^{(j)}\right)^{-1}\boldsymbol{\alpha}^{(j)}\right)^{-1}, \overline{\boldsymbol{\Omega}}_\beta\right)$ in (59)

5. If $j > B$ generate $\mathbf{u}_{T+1}^{(j)}, \ldots, \mathbf{u}_{T+H}^{(j)}$ from $\mathbf{u}_t \sim N\left(0, \boldsymbol{\Psi}^{(j)}\right)$ and calculate recursively

$$\widetilde{\mathbf{y}}_{T+h}^{(j)\prime} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}_{T+h-i}^{(j)\prime}\mathbf{A}_i^{(j)} + \sum_{i=h}^{p} \mathbf{y}_{T+h-i}'\mathbf{A}_i^{(j)} + \mathbf{x}_{T+h}'\mathbf{C}^{(j)} + \mathbf{u}_{T+h}^{(j)\prime}$$

using $\mathbf{A}_1 = \mathbf{I} + \mathbf{B}_1' + \boldsymbol{\beta}\boldsymbol{\alpha}'$, $\mathbf{A}_i = \mathbf{B}_i' - \mathbf{B}_{i-1}'$, $i = 2, \ldots, p-1$ and $\mathbf{A}_p = -\mathbf{B}_{p-1}'$.
Discarding the parameters yields $\left\{\widetilde{\mathbf{y}}_{T+1}^{(j)}, \ldots \widetilde{\mathbf{y}}_{T+H}^{(j)}\right\}_{j=B+1}^{B+R}$ as a sample from the joint predictive distribution.

---

matricvariate normal prior for $\boldsymbol{\beta}_*$. Villani (2005) derives closed form expressions for the marginal likelihoods $m\left(\mathbf{Y}_T|r=0\right)$ and $m\left(\mathbf{Y}_T|r=m\right)$ under the uniform prior on $\mathbb{G}_{r,m}$ for sp $(\boldsymbol{\beta})$ and linear normalization, i.e. the prior (54) but with $\pi\left(\boldsymbol{\Theta}\right) \propto 1$, and uses the Chib (1995) method to estimate the marginal likelihood for intermediate cases from the Gibbs sampler output, e.g. Algorithm 9. With the orthogonal normalization and an uninformative prior on $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$, $\pi\left(\boldsymbol{\Theta}, \boldsymbol{\Psi}\right) \propto |\boldsymbol{\Psi}|^{-(m+1)/2}$ Strachan and Inder (2004) derives the posterior $p\left(\boldsymbol{\beta}, r|\mathbf{Y}_T\right)$ and use a Laplace approximation to integrate out $\boldsymbol{\beta}$ to obtain the posterior distribution of the cointegrating rank. Sugita (2009) studies rank selection in a Monte Carlo experiment where the marginal likelihood is approximated with BIC or estimated using the Chib method and finds that BIC performs well when $T \geq 100$ and Chib's method requires considerably larger sample sizes, $T > 500$, to perform well.

Forecasting performance Villani (2001) forecasts the Swedish inflation rate with several versions of a 7 variable VECM with the Swedish GDP, CPI, interest rate, trade weighted exchange and "foreign" GDP, price level and interest rate. Villani considers several theory based cointegrating relations which are all rejected by the data in favour of a model with cointegrating rank 3 and unrestricted cointegrating relations. Nonetheless,

Algorithm 10 Gibbs sampler for VECM with a prior on sp $(\boldsymbol{\beta})$ and orthogonal normalization

With the orthogonal normalization $\boldsymbol{\beta}'\boldsymbol{\beta} = \mathbf{I}_r$ and the informative prior (60) and (61) the Gibbs sampler of Koop et al. (2010) is applicable. Select starting values $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$. For $j = 1, \ldots, B + R$

1. Generate $\boldsymbol{\Psi}^{(j)}$ from the conditional posterior $\boldsymbol{\Psi}|\mathbf{Y}_T, \boldsymbol{\alpha}^{(j-1)}, \boldsymbol{\beta}^{(j-1)} \sim iW\left(\overline{\mathbf{S}}, \overline{v}\right)$ in (63)

2. Generate $\boldsymbol{\Theta}^{(j)}$ from the full conditional posterior $\boldsymbol{\Theta}|\mathbf{Y}_T, \boldsymbol{\alpha}^{(j-1)}, \boldsymbol{\beta}^{(j-1)}, \boldsymbol{\Psi}^{(j)} \sim MN_{k,m}\left(\overline{\boldsymbol{\Theta}}, \boldsymbol{\Psi}, \overline{\boldsymbol{\Omega}}_\theta\right)$ in (56)

3. Generate $\boldsymbol{\alpha}^{(*)}$ from the full conditional posterior $\boldsymbol{\alpha}'|\boldsymbol{\beta}^{(j-1)}, \boldsymbol{\Theta}^{(j)}, \boldsymbol{\Psi}^{(j)} \sim MN_{r,m}\left(\overline{\boldsymbol{\alpha}}', \boldsymbol{\Psi}^{(j)}, \overline{\boldsymbol{\Omega}}_\alpha\right)$ in (64) and calculate $\mathbf{A}^{(*)} = \boldsymbol{\alpha}^{(*)}\left(\boldsymbol{\alpha}^{(*)\prime}\boldsymbol{\alpha}^{(*)}\right)^{-1/2}$

4. Generate $\mathbf{B}^{(j)}$ from the conditional posterior $\mathbf{B}|\mathbf{Y}_T, \mathbf{A}^{(*)}, \boldsymbol{\Theta}^{(j)}, \boldsymbol{\Psi}^{(j)} \sim MN_{m,r}\left(\overline{\mathbf{B}}, \left(\mathbf{A}^{(*)\prime}\left(\boldsymbol{\Psi}^{(j)}\right)^{-1}\mathbf{A}^{(*)}\right)^{-1}, \overline{\boldsymbol{\Omega}}_B\right)$ in (65) and calculate $\boldsymbol{\alpha}^{(j)} = \mathbf{A}^{(*)}\left(\mathbf{B}^{(j)\prime}\mathbf{B}^{(j)}\right)^{1/2}$ and $\boldsymbol{\beta}^{(j)} = \mathbf{B}^{(j)}\left(\mathbf{B}^{(j)\prime}\mathbf{B}^{(j)}\right)^{-1/2}$

5. If $j > B$ generate $\mathbf{u}_{T+1}^{(j)}, \ldots, \mathbf{u}_{T+H}^{(j)}$ from $\mathbf{u}_t \sim N\left(0, \boldsymbol{\Psi}^{(j)}\right)$ and calculate recursively

$$\widetilde{\mathbf{y}}_{T+h}^{(j)\prime} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}_{T+h-i}^{(j)\prime}\mathbf{A}_i^{(j)} + \sum_{i=h}^{p} \mathbf{y}_{T+h-i}'\mathbf{A}_i^{(j)} + \mathbf{x}_{T+h}'\mathbf{C}^{(j)} + \mathbf{u}_{T+h}^{(j)\prime}$$

using $\mathbf{A}_1 = \mathbf{I} + \mathbf{B}_1' + \boldsymbol{\beta}\boldsymbol{\alpha}'$, $\mathbf{A}_i = \mathbf{B}_i' - \mathbf{B}_{i-1}'$, $i = 2, \ldots, p-1$ and $\mathbf{A}_p = -\mathbf{B}_{p-1}'$. Discarding the parameters yields $\left\{\widetilde{\mathbf{y}}_{T+1}^{(j)}, \ldots \widetilde{\mathbf{y}}_{T+H}^{(j)}\right\}_{j=B+1}^{B+R}$ as a sample from the joint predictive distribution.

Villani goes ahead and forecasts with both theory based (PPP, stationary domestic and foreign interest rates) and unrestricted cointegrating vectors with stationary Minnesota type prior beliefs on the short run dynamics. Of the considered models, Bayesian VECMs and ML-estimated VECMs and an ARIMA model, the Bayesian VECMs do best and are very close to each other.

6 Conditional forecasts

It is often of interest to condition the forecasts on different scenarios, for example different trajectories for the world economy, different developments of the oil price or different paths for the interest rate considered by a central bank. Another use of conditional forecasts is to incorporate information from higher frequency data or judgement into the model. An early example of conditional forecasts is Doan et al. (1984) who note that conditioning on a specific path for a variable is (given the parameters of the model) equivalent to imposing a set of linear constraints on the future disturbances, $\mathbf{u}_{T+1,}$ $\mathbf{u}_{T+2}, \ldots$ Conditional forecasts

can then be constructed by using the conditional means, $\widehat{\mathbf{u}}_{T+i}$, in the forecasting recursions

$$\widehat{\mathbf{y}}'_{T+h} = \sum_{i=1}^{h-1} \widehat{\mathbf{y}}'_{T+h-i}\mathbf{A}_i + \sum_{i=h}^{p} \mathbf{y}'_{T+h-i}\mathbf{A}_i + \mathbf{x}'_{T+h}\mathbf{C} + \widehat{\mathbf{u}}'_{T+h}. \tag{66}$$

This approach, while straightforward, has two potential drawbacks. It conditions on specific parameter values (e.g. the posterior means) and does not produce minimum mean square error forecasts conditional on the restrictions. This can be overcome by simulation from the posterior distribution of the parameters and solving for the restricted distribution of the disturbances for each set of parameter values (the whole predictive distribution can be simulated by also drawing $\mathbf{u}_{T+i}$ from the restricted distribution). The second issue is that the posterior distribution of the parameters will, in general, not be consistent with the future path we are conditioning on.

Waggoner and Zha (1999) addresses both these issues. Let $\mathbf{y}_{T+1:T+H} = \left(\mathbf{y}'_{T+1}, \ldots, \mathbf{y}'_{T+H}\right)'$ denote the future values to be forecasted, we can then write the condition that some of the variables follow a specific path or takes a specific value at a give time point as

$$\mathbf{R}\mathbf{y}_{T+1:T+H} = \mathbf{r}.$$

To see how this implies a restriction on the future disturbances we use recursive substitution to rewrite the future $\mathbf{y}_{T+i}$ in terms of past $\mathbf{y}_t$ and future $\mathbf{u}_{T+j}$, $j = 1, \ldots, i$.

$$\mathbf{y}_{T+i} = E\left(\mathbf{y}_{T+i}|\mathbf{Y}_T, \mathbf{\Gamma}, \mathbf{\Psi}\right) + \sum_{j=0}^{i-1} \mathbf{B}'_j \mathbf{u}_{T+i-j} \tag{67}$$

where $\mathbf{B}_i$ are the parameter matrices in the MA-representation,

$$\begin{aligned} \mathbf{B}_0 &= \mathbf{I} \\ \mathbf{B}_i &= \sum_{m=1}^{q} \mathbf{A}_m \mathbf{B}_{i-m}, \ i > 0 \end{aligned}$$

and $\overline{\mathbf{y}}_{T+i} = E\left(\mathbf{y}_{T+i}|\mathbf{Y}_T, \mathbf{\Gamma}, \mathbf{\Psi}\right)$ can be obtained trough the recursion (66) with $\widehat{\mathbf{u}}_{T+i} = 0$. Stacking the equations (67) we obtain $\mathbf{y}_{T+1:T+H} = \overline{\mathbf{y}}_{T+1:T+H} + \mathbf{B}'\mathbf{u}_{T+1:T+H}$ for

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_0 & \mathbf{B}_1 & \cdots & \mathbf{B}_{H-1} \\ \mathbf{0} & \mathbf{B}_0 & \cdots & \mathbf{B}_{H-2} \\ \vdots & & \ddots & \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B}_0 \end{pmatrix}.$$

The restriction can then be written as

$$\begin{aligned} \mathbf{R}\mathbf{y}_{T+1:T+H} &= \mathbf{R}\left(\overline{\mathbf{y}}_{T+1:T+H} + \mathbf{B}'\mathbf{u}_{T+1:T+H}\right) = \mathbf{r} \\ \mathbf{D}\mathbf{u}_{T+1:T+H} &= \mathbf{R}\mathbf{B}'\mathbf{u}_{T+1:T+H} = \mathbf{r} - \mathbf{R}\overline{\mathbf{y}}_{T+1:T+H} = \mathbf{d}. \end{aligned}$$

Since $\mathbf{u}_{T+1:T+H} \sim N\left(\mathbf{0}, \mathbf{V}_H\right)$ with $\mathbf{V}_H = \mathbf{I}_H \otimes \mathbf{\Psi}$, normal theory implies that the conditional distribution of $\mathbf{u}_{T+1:T+H}$ is

$$\mathbf{u}_{T+1:T+H}|\mathbf{D}\mathbf{u}_{T+1:T+H} = \mathbf{d} \sim N\left(\mathbf{V}_H\mathbf{D}'\left(\mathbf{D}\mathbf{V}_H\mathbf{D}'\right)^{-1}\mathbf{d}, \mathbf{V}_H - \mathbf{V}_H\mathbf{D}'\left(\mathbf{D}\mathbf{V}_H\mathbf{D}'\right)^{-1}\mathbf{D}\mathbf{V}_H\right) \tag{68}$$

which can be used for the simulation of the predictive distribution discussed above. Note, however, that the variance matrix is singular and some care is needed when generating $\mathbf{u}_{T+1:T+H}$, see Jarociński (2010) for an efficient method to generate $\mathbf{u}_{T+1:T+H}$.[22]

This does not address the issue of the consistency of the posterior distribution of the parameters and the restriction $\mathbf{R}\mathbf{y}_{T+1:T+H} = \mathbf{r}$. The restriction is information that in principle should be incorporated in the prior. This is, however, not possible in practice due to the highly non-linear relationship between the parameters and $\mathbf{y}_{T+1:T+H}$. Instead Waggoner and Zha (1999) suggests treating $\mathbf{y}_{T+1:T+H}$ as latent variables and simulate the joint posterior distribution of the parameters and $\mathbf{y}_{T+1:T+H}$ subject to the restriction and gives a straightforward MCMC sampler for this. The sampler is reproduced as Algorithm 11.

In addition to the hard restrictions $\mathbf{R}\mathbf{y}_{T+1:T+H} = \mathbf{r}$ Waggoner and Zha (1999) also considers "soft" restrictions on the form $\mathbf{R}\mathbf{y}_{T+1:T+H} \in \mathbb{S}$ where $\mathbb{S}$ is some subset of $\mathbb{R}^{mH}$ indicating an interval or region the forecasts that the forecasts are restricted to. Andersson, Palmqvist and Waggoner (2010) generalizes the approach of Waggoner and Zha (1999) to restrictions on the distribution of the future values, e.g. $\mathbf{R}\mathbf{y}_{T+1:T+H} \sim N(\mathbf{r}, \mathbf{V}_r)$. Robertson, Tallman and Whiteman (2005) takes a different approach and use exponential tilting to modify the unrestricted predictive distribution to match moment conditions of the form $E[g(\mathbf{y}_{T+1:T+H})] = \overline{\mathbf{g}}$. An example of the use of the exponential tilting method is Cogley, Morozov and Sargent (2005) who used it to adapt the predictive distribution to the Bank of England target inflation rate and other information that is external to the estimated VAR.

Forecasting performance Bloor and Matheson (2011) forecasts New Zealand real GDP, tradable CPI, non-tradable CPI, 90 interest rate and the trade weighted exchange rates using a real time data set. The models considered includes univariate AR-models, a small 5 variable VAR and BVAR, a medium sized 13 variable structural BVAR and a large 35 variable structural BVAR. The prior specification for the VARs is based on the approach of Banbura, Giannone and Reichlin (2010) (see section 9.2). Overall the VAR models do better than the univariate forecasting models with the large VAR improving on the smaller models. Incorporating external information in the form of Reserve Bank of New Zealand forecasts for variables where current data has not been released or future trajectories of variables is found to improve the forecast performance of the models.

[22] Note that the formulation of Jarociński (like the one of Waggoner and Zha) is in terms of a structural VAR and generates structural form innovations rather than the reduced form used here. To see how the method of Jarociński maps to the results here let $\mathbf{\Lambda}^{-1}$ be a factor of $\mathbf{\Psi}$, i.e. $\mathbf{\Lambda}^{-1}\mathbf{\Lambda}^{-\mathsf{T}} = \mathbf{\Psi}$ where $\mathbf{\Lambda}$ might come from a structural VAR or is the Cholesky factor of $\mathbf{\Psi}^{-1}$ (which is generally available as part of generating $\mathbf{\Psi}$ from the full conditional posterior in a reduced form VAR). The reduced form disturbances is related to the structural form innovations by $\mathbf{u}_{T+1:T+H} = \left(\mathbf{I}_H \otimes \mathbf{\Lambda}^{-1}\right)\mathbf{e}_{T+1:T+H}$ and the restriction on the structural innovations is $\widetilde{\mathbf{R}}\mathbf{e}_{T+1:T+H} = \mathbf{d}$ for $\widetilde{\mathbf{R}} = \mathbf{D}\left(\mathbf{I}_H \otimes \mathbf{\Lambda}^{-1}\right)$. Since the uconditional distribution of $\mathbf{e}_{T+1:T+H}$ is $N(\mathbf{0}, \mathbf{I}_{mH})$ we get $\mathbf{e}_{T+1:T+H}|\widetilde{\mathbf{R}}\mathbf{e}_{T+1:T+H} = \mathbf{d} \sim N\left(\widetilde{\mathbf{R}}'\left(\widetilde{\mathbf{R}}'\widetilde{\mathbf{R}}\right)\mathbf{d}, \mathbf{I}_{mH} - \widetilde{\mathbf{R}}'\left(\widetilde{\mathbf{R}}'\widetilde{\mathbf{R}}\right)^{-1}\widetilde{\mathbf{R}}\right)$ and the method of Jarociński can be used to generate first $\mathbf{e}_{T+1:T+H}$ and then $\mathbf{u}_{T+1:T+H}$.

Algorithm 11 MCMC sampler for VAR subject to "hard" restrictions

For a VAR subject to the restrictions $\mathbf{R}\mathbf{y}_{T+1:T+H} = \mathbf{r}$ select starting values $\mathbf{\Gamma}^{(0)}$ and $\mathbf{\Psi}^{(0)}$. The starting values can be taken from a separate simulation run on the historical data. For $i = 1, \ldots, B + R$

1. Generate $\mathbf{u}^{(j)}_{T+1:T+H}$ from the conditional distribution $\mathbf{u}_{T+1:T+H}|\mathbf{D}\mathbf{u}_{T+1:T+H} = \mathbf{d}, \mathbf{\Gamma}^{(j-1)}, \mathbf{\Psi}^{(j-1)}$ in (68) and calculate recursively

$$\widetilde{\mathbf{y}}^{(j)\prime}_{T+h} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}^{(j)\prime}_{T+h-i}\mathbf{A}^{(j-1)}_i + \sum_{i=h}^{p} \mathbf{y}'_{T+h-i}\mathbf{A}^{(j-1)}_i + \mathbf{x}'_{T+h}\mathbf{C}^{(j-1)} + \mathbf{u}^{(j)\prime}_{T+h}$$

2. Augment the data with $\widetilde{\mathbf{y}}^{(j)}_{T+1:T+H}$ and generate the parameters $\mathbf{\Psi}^{(j)}$ and $\mathbf{\Gamma}^{(j)}$ from the full conditional posteriors $\mathbf{\Psi}|\mathbf{Y}_T, \widetilde{\mathbf{y}}^{(j)}_{T+1:T+H}, \mathbf{\Gamma}^{(j-1)}$ and $\mathbf{\Gamma}|\mathbf{Y}_T, \widetilde{\mathbf{y}}^{(j)}_{T+1:T+H}, \mathbf{\Psi}^{(j)}$ using the relevant steps from one of the samplers discussed in this chapter depending on the choice of model structure and prior.

Discarding the parameters and keeping yields $\left\{\widetilde{\mathbf{y}}^{(j)}_{T+1}, \ldots \widetilde{\mathbf{y}}^{(j)}_{T+H}\right\}_{j=B+1}^{B+R}$ as a sample from the joint predictive distribution subject to the restrictions.

7 Time-varying parameters and stochastic volatility

7.1 Time-varying parameters

The constant parameter assumption implicit in the formulation of the VAR model (6) is often, but not always, reasonable. Parameter constancy might fail if the data covers a long time period, if there are changes in economic policy (e.g. monetary policy) as well as for many other reasons. It can thus be useful to allow for the parameters to change over time and we write the VAR as

$$\mathbf{y}_t = \mathbf{W}_t\boldsymbol{\gamma}_t + \mathbf{u}_t \tag{69}$$

for $\mathbf{W}_t = \mathbf{I}_m \otimes \mathbf{z}'_t$.

Doan et al. (1984), Highfield (1987) and Sims (1993) were among the earliest to introduce parameter variation in VAR models. Sims (1993), Doan et al. (1984) retains the equation by equation estimation strategy of Litterman while allowing the regression parameters to follow an AR(1) process

$$\boldsymbol{\gamma}_{t+1} = \pi_8\boldsymbol{\gamma}_t + (1 - \pi_8)\underline{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}_t \tag{70}$$

with $0 \leq \pi_8 \leq 1$. Doan et al. (1984) shows how the estimation can be conducted using the Kalman filter to update the state vector $\boldsymbol{\gamma}_t$ and conducts a search over a subset of the hyperparameters to find the combinations that provides the best forecast accuracy in a 10 variable VAR. Highfield (1987) relaxes the assumption of a known diagonal error variance-covariance matrix and uses the normal-Whishart conjugate prior in a state space formulation of the model. These are all examples of the type of time-varying parameter VAR models (TVP-VAR) formulated as state space models that we will focus on. There

are, of course, other ways to formulate a model where the parameters are allowed to change over time. This includes models accommodating structural breaks by including dummy variables that interact with some or all of the right hand side variables and Markov switching models with a fixed number of regimes (Chib (1998)) or an evolving number of regimes (Pesaran, Petenuzzo and Timmermann (2006), Koop and Potter (2007)).

The popularity of the Bayesian approach to TVP-VARs owes much to Cogley and Sargent (2002, 2005) and Primiceri (2005) who, although not primarily concerned with forecasting, provides the foundations for Bayesian inference in these models. Koop and Korobilis (2009) provides a good introduction to TVP-VARs.

The basic TVP-VAR complements the observation equation (69) with the state equation[23]

$$\boldsymbol{\gamma}_{t+1} = \boldsymbol{\gamma}_t + \boldsymbol{\varepsilon}_t. \tag{71}$$

That is, the parameters are assumed to follow a random walk and evolve smoothly over time. $\boldsymbol{\varepsilon}_t$ is assumed to be normally distributed, $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \mathbf{Q})$ and independent of the error term in the observation equation which is also normal, $\mathbf{u}_t \sim N(\mathbf{0}, \boldsymbol{\Psi})$. Note that the state equation implies that $\boldsymbol{\gamma}_{t+1}|\boldsymbol{\gamma}_t, \mathbf{Q} \sim N(\boldsymbol{\gamma}_t, \mathbf{Q})$ and that this in a sense serves as prior distribution for $\boldsymbol{\gamma}_{t+1}$ and the prior for all the states (parameters) is simply a product of normal distributions that needs to be complemented with a prior for the first state, $\pi(\boldsymbol{\gamma}_1)$, which is then usefully also taken to be normal,

$$\boldsymbol{\gamma}_1 \sim N\left(\mathbf{s}_{1|0}, \mathbf{P}_{1|0}\right). \tag{72}$$

The prior specification is completed with independent inverse Wishart priors for $\boldsymbol{\Psi}$ and $\mathbf{Q}$,

$$\begin{aligned} \boldsymbol{\Psi} &\sim iW\left(\underline{\boldsymbol{\Psi}}, \underline{v}\right) \\ \mathbf{Q} &\sim iW\left(\underline{\mathbf{Q}}, \underline{v}_Q\right). \end{aligned} \tag{73}$$

The time-varying parameter specification introduces an additional layer of complication when forecasting since the parameters can not be assumed to be constant in the forecast period. This contributes to additional variability in the predictive distribution and we must simulate $\boldsymbol{\gamma}_{T+h}$ from the state equation 71 in order to simulate the predictive distribution. See Cogley et al. (2005) for a discussion of these issues.

It is straightforward to set up a Gibbs sampler for the joint posterior distribution of $\left(\boldsymbol{\gamma}^T, \boldsymbol{\Psi}, \mathbf{Q}\right)$ (as a notational convention we will use superscripts to refer to sequences of variables and parameters, i.e. x^t to refer to the sequence $x_1, \ldots, x_t$). Conditional on the unobserved time varying parameters (states), $\boldsymbol{\gamma}_t$, posterior inference for $\boldsymbol{\Psi}$ and $\mathbf{Q}$ is standard and we have the full conditional posteriors as

$$\boldsymbol{\Psi}|\mathbf{y}^T, \boldsymbol{\gamma}^T \sim iW\left(\overline{\boldsymbol{\Psi}}, \overline{v}\right), \ \overline{\boldsymbol{\Psi}} = \underline{\boldsymbol{\Psi}} + \sum_{i=1}^{T} \left(\mathbf{y}_t - \mathbf{W}_t\boldsymbol{\gamma}_t\right)\left(\mathbf{y}_t - \mathbf{W}_t\boldsymbol{\gamma}_t\right)', \ \overline{v} = \underline{v} + T \tag{74}$$

and

$$\mathbf{Q}|\boldsymbol{\gamma}^T \sim iW\left(\overline{\mathbf{Q}}, \overline{v}\right), \ \overline{\mathbf{Q}} = \underline{\mathbf{Q}} + \sum_{i=1}^{T} \left(\boldsymbol{\gamma}_{t+1} - \boldsymbol{\gamma}_t\right)\left(\boldsymbol{\gamma}_{t+1} - \boldsymbol{\gamma}_t\right)', \ \overline{v}_Q = \underline{v}_Q + T. \tag{75}$$

[23] See Appendix B for a (very) brief introduction to state space models.

Algorithm 12 Gibbs sampler for the TVP-VAR

For the TVP-VAR (69), (71) and the priors (72), 73 select starting values $\boldsymbol{\Psi}^{(0)}$ and $\mathbf{Q}^{(0)}$. For $j = 1, \ldots, B + R$

1. Draw $\boldsymbol{\gamma}_T^{(j)}$ from the full conditional posterior $\boldsymbol{\gamma}_T|\mathbf{y}^T, \boldsymbol{\Psi}^{(j-1)}, \mathbf{Q}^{(j-1)} \sim N\left(\mathbf{s}_{T|T}, \mathbf{P}_{T|T}\right)$ obtained from the Kalman filter (equation (126) in Appendix B). For $t = T-1, \ldots, 1$ draw $\boldsymbol{\gamma}_t^{(j)}$ from the full conditional $\boldsymbol{\gamma}_t|\mathbf{y}^T, \boldsymbol{\Psi}^{(j-1)}, \mathbf{Q}^{(j-1)}, \boldsymbol{\gamma}_{t+1}^{(j)}$ in (76) by running the simulation smoother (Algorithm B.3 in Appendix B).

2. Draw $\mathbf{Q}^{(j)}$ from the full conditional $\mathbf{Q}|\boldsymbol{\gamma}^{T(j)}$ in (75).

3. Draw $\boldsymbol{\Psi}^{(j)}$ from the full conditional $\boldsymbol{\Psi}|\mathbf{y}^T, \boldsymbol{\gamma}^{T(j)}$ in (74).

4. If $j > B$
   Generate $\mathbf{u}_{T+1}^{(j)}, \ldots, \mathbf{u}_{T+H}^{(j)}$ from $\mathbf{u}_t \sim N\left(0, \boldsymbol{\Psi}^{(j)}\right)$, for $h = 1, \ldots, H$, generate $\boldsymbol{\gamma}_{T+h}^{(j)}$ from the state equation (71) and calculate recursively

$$\widetilde{\mathbf{y}}_{T+h}^{(j)\prime} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}_{T+h-i}^{(j)\prime} \mathbf{A}_{i,T+h}^{(j)} + \sum_{i=h}^{p} \mathbf{y}_{T+h-i}^{\prime} \mathbf{A}_{i,T+h}^{(j)} + \mathbf{x}_{T+h}^{\prime} \mathbf{C}_{T+h}^{(j)} + \mathbf{u}_{T+h}^{(j)\prime}. \tag{77}$$

   Discarding the parameters yields $\left\{\widetilde{\mathbf{y}}_{T+1}^{(j)}, \ldots \widetilde{\mathbf{y}}_{T+H}^{(j)}\right\}_{j=1}^{R}$ as a sample of independent draws from the joint predictive distribution.

Generating $\boldsymbol{\gamma}^T$ from the full conditional posterior is somewhat more involved. The basic idea is to make use of the linear state space structure. Given $\boldsymbol{\Psi}$ and $\mathbf{Q}$ a run of the (forward) Kalman filter (equation (126) in Appendix B[24]) produces the sequence of conditional distributions $\boldsymbol{\gamma}_t|\mathbf{y}^t, \boldsymbol{\Psi}, \mathbf{Q} \sim N\left(\mathbf{s}_{t|t}, \mathbf{P}_{t|t}\right)$, $t = 1, \ldots, T$. This gives the full conditional posterior for the last state, $\boldsymbol{\gamma}_T|\mathbf{y}^T, \boldsymbol{\Psi}, \mathbf{Q} \sim N\left(\mathbf{s}_{T|T}, \mathbf{P}_{T|T}\right)$ from which a draw of $\boldsymbol{\gamma}_T$ can be made. To obtain the full conditional posterior for all the states we decompose the joint distribution as $p\left(\boldsymbol{\gamma}^T|\mathbf{y}^T, \boldsymbol{\Psi}, \mathbf{Q}\right) = p\left(\boldsymbol{\gamma}_T|\mathbf{y}^T, \boldsymbol{\Psi}, \mathbf{Q}\right) \prod_{t=1}^{T-1} p\left(\boldsymbol{\gamma}_t|\mathbf{y}^T, \boldsymbol{\Psi}, \mathbf{Q}, \boldsymbol{\gamma}_{t+1}\right)$ where

$$\boldsymbol{\gamma}_t|\mathbf{y}^T, \boldsymbol{\Psi}, \mathbf{Q}, \boldsymbol{\gamma}_{t+1} \sim N\left(\mathbf{s}_{t|T}, \mathbf{P}_{t|T}\right) \tag{76}$$

and the moments are obtained from the backwards recursions in the simulation smoother (Algorithm B.3 in Appendix B). Algorithm 12 summarizes the Gibbs sampler for the TVP-VAR.

Specifying the prior The random walk nature of the state equation (71) puts little structure on the behavior of $\boldsymbol{\gamma}_t$ and the implied prior for the sequence of parameters, $\boldsymbol{\gamma}^T$, gets increasingly loose as the unconditional variance increases at the rate $t\mathbf{Q}$. To enforce some smoothness in $\boldsymbol{\gamma}_t$ it is useful to focus the prior for $\mathbf{Q}$ on small values of the variances. In addition, the random walk nature can lead to explosive behavior at some

[24]Set $\mathbf{Z}_t = \mathbf{W}_t$, $\mathbf{H}_t = \boldsymbol{\Psi}$, $\mathbf{d}_t = 0$, $\mathbf{T}_t = \mathbf{I}$ and $\mathbf{Q}_t = \mathbf{Q}$ in the Kalman filter equations.

time points which can be undesirable if the data is believed to be stationary. To prevent this Cogley and Sargent (2002, 2005) truncates the prior for $\boldsymbol{\gamma}^T$ to the region where $\boldsymbol{\gamma}_t$, $\forall t$, is stationary. Truncated prior distributions like this are easily incorporated in a Gibbs sampler, simply check the stationarity condition for all t at the end of step 1 of Algorithm 12 and redo step 1 if it fails for at least one time point.

It is common to use a training sample prior for the first state, $\boldsymbol{\gamma}_1$. The first, say $k+20$, observations are set aside as a training sample and the prior mean and variance are based on the OLS estimates using the training sample. The prior variance should, in general, be relatively large so as not to make the prior too informative, see, for example Primiceri (2005).

An alternative is to base the prior for $\boldsymbol{\gamma}_1$ on the Minnesota prior. For this to be effective a modified state equation along the line of the Doan et al. (1984) specification (70) is useful. Generalizing this we can write the state equation as

$$\boldsymbol{\gamma}_{t+1} = \mathbf{s}_{1|0} + \boldsymbol{\Phi}\left(\boldsymbol{\gamma}_t - \mathbf{s}_{1|0}\right) + \boldsymbol{\varepsilon}_t \tag{78}$$

with $\boldsymbol{\Phi}$ a diagonal matrix. The state equation is stationary and mean reverting if $|\phi_{ii}| < 1$. The diagonal elements can be taken as fixed and specified along with the other prior parameters or estimated. Inference on ϕ_{ii} is straightforward in the latter case. Conditional on $\boldsymbol{\gamma}^T$(78) is just a multivariate regression and it is easy to add a block to the Gibbs sampler drawing from the full conditional posterior for $\boldsymbol{\Phi}$.

Forecasting performance Sims (1993) reports on the enhancements made to the original Litterman forecasting model where he allows for conditional heteroskedasticity and non-normal errors in addition to the time varying regression parameters. The result of these modifications and the addition of three more variables, the trade weighted value of the dollar, the S&P 500 stock index and the commodity price index, led to an improved forecasting performance for the price variable, comparable or slightly better forecasts for the real variables and slightly worse forecasts for interest rates compared to the original Litterman model.

Canova (2007) forecasts the inflation rate of the G7 countries using a range of models. First there is a set of country specific models, univariate ARMA, several bivariate VARs with the additional variable suggested by theory, trivariate VARs where the two additional variables are selected to minimize the in sample mean square error for inflation. The trivariate VAR is estimated by OLS, as a BVAR with Minnesota style prior beliefs and also as a Bayesian TVP-VAR and a Bayesian TVP-AR with mean reverting state equations (70). Canova also use several "international" models, three variables controlling for international demand are added as predetermined variables to the country specific BVARs, a TVP-BVAR for the 7 inflation rates with the same international variables as predetermined, a Bayesian panel VAR and a dynamic factor model for the inflation rates where the factors are principal components of the variables in the panel VAR. All the models are formulated in terms of direct rather than iterated forecasts, i.e. $y_{t+h} = \phi y_t + u_{t+h}$ for an AR(1). The models differ in two important dimensions, the richness of the information set and the flexibility of the specification. Overall the model with the largest information set and the most general specification, the Bayesian panel VAR does best. Comparing models with similar information sets, a BVAR improves on a VAR

estimated with OLS and time-varying parameters improve the forecasts for univariate models but not for the BVARs.

Clark and McCracken (2010) use a real time data set and forecasts the US inflation, interest rate and output using a wide range of trivariate VAR models based on different approaches to allowing for structural change. This includes models in levels, in differences, estimated on rolling windows of the data, estimated by OLS or as BVARs with Minnesota type priors and TVP-BVARs with random walk state equations. While the focus is on different methods for combining forecasts and how well they cope with structural changes Clark and McCracken (2010) do report some results for individual models. Of these a BVAR with detrended inflation does best and, while not directly comparable, considerably better than a TVP-BVAR where inflation has not been detrended.

7.2 Stochastic volatility

The constant error variance assumption can also be questioned, especially in light of the so called "great moderation" with considerable lower variability in key macroeconomic variables since the mid 1980s. It can also be difficult to empirically distinguish between a model with constant parameters and time-varying variances and a model with time-varying parameters and constant error variance. It can thus be prudent to allow for both. Cogley and Sargent (2005) and Primiceri (2005) construct TVPSV-VAR models by adding stochastic volatility to the TVP-VAR. The setup in Primiceri (2005) is more general and the overview here is based on Primiceri. See Koop and Korobilis (2009) for a more in-depth discussion of stochastic volatility in VAR models.

To introduce time-varying volatilities and correlations we decompose the error variance matrix into

$$\mathbf{\Psi}_t = \mathbf{L}_t^{-1}\mathbf{D}_t\mathbf{L}_t^{-\mathsf{T}}$$

where $\mathbf{D}_t$ is a diagonal matrix, $\mathbf{D}_t = \text{diag}\left(d_{1t}, \ldots, d_{mt}\right)$ with a stochastic volatility specification,

$$\begin{aligned} d_{it} &= \exp\left(h_{it}/2\right), \\ h_{i,t+1} &= \mu_i + \phi_i\left(h_{it} - \mu_i\right) + \eta_{it} \end{aligned} \tag{79}$$

where $\boldsymbol{\eta}_t = \left(\eta_{1t}, \ldots, \eta_{mt}\right)$ is iid $N\left(\mathbf{0}, \mathbf{V}_\eta\right)$. The purpose of $\mathbf{L}_t$ is to allow for an arbitrary time-varying correlation structure. It is a lower triangular matrix with 1 on the diagonal, e.g. for $m = 3$

$$\mathbf{L}_t = \begin{pmatrix} 1 & 0 & 0 \\ l_{21,t} & 1 & 0 \\ l_{31,t} & l_{32,t} & 1 \end{pmatrix}, \tag{80}$$

where the time-varying elements under the diagonal follow a random walk,

$$\mathbf{l}_{t+1} = \mathbf{l}_t + \boldsymbol{\zeta}_t \tag{81}$$

for $\mathbf{l}_t$ a $m\left(m-1\right)/2$ vector that collects $l_{ij,t}$, $i > j$ in row major order, $\boldsymbol{\zeta}_t$ iid $N\left(\mathbf{0}, \mathbf{V}_\zeta\right)$ and $\mathbf{V}_\zeta$ block diagonal with blocks corresponding to the rows of $\mathbf{L}_t$. The triangular specification (80) is convenient and can also be interpreted as a structural VAR with time-varying parameters.

Prior specification Prior distributions for the parameters μ_i, ϕ_i, $\mathbf{V}_\eta$ and $\mathbf{V}_\zeta$ are needed in order to complete the model. μ_i is the unconditional expectation of the log volatilities and in absence of specific information about the scale of the parameters a noninformative normal prior can be used, $\boldsymbol{\mu} \sim N\left(0, \underline{\boldsymbol{\Sigma}}_\mu\right)$ with $\underline{\boldsymbol{\Sigma}}_\mu$ diagonal. For the autoregressive parameter it is common to restrict this to the stationary region and specify a truncated normal, $\boldsymbol{\phi} \sim N\left(\underline{\boldsymbol{\phi}}, \underline{\boldsymbol{\Sigma}}_\phi\right) I\left(|\phi_i| < 1\right)$ with $\underline{\boldsymbol{\Sigma}}_\phi$ diagonal. Alternatively one can, as in Primiceri (2005), work with a random walk specification for h_{it} with $\phi_i = 1$ and where μ_i drops out of the model. For the state equation variance $\mathbf{V}_\eta$ an inverse Wishart prior, $\mathbf{V}_\eta \sim iW\left(\underline{\mathbf{S}}_\eta, \underline{\upsilon}_\eta\right)$, is conditionally conjugate and convenient. For the log volatilities an initial condition (prior) is needed. With ϕ_i restricted to the stationary region, $h_{i1}|\mu_i, \phi_i, \sigma^2_{\eta_i} \sim N\left(\mu_i, \sigma^2_{\eta_i}/\left(1-\phi_i^2\right)\right)$ for $\sigma^2_{\eta_i}$ the i^{th} diagonal element of $\mathbf{V}_\eta$, is a natural choice, and with $\phi_i = 1$ a noninformative normal distribution can be used for the initial condition.

$\mathbf{V}_\zeta$ is assumed to be block diagonal in order to simplify the posterior sampler (Primiceri (2005) shows how to relax this). With a block diagonal structure the prior for the blocks $\mathbf{V}_{\zeta,i}$, $i = 2, \ldots, m$, can be specified with independent inverse Wishart distributions, $\mathbf{V}_{\zeta,i} \sim iW\left(\underline{\mathbf{S}}_{\zeta,i}, \underline{\upsilon}_{\zeta,i}\right)$. In addition, an initial condition is needed for the elements of $\mathbf{L}_1$ collected in $\mathbf{l}_1$. For simplicity, this can be taken as a noninformative normal distribution. For some additional simplification, $\mathbf{V}_\eta$ and $\mathbf{V}_\zeta$ can be specified as diagonal matrices with inverse Gamma priors for the diagonal elements.

The exact choices for the parameters of the prior distribution can be based on a training sample as for the TVP-VAR model, see Primiceri (2005) for an example.

Sampling from the posterior When discussing inference on the variance parameters $\mathbf{L}_t$, $\mathbf{V}_\zeta$, $\mathbf{D}_t$, μ_i, ϕ_i and $\mathbf{V}_\eta$ we condition on the other parameters in the model and simply take $\widetilde{\mathbf{y}}_t = \mathbf{y}_t - \mathbf{W}_t\boldsymbol{\gamma}_t$ as our data. This implies that the (conditional) inference procedure for the variance parameters does not depend on if the other parameters are time-varying or constant. It will consist of a few blocks of a Gibbs sampler that can be combined with a MCMC sampler for a VAR with constant or time-varying parameters. The inference procedure for the remaining parameters is, on the other, affected by the introduction of time-varying variances. For the TVP-VAR this amounts to noting that $\boldsymbol{\Psi}_t$ ($\mathbf{H}_t$ in the Kalman filter equations (126) in Appendix B) is now time-varying. The constant parameter case can also be handled with the help of the Kalman filter by setting $\mathbf{Q}_t = \mathbf{0}$, $\mathbf{d}_t = \mathbf{0}$ and $\mathbf{T}_t = \mathbf{I}$ in addition to allowing for time-varying variances in the Kalman filter. By setting $\mathbf{Q}_t$ to zero the parameter variation is shut down and $\boldsymbol{\gamma}_t = \boldsymbol{\gamma}_1$, $\forall t$, is enforced. The prior (72) for $\boldsymbol{\gamma}_1$ is then a prior for the constant parameter. After running the Kalman filter the (conditional) posterior mean and variance of $\boldsymbol{\gamma}$ is returned as $\mathbf{s}_{T|T}$ and $\mathbf{P}_{T|T}$ and no smoothing is necessary. The conditional posterior for a constant $\boldsymbol{\gamma}$ can of course also be derived analytically. Write the constant parameter VAR as

$$\mathbf{y}_t = \mathbf{W}_t\boldsymbol{\gamma} + \mathbf{u}_t$$

with $\mathbf{u}_t \sim N(\mathbf{0}, \mathbf{\Psi}_t)$ and $\mathbf{W}_t = \mathbf{I}_m \otimes \mathbf{z}_t'$. With an independent normal prior $\boldsymbol{\gamma} \sim N\left(\underline{\boldsymbol{\gamma}}, \underline{\boldsymbol{\Sigma}}_\gamma\right)$ as in section 3.2.2 the conditional posterior is

$$\begin{aligned} \boldsymbol{\gamma}|\mathbf{Y}_T, \mathbf{\Psi}^T &\sim N\left(\overline{\boldsymbol{\gamma}}, \overline{\boldsymbol{\Sigma}}_\gamma\right) \\ \overline{\boldsymbol{\Sigma}}_\gamma &= \left(\underline{\boldsymbol{\Sigma}}_\gamma^{-1} + \sum_{i=1}^{T} \mathbf{W}_t' \mathbf{\Psi}_T^{-1} \mathbf{W}_t\right)^{-1}, \\ \overline{\boldsymbol{\gamma}} &= \overline{\boldsymbol{\Sigma}}_\gamma \left[\underline{\boldsymbol{\Sigma}}_\gamma^{-1} \underline{\boldsymbol{\gamma}} + \left(\sum_{i=1}^{T} \mathbf{W}_t' \mathbf{\Psi}_T^{-1} \mathbf{W}_t\right) \widehat{\boldsymbol{\gamma}}\right] \end{aligned} \tag{82}$$

where $\widehat{\boldsymbol{\gamma}}$ is the GLS estimate $\widehat{\boldsymbol{\gamma}} = \left(\sum_{i=1}^{T} \mathbf{W}_t' \mathbf{\Psi}_T^{-1} \mathbf{W}_t\right)^{-1} \sum_{i=1}^{T} \mathbf{W}_t' \mathbf{\Psi}_T^{-1} \mathbf{y}_t$.

Turning to the conditional posteriors for the variance parameters, the one for the correlation parameters in $\mathbf{L}_t$ is relatively straightforward and replicates the treatment of time-varying parameters in the TVP-VAR. Multiplying each observation with $\mathbf{L}_t$ we obtain $\mathbf{L}_t \widetilde{\mathbf{y}}_t = \mathbf{L}_t \mathbf{u}_t = \mathbf{e}_t$ with $V(\mathbf{e}_t) = \mathbf{D}_t$. This yields $m-1$ uncorrelated equations in a triangular equation system,

$$\widetilde{y}_{it} = -\sum_{j=1}^{i-1} \widetilde{y}_{jt} l_{ij,t} + e_{it}, \; i = 2, \ldots, m. \tag{83}$$

This, together with the assumption that $\mathbf{V}_\zeta$ is block diagonal, means that the full conditional posterior for $\mathbf{L}_t, t = 1, \ldots, T$ can be recovered by running the corresponding Kalman filter and simulation smoother for each equation in turn, setting $\mathbf{Z}_t = (\widetilde{y}_{1t}, \ldots, \widetilde{y}_{i-1,t})$, $\mathbf{H}_t = \exp(h_{it})$, $\mathbf{d}_t = 0$, $\mathbf{T}_t = \mathbf{I}$ and $\mathbf{Q}_t$ to the relevant block of $\mathbf{V}_\zeta$ in the Kalman filter equations.

The posterior for $\mathbf{V}_\zeta$ is straightforward conditional on $\mathbf{L}_t$. With the block diagonal structure the blocks are inverse Wishart

$$\begin{aligned} \mathbf{V}_{\zeta,i}|\mathbf{l}^T &\sim iW\left(\overline{\mathbf{S}}_{\zeta,i}, \overline{v}_{\zeta,i}\right), \; i = 2, \ldots, m \\ \overline{\mathbf{S}}_{\zeta,i} &= \underline{\mathbf{S}}_{\zeta,i} + \sum_{t=1}^{T} (\mathbf{l}_{i,t} - \mathbf{l}_{i,t-1})(\mathbf{l}_{i,t} - \mathbf{l}_{i,t-1})', \; \overline{v}_{\zeta,i} = \underline{v}_{\zeta,i} + T \end{aligned} \tag{84}$$

for $\mathbf{l}_{i,t} = (l_{i1,t}, \ldots, l_{i,i-1,t})'$.

The posterior analysis of the time-varying volatilities is complicated by the fact that the observation equation is non-linear in the states $h_{i,t}$.Let $\mathbf{y}_t^* = \mathbf{L}_t \widetilde{\mathbf{y}}_t$, we then have

$$y_{it}^* = \exp(h_{it}/2) v_{it}$$

where v_{it} is iid $N(0,1)$. Squaring and then taking logarithms yields

$$y_{it}^{**} = h_{it} + v_{it}^*$$

where $y_{it}^{**} = \ln\left[(y_{it}^*)^2 + c\right]$ for c a small positive constant and with $v_{it}^* = \ln v_{it}^2$ distributed as the logarithm of a χ_1^2 random variable. Kim, Shephard and Chib (1998) show that the distribution of v_{it}^* is well approximated by a mixture of normals, $p(v_{it}^*) \approx \sum_{j=1}^{7} q_j N\left(v_{it}^*; m_j - 1.2704, \tau_j^2\right)$. The mixture coefficients obtained by Kim et al. (1998)

are reproduced in Table 1. By introducing a latent indicator variable δ_{it} for which component in the mixture v_{it}^* has been drawn from, the mixture can be rewritten as $v_{it}^*|\delta_{it} = j \sim N\left(m_j - 1.2704, \tau_j^2\right)$ with $P\left(\delta_{it} = j\right) = q_j$ the problem can thus be transformed into a linear and normal filtering problem conditional on δ_{it}. Kim et al. (1998) develops an MCMC algorithm for sampling from the posterior distribution of the states h_{it}.

Conditional on the sequence of states, $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_T$, we can, using the notation of Appendix B, write the stochastic volatility part of the model as a linear state space system,

$$\begin{aligned} \mathbf{y}_t^{**} &= \mathbf{Z}_t \begin{pmatrix} 1 \\ \mathbf{h}_t \end{pmatrix} + \mathbf{e}_t^* \\ \mathbf{h}_{t+1} &= \mathbf{d} + \mathbf{T}\mathbf{h}_t + \boldsymbol{\eta}, \end{aligned} \tag{85}$$

for $\mathbf{Z}_t = \left(\mathbf{m}_t, \mathbf{I}_m\right)$, $d_i = \mu_i\left(1 - \phi_i\right)$, $\mathbf{T} = \operatorname{diag}\left(\phi_i\right)$. The elements of $\mathbf{m}_t$ are the conditional means $E\left(v_{it}^*|\delta_{it} = j\right) = m_j - 1.2704$ and $\mathbf{e}_t^* \sim N\left(\mathbf{0}, \mathbf{H}_t\right)$ for $\mathbf{H}_t$ diagonal with diagonal elements given by the conditional variances, $V\left(v_{it}^*|\delta_{it} = j\right) = \tau_j^2$. Running the Kalman filter and then the simulation smoother on (85) yields a draw of h_{it}, $i = 1, \ldots, m$, $t = 1, \ldots, T$ from the full conditional posterior.

The full conditional posterior for $\mathbf{V}_\eta$ is straightforward as an inverse Wishart,

$$\mathbf{V}_\eta|\mathbf{h}^T, \boldsymbol{\mu}, \boldsymbol{\phi} \sim iW\left(\overline{\mathbf{S}}_\eta, \overline{v}_\eta\right), \ \overline{\mathbf{S}}_\eta = \underline{\mathbf{S}}_\eta + \sum_{t=1}^{T}\left(\mathbf{h}_t - \mathbf{d} - \mathbf{T}\mathbf{h}_{t-1}\right)\left(\mathbf{h}_t - \mathbf{d} - \mathbf{T}\mathbf{h}_{t-1}\right)', \ \overline{v}_\eta = \underline{v}_\eta + T. \tag{86}$$

The states, δ_{it}, can be sampled from the full conditional posterior

$$p\left(\delta_{it} = j|y_{it}^{**}, h_{it}\right) \propto q_j \frac{1}{\tau_j} \exp\left(\frac{\left(y_{i,t}^{**} - m_j - 1.2704 - h_{it}\right)^2}{2\tau_j^2}\right). \tag{87}$$

For ϕ_i and μ_i, finally, write

$$\mathbf{h}_t^* = \mathbf{h}_t - \mu = \mathbf{X}_t\boldsymbol{\phi} + \boldsymbol{\eta}$$

for $\mathbf{X}_t = \operatorname{diag}\left(h_{i,t-1} - \mu_i\right)$. Stacking the observations and performing the usual calculations yields the full conditional posterior for $\boldsymbol{\phi}$ as

$$\begin{aligned} \boldsymbol{\phi}|\mathbf{h}^T, \boldsymbol{\mu}, \mathbf{V}_\eta &\sim N\left(\overline{\boldsymbol{\phi}}, \overline{\boldsymbol{\Sigma}}_\phi\right) I\left(|\phi_i| < 1\right) \\ \overline{\boldsymbol{\Sigma}}_\phi &= \left(\underline{\boldsymbol{\Sigma}}_\phi^{-1} + \sum_{t=1}^{T} \mathbf{X}_t' \mathbf{V}_\eta^{-1} \mathbf{X}_t\right)^{-1} \\ \overline{\boldsymbol{\phi}} &= \overline{\boldsymbol{\Sigma}}_\phi\left(\underline{\boldsymbol{\Sigma}}_\phi^{-1}\underline{\boldsymbol{\phi}} + \sum_{t=1}^{T} \mathbf{X}_t' \mathbf{V}_\eta^{-1} \mathbf{h}_t^*\right). \end{aligned} \tag{88}$$

The full conditional posterior for $\boldsymbol{\mu}$ is obtained in a similar fashion as

$$\begin{aligned} \boldsymbol{\mu}|\mathbf{h}^T, \boldsymbol{\phi}, \mathbf{V}_\eta &\sim N\left(\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\Sigma}}_\mu\right) \\ \overline{\boldsymbol{\Sigma}}_\mu &= \left(\underline{\boldsymbol{\Sigma}}_\mu^{-1} + T\mathbf{X}'\mathbf{V}_\eta^{-1}\mathbf{X}\right)^{-1} \\ \overline{\boldsymbol{\mu}} &= \overline{\boldsymbol{\Sigma}}_\mu\left(\underline{\boldsymbol{\Sigma}}_\mu^{-1}\underline{\boldsymbol{\phi}} + \mathbf{X}'\mathbf{V}_\eta^{-1}\sum_{t=1}^{T} \mathbf{h}_t^{**}\right). \end{aligned} \tag{89}$$

Table 1 Normal mixture coefficient for $\ln \chi_1^2$

Component, δ	q_j	m_j	τ_j^2
1	0.00730	−10.12999	5.79596
2	0.10556	−3.97281	2.61369
3	0.00002	−8.56686	5.17950
4	0.04395	2.77786	0.16735
5	0.34001	0.61942	0.64009
6	0.24566	1.79518	0.34023
7	0.25750	−1.08819	1.26261

Source: Kim et al. (1998)

by writing

$$\mathbf{h}_t^{**} = \mathbf{h}_t - \operatorname{diag}(\boldsymbol{\phi})\,\mathbf{h}_{t-1} = \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\eta}$$

for $\mathbf{X} = \operatorname{diag}(1 - \phi_i)$.

The steps of the resulting Gibbs sampler are summarized in algorithm 13.

Forecast performance Clark (2011) studies point and density forecasts using real time data on the US output growth, unemployment rate, inflation and federal funds rate. With a focus on the effects of changing data variability motivated by the possible end of the great moderation Clark introduces a new model, the steady state BVAR of Villani (2009) combined with stochastic volatility following the approach of Cogley and Sargent (2005). The forecast performance of the new model is compared to univariate AR-models with and without stochastic volatility, a standard Minnesota type BVAR with a normal-diffuse prior and a Minnesota type steady state BVAR. The BVAR includes the four variables to forecast whereas the steady state BVARs includes the detrended unemployment rate, inflation and the interest rate less the long run inflation expectation and the long run inflations expectation as an additional variable. The three BVAR variants are estimated on both a recursively updated and rolling data window. When the point forecasts are evaluated by their RMSEs, the BVARs generally do worse than the benchmark univariate AR without stochastic volatilities at shorter lead times (1 and 2 quarters) but improve on the benchmark for lead times of 1 and 2 years with the exception for the forecasts of inflation. The steady state BVAR tend to do better than the standard BVAR and adding stochastic volatility brings further improvements. In addition, the rolling updating scheme with an 80 observation window tends to produce better forecasts than recursive updating. The density forecasts are evaluated using several criteria. In terms of the empirical coverage of the prediction intervals it is found that the models without stochastic volatility produces to wide intervals while the models with stochastic volatility produces better calibrated prediction intervals. When evaluated using the probability integral transform (PIT, see Corradi and Swanson (2006)) the hypothesis of a correctly specified predictive distribution is rejected in almost all cases for models without stochastic volatility whereas the stochastic volatility models pass the test with only a few exceptions. Finally, evaluating the density forecasts using the log predictive density score, the stochastic volatility models do considerably better at the shorter lead times while the differences are quite small for the longer lead times. Similarly the steady state BVAR outperforms the standard BVAR at short lead times and the steady state BVAR with stochastic volatility outperforms the

Algorithm 13 Gibbs sampler for VAR with stochastic volatility

For the VAR with stochastic volatility select starting values $\boldsymbol{\Psi}^{T(0)} = \boldsymbol{\Psi}_1^{(0)}, \ldots, \boldsymbol{\Psi}_T^{(0)}, \mathbf{V}_{\zeta,i}^{(0)}$, $i = 2, \ldots, m$, $\boldsymbol{\mu}^{(0)}, \boldsymbol{\phi}^{(0)}$, $\mathbf{V}_\eta^{(0)}$ and $\boldsymbol{\delta}^{T(0)}$.
For $j = 1, \ldots, B + R$

1. Draw the regression $\boldsymbol{\gamma}^{(j)}$ from the full conditional posterior $\boldsymbol{\gamma}|\mathbf{Y}_T, \boldsymbol{\Psi}^{T(j-1)}$ in (82) for the constant parameter case or using the Kalman filter and simulation smoother for the time-varying parameter case as in Algorithm 12.

2. For $i = 2, \ldots, m$ run the Kalman filter for the observation equation (83) and state equation 81) and generate $\mathbf{l}_{i,T}^{(j)}$ from the normal full conditional posterior $\mathbf{l}_{i,T}|\mathbf{Y}_T, \mathbf{V}_{\zeta,i}^{(j-1)}$ with parameters given by $\mathbf{s}_{T|T}$ and $\mathbf{P}_{T|T}$ from the Kalman filter. For $k = T-1, \ldots, 1$ draw $\mathbf{l}_{i,t}^{(j)}$ from the normal full conditional posterior $\mathbf{l}_{i,t}^{(j)}|\mathbf{y}^t, \mathbf{V}_{\zeta,i}^{(j-1)}, \mathbf{l}_{i,t+1}^{(j)}$ obtained from the simulation smoother.

3. For $i = 2, \ldots, m$, draw $\mathbf{V}_{\zeta,i}^{(j)}$ from the full conditional posterior $\mathbf{V}_{\zeta,i}|\mathbf{l}^{T(j)}$ in (84).

4. Draw the log volatilities $\mathbf{h}_T^{(j)}$ from the normal full conditional posterior $h_T|\mathbf{Y}_T, \boldsymbol{\mu}^{(j-1)}, \boldsymbol{\phi}^{(j-1)}, \mathbf{V}_\eta^{(j-1)}, \boldsymbol{\delta}^{T(j-1)}$ with parameters $\mathbf{s}_{T|T}$ and $\mathbf{P}_{T|T}$ obtained by running the Kalman filter for the state space system (85). For $t = T-1, \ldots, 1$ draw $\mathbf{h}_t^{(j)}$ from the normal full conditional posterior $h_T|\mathbf{y}^t, \boldsymbol{\mu}^{(j-1)}, \boldsymbol{\phi}^{(j-1)}, \mathbf{V}_\eta^{(j-1)}, \boldsymbol{\delta}^{T(j-1)}, \mathbf{h}_{t+1}^{(j)}$ with parameters obtained from the simulation smoother.

5. For $i = 1, \ldots, m$, $t = 1, \ldots, T$ draw the states $\delta_{it}^{(j)}$ from the full conditional posterior $\delta_{ij}|y_{it}^{**}, h_{it}$ in (87).

6. Draw $\mathbf{V}_\eta^{(j)}$ from the full conditional posterior $\mathbf{V}_\eta|\mathbf{h}^{T(j)}, \boldsymbol{\mu}^{(j-1)}, \boldsymbol{\phi}^{(j-1)}$ in (86).

7. Draw $\boldsymbol{\phi}^{(j)}$ from the full conditional posterior $\boldsymbol{\phi}|\mathbf{h}^{T(j)}, \mathbf{V}_\eta^{(j)}, \boldsymbol{\mu}^{(j-1)}$ in (88).

8. Draw $\boldsymbol{\mu}^{(j)}$ from the full conditional posterior $\boldsymbol{\mu}|\mathbf{h}^{T(j)}, \mathbf{V}_\eta^{(j)}, \boldsymbol{\phi}^{(j)}$ in (89).

9. If $j > B$
   For $h = 1, \ldots, H$, generate $\boldsymbol{\gamma}_{T+h}^{(j)}$ from the state equation if $\boldsymbol{\gamma}$ is time varying, generate $\mathbf{l}_{i,T+h}^{(j)}$ from (81) and $\mathbf{h}_{T+h}^{(j)}$ from (79), form $\boldsymbol{\Psi}_{T+h}^{(j)}$ and generate $\mathbf{u}_{T+h}^{(j)}$ from $\mathbf{u}_{T+h} \sim N\left(0, \boldsymbol{\Psi}_{T+h}^{(j)}\right)$ and calculate recursively

$$\widetilde{\mathbf{y}}_{T+h}^{(j)\prime} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}_{T+h-i}^{(j)\prime} \mathbf{A}_{i,T+h}^{(j)} + \sum_{i=h}^{p} \mathbf{y}_{T+h-i}^{\prime} \mathbf{A}_{i,T+h}^{(j)} + \mathbf{x}_{T+h}^{\prime} \mathbf{C}_{T+h}^{(j)} + \mathbf{u}_{T+h}^{(j)\prime}. \tag{90}$$

   Discarding the parameters yields $\left\{\widetilde{\mathbf{y}}_{T+1}^{(j)}, \ldots \widetilde{\mathbf{y}}_{T+H}^{(j)}\right\}_{j=1}^{R}$ as a sample of independent draws from the joint predictive distribution.

steady state BVAR at shorter lead times.

Using a real time data set D'Agostino, Gambetti and Giannone (forthcoming) forecasts the US unemployment rate, inflation and a short term interest rate. The aim of the forecasting exercise is to investigate how important it is to allow for time-varying parameters and stochastic volatility. Forecasts are thus made with a number of models which incorporate these futures to a varying degree: A univariate TVPSV-AR model, SV-AR and SV-VAR models, standard AR and VAR models estimated using recursive and rolling data windows and a TVPSV-VAR using the specification of Primiceri (2005). The inference is Bayesian for all the models and the prior beliefs are based on the Minnesota prior. Overall the TVPSV-VAR does best both in terms of point forecasts and density forecasts. The SV-AR and SV-VAR models improve on their constant variance counterparts and D'Agostino et al. (forthcoming) concludes that there is a role for both time-varying parameters and time-varying error variances when forecasting these variables.

8 Model and variable selection

Model specification in VAR models essentially consists of two questions: Which variables should be modelled (included in $\mathbf{y}_t$) and, given the content of $\mathbf{y}_t$, how many lags of $\mathbf{y}_t$ should be included? The answer to the first question obviously depends on the objective of the exercise which mandates the inclusion of some variables (e.g. the variables to be forecasted) but there is usually also a secondary set of variables where the choice is not so clear cut. These are variables that are not of primary interest but could be included in the model if it leads to a better specified model that improves the forecasts of the variables of interest or clearer inference by avoiding omitted variable bias.

The Litterman prior tries to answer the second question in its general form, how many lags, by making the prior tighter around zero for coefficients on larger lags and thus allowing for a comfortably large lag length while reducing the risk of overfitting. The question can, however, be made more specific: Which lags of which variables should be included in each equation? This opens up for a huge number of different model specifications and requires new tools.

8.1 Restricting the parameter matrices - SSVS

George, Sun and Ni (2008) considers soft and hard restrictions on the parameters of the VAR as a means of reducing the effects of overparameterization and sharpening the inference. This has some similarity with the structural VAR models discussed in section 4 but can also bee seen as a matter of selecting which lags of which dependent variable to include in the model. In contrast with the SVAR approach the restrictions are determined by the data rather than economic theory and are applied to a mixture of reduced form and structural form parameters. Taking the reduced form VAR

$$\mathbf{y}_t' = \sum_{i=1}^{p} \mathbf{y}_{t-i}' \mathbf{A}_i + \mathbf{x}_t' \mathbf{C} + \mathbf{u}_t' = \mathbf{z}_t' \mathbf{\Gamma} + \mathbf{u}_t'$$

as the starting point George et al. (2008) considers restrictions on $\mathbf{\Gamma}$ and the Cholesky factor $\mathbf{\Lambda}$ of the inverse variance matrix of $\mathbf{u}_t$, $\mathbf{\Psi}^{-1} = \mathbf{\Lambda}\mathbf{\Lambda}'$. The $\mathbf{\Lambda}$-matrix plays the

same role as in the SVAR and restrictions on $\mathbf{\Lambda}$ can, in contrast to restrictions on $\mathbf{\Gamma}$, be given a structural interpretation. Allowing for zero or "near-zero" restrictions on arbitrary parameters there is vast number of combinations of restrictions or models, $2^{mk+m(m-1)/2}$, to consider. It is clearly impossible to evaluate all of them and George et al. (2008) proposes a stochastic search variable selection (SSVS) procedure that will focus on the restrictions with empirical support. The SSVS (George and McCulloch (1993)) is a MCMC algorithm for simulating from the joint posterior distribution of the set of restrictions (or models) and parameters based on a specific form of hierarchical prior distribution. For the regression coefficients γ_{ij} let δ_{ij} be an indicator variable, the prior conditional on δ_{ij} is then $\gamma_{ij} \sim N\left(\underline{\gamma}_{ij}\delta_{ij}, h_{ij}^2\right)$ for

$$h_{ij} = \begin{cases} \tau_{0,ij} \text{ if } \delta_{ij} = 0 \\ \tau_{1,ij} \text{ if } \delta_{ij} = 1 \end{cases} \tag{91}$$

with $\tau_{0,ij} \ll \tau_{1,ij}$. The idea being that the prior shrinks aggressively towards zero if $\delta_{ij} = 0$ (or imposes $\gamma_{ij} = 0$ if $\tau_{0,ij}^2 = 0$) and allows for a non-zero γ_{ij} if $\delta_{ij} = 1$ by setting $\tau_{1,ij}$ relatively large. The hierarchical prior is completed by specifying independent Bernoulli priors for $\delta_{ij}, P(\delta_{ij} = 1) = p_{ij}$, where p_{ij} reflects the strength of the prior belief that γ_{ij} differs from zero in a meaningful way. Note that a parameter/variable can be forced into the model by setting p_{ij} to 1. For convenience (and to allow for prior correlation) we write the prior as a multivariate normal distribution for $\boldsymbol{\gamma} = \text{vec}(\mathbf{\Gamma})$,

$$\boldsymbol{\gamma}|\boldsymbol{\delta} \sim N\left(\mathbf{D}\underline{\gamma}, \mathbf{HRH}\right), \tag{92}$$

where $\mathbf{D} = \text{diag}(\delta_{11}, \delta_{21}, \ldots, \delta_{km})$, $\mathbf{H} = \text{diag}(h_{11}, h_{21}, \ldots, h_{km})$ and $\mathbf{R}$ is a known correlation matrix. With prior independence between $\mathbf{\Gamma}$, $\boldsymbol{\delta}$ and $\mathbf{\Psi}$ (or $\mathbf{\Lambda}$) the full conditional posterior for $\boldsymbol{\gamma}$ is standard and has the same form as with the independent normal-Wishart prior,

$$\begin{aligned} \boldsymbol{\gamma}|\mathbf{Y}_T, \mathbf{\Lambda}, \boldsymbol{\delta} &\sim N\left(\overline{\boldsymbol{\gamma}}, \overline{\mathbf{\Sigma}}_\gamma\right) \\ \overline{\mathbf{\Sigma}}_\gamma &= \left[(\mathbf{HRH})^{-1} + \mathbf{\Lambda}\mathbf{\Lambda}' \otimes \mathbf{Z}'\mathbf{Z}\right]^{-1} \\ \overline{\boldsymbol{\gamma}} &= \overline{\mathbf{\Sigma}}_\gamma \left[(\mathbf{HRH})^{-1} \mathbf{D}\overline{\boldsymbol{\gamma}} + \text{vec}\left(\mathbf{Z}'\mathbf{Y}\mathbf{\Lambda}\mathbf{\Lambda}'\right)\right]. \end{aligned} \tag{93}$$

George et al. (2008) gives the full conditional posterior for δ_{ij} as a Bernoulli distribution,

$$\begin{aligned} P(\delta_{ij} = 1|\mathbf{Y}_T, \mathbf{\Gamma}, \mathbf{\Lambda}, \boldsymbol{\delta}_{-ij}) &= \frac{u_{1,ij}}{u_{1,ij} + u_{0,ij}} \\ u_{1,ij} &= \pi(\mathbf{\Gamma}|\boldsymbol{\delta}_{-ij}, \delta_{ij} = 1)\, p_{ij} \\ u_{0,ij} &= \pi(\mathbf{\Gamma}|\boldsymbol{\delta}_{-ij}, \delta_{ij} = 0)\,(1 - p_{ij}) \end{aligned} \tag{94}$$

where $\pi(\mathbf{\Gamma}|\cdot)$ is the prior distribution (92). The simple form follows since, with the hierarchical prior structure, δ_{ij} is independent of the data once we condition on $\mathbf{\Gamma}$ (see George and McCulloch (1993)). If, as is frequently the case, the prior for $\boldsymbol{\gamma}$ is specified with no correlation between the elements, $\mathbf{R} = \mathbf{I}$, the expressions simplify further and we

have

$$u_{1,ij} = \frac{1}{\tau_{1,ij}} \exp\left(-\frac{\left(\gamma_{ij} - \underline{\gamma}_{ij}\right)^2}{2\tau_{1,ij}^2} \right) p_{ij}$$
$$u_{0,ij} = \frac{1}{\tau_{0,ij}} \exp\left(-\frac{\gamma_{ij}^2}{2\tau_{0,ij}^2} \right) (1 - p_{ij}) .$$

To facilitate similar selection among the off-diagonal elements of $\mathbf{\Lambda}$, collect these in vectors $\boldsymbol{\eta}_j = (\lambda_{1j}, \ldots, \lambda_{j-1,j})'$, $j = 2, \ldots, m$, (George et al. (2008) works with $\mathbf{\Lambda}$ upper triangular) and let $\boldsymbol{\omega}_j = (\omega_{1j}, \ldots, \omega_{j-1,j})'$ be the corresponding indicators. The conditional prior for $\boldsymbol{\eta}_j$ is then specified in the same fashion as the prior for $\boldsymbol{\gamma}$,

$$\boldsymbol{\eta}_j | \boldsymbol{\omega}_j \sim N\left(\mathbf{0}, \mathbf{G}_j \mathbf{R}_j \mathbf{G}_j\right) \tag{95}$$

with $\mathbf{G}_j = \operatorname{diag}\left(g_{1j}, \ldots, g_{j-1,j}\right)$, for

$$g_{ij} = \begin{cases} \kappa_{0,ij} \text{ if } \omega_{ij} = 0 \\ \kappa_{1,ij} \text{ if } \omega_{ij} = 1 \end{cases},$$

with $\kappa_{0,ij} \ll \kappa_{1,ij}$, and $\mathbf{R}_j$ a known correlation matrix. As for δ_{ij}, the prior for ω_{ij} is specified as independent Bernoulli distributions with $P\left(\omega_{ij} = 1\right) = q_{ij}$. For the diagonal elements of $\mathbf{\Lambda}$, $\boldsymbol{\lambda} = (\lambda_{11}, \ldots, \lambda_{mm})'$, George et al. (2008) specifies independent Gamma distributions for the square of the diagonal as the prior,

$$\lambda_{ii}^2 \sim G\left(a_i, b_i\right). \tag{96}$$

Note that an inverse-Wishart prior for $\mathbf{\Psi}$ can be obtained as a special case when there is no selection of zero elements in $\mathbf{\Lambda}$,i.e. $q_{ij} = 1 \ \forall i, j$, see Algorithm 20 for details.

In order to derive the full conditional posteriors George et al. (2008) rewrites the reduced form likelihood (7) as

$$\begin{aligned} L\left(\mathbf{Y} | \mathbf{\Gamma}, \mathbf{\Lambda}\right) &\propto \left|\det \mathbf{\Lambda}\right|^T \exp\left\{ -\frac{1}{2} \operatorname{tr}\left[\mathbf{\Lambda}' \mathbf{S} \mathbf{\Lambda}\right] \right\} \\ &= \prod_{i=1}^{m} \lambda_{ii}^T \exp\left\{ -\frac{1}{2} \left[\sum_{i=1}^{m} \lambda_{ii}^2 v_i + \sum_{j=2}^{m} \left(\boldsymbol{\eta}_j + \lambda_{jj} \mathbf{S}_{j-1}^{-1} \mathbf{s}_j\right)' \mathbf{S}_{j-1} \left(\boldsymbol{\eta}_j + \lambda_{jj} \mathbf{S}_{j-1}^{-1} \mathbf{s}_j\right) \right] \right\} \end{aligned}$$

where $\mathbf{S} = \left(\mathbf{Y} - \mathbf{Z}\mathbf{\Gamma}\right)' \left(\mathbf{Y} - \mathbf{Z}\mathbf{\Gamma}\right)$, $\mathbf{S}_j$ the upper left $j \times j$ submatrix of $\mathbf{S}$, $\mathbf{s}_j = (s_{1j}, \ldots, s_{j-1,j})'$, $v_1 = s_{11}$ and $v_j = \left|\mathbf{S}_j\right| / \left|\mathbf{S}_{j-1}\right| = s_{jj} - \mathbf{s}_j' \mathbf{S}_{j-1}^{-1} \mathbf{s}_j$ for $j = 2, \ldots, m$. It is then easy to show that the conditional posteriors for $\boldsymbol{\eta}_j$ are independent and normal,

$$\boldsymbol{\eta}_j | \mathbf{Y}_T, \mathbf{\Gamma}, \boldsymbol{\omega}, \boldsymbol{\lambda} \sim N\left(\overline{\boldsymbol{\eta}}_j, \overline{\mathbf{\Sigma}}_j\right) \tag{97}$$

with

$$\begin{aligned} \overline{\mathbf{\Sigma}}_j &= \left[\left(\mathbf{G}_j \mathbf{R}_j \mathbf{G}_j\right)^{-1} + \mathbf{S}_{j-1}\right]^{-1} \\ \overline{\boldsymbol{\eta}}_j &= -\overline{\mathbf{\Sigma}}_j \lambda_{jj} \mathbf{s}_j, \end{aligned}$$

and that the conditional posteriors for λ_{jj}^2 are independent Gamma distributions,

$$\begin{aligned}\lambda_{jj}^2|\mathbf{Y}_T,\boldsymbol{\Gamma},\boldsymbol{\omega} &\sim G\left(\overline{a}_j,\overline{b}_j\right) \\ \overline{a}_j &= a_j + T/2 \\ \overline{b}_j &= \begin{cases} b_1 + s_{11}/2, \ j=1 \\ b_j + \left(s_{jj} - \mathbf{s}_j'\overline{\boldsymbol{\Sigma}}_j^{-1}\mathbf{s}_j\right)/2, \ j=2,\ldots,m \end{cases}.\end{aligned} \tag{98}$$

The full conditional posteriors for ω_{ij}, finally, are Bernoulli distributions,

$$\begin{aligned}P\left(\omega_{ij}=1|\mathbf{Y}_T,\boldsymbol{\Gamma},\boldsymbol{\Lambda},\boldsymbol{\omega}_{-ij}\right) &= \frac{v_{1,ij}}{v_{1,ij}+v_{0,ij}} \\ v_{1,ij} &= \pi\left(\boldsymbol{\eta}_j|\boldsymbol{\omega}_{-ij},\omega_{ij}=1\right)q_{ij} \\ v_{0,ij} &= \pi\left(\boldsymbol{\eta}_j|\boldsymbol{\omega}_{-ij},\omega_{ij}=0\right)(1-q_{ij})\end{aligned} \tag{99}$$

where $\pi\left(\boldsymbol{\eta}_j|\cdot\right)$ is the prior distribution (95). If the elements of $\boldsymbol{\eta}_j$ are uncorrelated a prior ($\mathbf{R}_j=\mathbf{I}$) the expressions simplify

$$\begin{aligned}v_{1,ij} &= \frac{1}{\kappa_{1,ij}}\exp\left(-\frac{\eta_{ij}^2}{2\kappa_{1,ij}^2}\right)q_{ij} \\ v_{0,ij} &= \frac{1}{\kappa_{0,ij}}\exp\left(-\frac{\eta_{ij}^2}{2\kappa_{0,ij}^2}\right)(1-q_{ij}).\end{aligned}$$

Specifying the prior beliefs The prior "inclusion probabilities" determines the prior expected model size (number of non-zero parameters) and influences how aggressively the restrictions are applied. Setting $p_{ij}=q_{ij}=1/2$ is a reasonable starting point but a smaller value can be useful with large and richly parameterized models. There are usually some parameters, such as the constant term, that should always be in the model. This is achieved by setting the corresponding prior inclusion probability to 1. We might also have substantive information about how likely it is that a parameter will contribute to model fit and forecast performance. In VAR models it could, for example, be useful to let the inclusion probability p_{ij} decrease with the lag length in the spirit of the Minnesota prior. The choice of inclusion probabilities for the variance parameters could be guided by the same type of considerations that leads to restrictions in structural VAR models.

The prior variances $\tau_{0,ij}$ and $\kappa_{0,ij}$ should be sufficiently small to effectively shrink the parameter to zero when δ_{ij} or ω_{ij} are zero. The choice of $\tau_{1,ij}$ and $\kappa_{1,ij}$ is more difficult. George et al. (2008) suggests a semiautomatic choice with $\tau_{0,ij}=\widehat{\sigma}_{\gamma_{ij}}/10$ and $\tau_{1,ij}=10\widehat{\sigma}_{\gamma_{ij}}$where $\widehat{\sigma}_{\gamma_{ij}}$ is the standard error of the OLS estimate of γ_{ij} in the unrestricted model. Alternatively $\tau_{1,ij}$ can be based on the Minnesota prior and set as in (14).

The correlation matrices $\mathbf{R}$ and $\mathbf{R}_j$, $j=2,\ldots,m$, are usefully set to the identity matrix unless there is substantial prior information about the correlation structure. It is also standard practice in SSVS applications to set the prior means $\underline{\gamma}_{ij}$ to zero when $\delta_{ij}=1$ in addition to when $\delta_{ij}=0$. With VAR models it can be useful to deviate from this and set the prior mean for the first own lag to a non-zero value in the spirit of the Minnesota prior.

If no restriction search is wanted for the regression parameters the prior for $\boldsymbol{\Gamma}$ reduces to the independent normal prior in section 3.2.2. The prior for $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ can be overly

complicated if no restriction search is to be conducted on $\boldsymbol{\eta}$ and these priors can usefully be replaced by a Jeffreys' prior or an inverse Wishart prior on $\boldsymbol{\Psi}$ as in section 3.2.2.

Simulating from the posterior With the conditional posterior distributions in hand it is straightforward to implement a Gibbs sampler (see Algorithm 14) for the joint posterior distribution and the predictive distributions needed in forecasting applications. This will effectively conduct model averaging over the different models implied by the restrictions and produces the model averaged posterior distribution. If the variable/restriction selection is of interest the indicator variables δ_{ij} and ω_{ij} will provide evidence on this. The posterior probability that a parameter is non-zero can be estimated by averaging δ_{ij} and ω_{ij} (or, for a more precise estimate, average the posterior probabilities $P(\delta_{ij} = 1)$ and $P(\omega_{ij} = 1)$ in (94) and (99)) over the output of the sampler.

Note that the sampler in Algorithm 14 will not converge to the joint posterior if hard restrictions ($\tau_{0,ij} = 0$ or $\kappa_{0,ij} = 0$) are used and will converge very slowly if the ratios $\tau_{1,ij}/\tau_{0,ij}$ or $\kappa_{1,ij}/\kappa_{0,ij}$ are very large. The MCMC algorithm suggested by Geweke (1996*b*) is a better choice in these cases. Korobilis (forthcoming*b*) suggests a convenient algorithm for the case with hard restrictions on the regression parameters and no restriction search on the variance parameters.

Forecast performance Korobilis (2008) applies the SSVS in a forecasting exercise where the base model is a VAR with eight US macroeconomic variables which is augmented with an additional 124 exogenous variables that are entered into the model in the form of their principal components. He finds that the SSVS model averaged predictions, as well as the predictions from the "median model" (i.e. the model containing the variables with posterior inclusion probabilities greater than 0.5, Barbieri and Berger (2004)), improves on the forecasts from OLS estimated VARs without the additional variables and model selection using BIC, the Bayesian information criteria of Schwarz (1978).

Jochmann, Koop and Strachan (2010) extends the SSVS restriction search to VAR models with Markov switching to allow for structural breaks and conducts a forecasting exercise comparing models allowing for different combinations of restriction searches and breaks in the regression and variance parameters. Using a 4 lag VAR with US unemployment, interest rate and inflation they find that the restriction search which effectively sets a large number of parameters to zero results in improved forecasts compared to BVARs with a "loose" prior (obtained by forcing $\delta_{ij} = 1$ and $\omega_{ij} = 1$ for all parameters in the SSVS prior) and a Minnesota prior. Allowing for structural breaks also improves on performance in combination with SSVS if only a subset (either $\boldsymbol{\Gamma}$ or $\boldsymbol{\Lambda}$) of the parameters are allowed to change.

Korobilis (forthcoming*b*) considers SSVS in a richer class of multivariate time series models than just linear VAR models but limits the restriction search to the conditional mean parameters $\boldsymbol{\Gamma}$ and consider hard rather than soft restrictions, corresponding to $\tau_{0,ij} = 0$ in (91). In a forecasting exercise where the aim is to forecast UK unemployment, interest rate and inflation Korobilis uses a range of models, allowing for structural breaks or time varying parameters, and prior specifications. The general conclusion is that the restriction search does improve forecast performance when the prior is informative, the model is richly parameterized or quite large.

Algorithm 14 Gibbs sampler for stochastic restriction search (SSVS)

For the priors (92), (95), (96) and independent Bernoulli priors on δ_{ij} and ω_{ij} the following Gibbs sampler (George et al. (2008)) can be used to simulate the joint posterior distribution of $\mathbf{\Gamma}$, $\mathbf{\Lambda}$, $\boldsymbol{\delta}$ and $\boldsymbol{\omega}$. Select starting values $\boldsymbol{\gamma}^{(0)}$, $\boldsymbol{\delta}^{(0)}$, $\boldsymbol{\eta}^{(0)}$ and $\boldsymbol{\omega}^{(0)}$.
For $j = 1, \ldots, B + R$

1. Generate $\boldsymbol{\lambda}^{(j)}$ by drawing λ_{ii}^2, $i = 1, \ldots, m$ from the full conditional $\lambda_{ii}^2 | \mathbf{Y}_T, \boldsymbol{\gamma}^{(j-1)}, \boldsymbol{\omega}^{(j-1)} \sim G\left(\overline{a}_j, \overline{b}_j\right)$ in (98).

2. Generate $\boldsymbol{\eta}_i^{(j)}$, $i = 2, \ldots, m$ from the full conditional $\boldsymbol{\eta}_i | \mathbf{Y}_T, \boldsymbol{\gamma}^{(j-1)}, \boldsymbol{\omega}^{(j-1)}, \boldsymbol{\lambda}^{(j)} \sim N\left(\overline{\boldsymbol{\eta}}_j, \overline{\mathbf{\Sigma}}_j\right)$ in (97)

3. Generate $\omega_{ik}^{(j)}$, $i = 1, \ldots, k-1$, $k = 2, \ldots, m$ from the full conditional $\omega_{ik} | \boldsymbol{\eta}_k^{(j)}, \omega_{1k}^{(j)}, \ldots \omega_{i-1,k}^{(j)}, \omega_{i+1,k}^{(j-1)}, \ldots, \omega_{k-1,k}^{(j-1)} \sim Ber\left(v_{1,ik} / \left(v_{1,ik} + v_{0,ik}\right)\right)$ in (99).

4. Generate $\boldsymbol{\gamma}^{(j)}$ from the full conditional $\boldsymbol{\gamma} | \mathbf{Y}_T, \boldsymbol{\eta}^{(j)}, \boldsymbol{\lambda}^{(j)}, \boldsymbol{\delta}^{(j-1)} \sim N\left(\overline{\boldsymbol{\gamma}}, \overline{\mathbf{\Sigma}}_\gamma\right)$ in (93).

5. Generate $\delta_{il}^{(j)}$, $i = 1, \ldots, k$, $l = 1, \ldots, m$ from the full conditional posterior $\delta_{il} | \boldsymbol{\gamma}^{(j)}, \delta_{11}^{(j)}, \ldots \delta_{i-1,l}^{(j)}, \delta_{i+1,l}^{(j-1)}, \ldots, \delta_{mk}^{(j-1)} \sim Ber\left(u_{1,il} / \left(u_{1,il} + u_{0,il}\right)\right)$ in (94)

6. If $j > B$ form $\mathbf{\Psi}^{(j)}$ from $\boldsymbol{\eta}^{(j)}$ and $\boldsymbol{\lambda}^{(j)}$, generate $\mathbf{u}_{T+1}^{(j)}, \ldots, \mathbf{u}_{T+H}^{(j)}$ from $\mathbf{u}_t \sim N\left(0, \mathbf{\Psi}^{(j)}\right)$ and calculate recursively

$$\widetilde{\mathbf{y}}_{T+h}^{(j)\prime} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}_{T+h-i}^{(j)\prime} \mathbf{A}_i^{(j)} + \sum_{i=h}^{p} \mathbf{y}_{T+h-i}^{\prime} \mathbf{A}_i^{(j)} + \mathbf{x}_{T+h}^{\prime} \mathbf{C}^{(j)} + \mathbf{u}_{T+h}^{(j)\prime}.$$

   Discarding the parameters yields $\left\{\widetilde{\mathbf{y}}_{T+1}^{(j)}, \ldots \widetilde{\mathbf{y}}_{T+H}^{(j)}\right\}_{j=B+1}^{B+R}$ as a sample from the joint predictive distribution.

If there is no restriction search on the elements of $\boldsymbol{\eta}$, the priors (95) and (96) can be replaced by a Jeffreys' prior (23) for $\mathbf{\Psi}$ or an inverse Wishart, $\mathbf{\Psi} \sim iW\left(\underline{\mathbf{S}}, \underline{v}\right)$. Steps 1-3 can then be replaced by a draw from the full conditional posterior $\mathbf{\Psi} | \mathbf{Y}_T, \mathbf{\Gamma}^{(j-1)} \sim iW\left(\overline{\mathbf{S}}, \overline{v}\right)$ with parameters given in (25) or (26).
If there is no restriction search on the elements of $\mathbf{\Gamma}$, the prior (92) reduces to an independent normal prior that does not depend on $\boldsymbol{\delta}$ and step 5 can be omitted.

8.2 Selecting variables to model

The standard Bayesian approach to model selection is based on the marginal likelihood (4) and runs into problems when the issue is which variables to include as dependent variables in a multivariate model. The likelihoods are simply not comparable when variables are added to or dropped from $\mathbf{y}_t$. In forecasting applications there is an additional consideration. We are, in general, not interested in how well the model as a whole fits the data only in how well it forecasts a core set of variables of interest. Other variables are then only included if they are expected to improve the forecast performance.

8.2.1 Marginalized predictive likelihoods

Andersson and Karlsson (2009) suggests replacing the marginal likelihood with the predictive likelihood for the variables of interest, that is after marginalizing out the other variables, in the calculation of posterior "probabilities" or model weights. This creates a focused measure that can be used for model selection or forecast combination and is attractive in a forecasting context since it directly addresses the forecasting performance of the different models

The predictive likelihood approach is based on a split of the data into two parts, the training sample, $\mathbf{Y}_n^* = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_n')'$ of size n, and an evaluation or hold out sample, $\widetilde{\mathbf{Y}}_n = \left(\mathbf{y}_{n+1}', \mathbf{y}_{n+2}', \ldots, \mathbf{y}_T'\right)'$ of size $T-n$. The training sample is used to convert the prior into a posterior and the predictive likelihood for the hold out sample is obtained by marginalizing out the parameters from the joint distribution of data and parameters,

$$p\left(\widetilde{\mathbf{Y}}_n \middle| \mathbf{Y}_n^*, \mathcal{M}_i\right) = \int L\left(\widetilde{\mathbf{Y}}_n \middle| \boldsymbol{\theta}_i, \mathbf{Y}_n^*, \mathcal{M}_i\right) p\left(\boldsymbol{\theta}_i \middle| \mathbf{Y}_n^*, \mathcal{M}_i\right) d\boldsymbol{\theta}_i.$$

Partitioning the hold out sample data into the variables of interest and the remaining variables, $\widetilde{\mathbf{Y}}_n = \left(\widetilde{\mathbf{Y}}_{1,n}, \widetilde{\mathbf{Y}}_{2,n}\right)$, the marginalized predictive likelihood for the variables of interest is obtained by marginalizing out $\widetilde{\mathbf{Y}}_{2,n}$,

$$MPL\left(\widetilde{\mathbf{Y}}_{1,n} \middle| \mathbf{Y}_n^*, \mathcal{M}_i\right) = \int p\left(\widetilde{\mathbf{Y}}_n \middle| \mathbf{Y}_n^*, \mathcal{M}_i\right) d\widetilde{\mathbf{Y}}_{2,n}.$$

Predictive weights that can be used for model averaging or model selection are then calculated as

$$w\left(\mathcal{M}_i \middle| \widetilde{\mathbf{Y}}_{1,n}, \mathbf{Y}_n^*\right) = \frac{MPL\left(\widetilde{\mathbf{Y}}_{1,n} \middle| \mathbf{Y}_n^*, \mathcal{M}_i\right) p\left(\mathcal{M}_i\right)}{\sum_{j=1}^{M} MPL\left(\widetilde{\mathbf{Y}}_{1,n} \middle| \mathbf{Y}_n^*, \mathcal{M}_j\right) p\left(\mathcal{M}_j\right)} \tag{100}$$

where $MPL\left(\widetilde{\mathbf{Y}}_{1,n} \middle| \mathbf{Y}_n^*, \mathcal{M}_i\right)$ is evaluated at the observed values of the variables of interest in the hold out sample.

While the predictive weights (100) strictly speaking can not be interpreted as posterior probabilities they have the advantage that proper prior distributions are not required for the parameters. The predictive likelihood is, in contrast to the marginal likelihood, well defined as long as the posterior distribution of the parameters conditioned on the training sample is proper.

The use of the predictive likelihood is complicated by the dynamic nature of VAR models. As noted by Andersson and Karlsson (2009) the predictive likelihood is the joint predictive distribution over lead times $h = 1$ to $T - n$. This will become increasingly uninformative for larger lead times and unrepresentative of lead times such as $h = 4$ or 8 usually considered in macroeconomic forecasting. At the same time the hold out sample needs to be relatively large in order to provide a sound basis for assessing the forecast performance of the models. To overcome this Andersson and Karlsson suggested focusing the measure to specific lead times $h_1, \ldots, h_k$ and using a series of predictive likelihoods,

$$g\left(\widetilde{\mathbf{Y}}_{1,n}|\mathcal{M}_i\right) = \prod_{t=n}^{T-h_k} MPL\left(y_{1,t+h_1}, \ldots, y_{1,t+h_k}|\,\mathbf{Y}_t^*, \mathcal{M}_i\right), \tag{101}$$

in the calculation of the predictive weights.

A final complication is that the predictive likelihood is not available in closed form for lead times $h > 1$ and must be estimated using simulation methods. With a normal likelihood the predictive likelihood for a VAR model will be normal conditional on the parameters and easy to evaluate. Andersson and Karlsson suggested estimating the multiple horizon marginalized predictive likelihood using a Rao-Blackwellization technique as

$$\widehat{MPL}\left(y_{1,t+h_1}, \ldots, y_{1,t+h_k}|\,\mathbf{Y}_t^*, \mathcal{M}_i\right) = \frac{1}{R}\sum_{i=1}^{R} p\left(y_{1,t+h_1}, \ldots, y_{1,t+h_k}|\,\mathbf{Y}_t^*, \mathcal{M}_i, \boldsymbol{\theta}_i^{(j)}\right)$$

by averaging the conditional predictive likelihood $p\left(y_{1,t+h_1}, \ldots, y_{1,t+h_k}|\,\mathbf{Y}_t^*, \mathcal{M}_i, \boldsymbol{\theta}_i\right)$ over draws, $\boldsymbol{\theta}_i^{(j)}$, of the parameters from the posterior distribution based on $\mathbf{Y}_t^*$. This leads to estimated predictive weights

$$\widehat{w}\left(\mathcal{M}_i|\,\widetilde{\mathbf{Y}}_{1,n}, \mathbf{Y}_n^*\right) = \frac{\widehat{g}\left(\mathbf{Y}, n|\mathcal{M}_i\right) p\left(\mathcal{M}_i\right)}{\sum_{j=1}^{M} \widehat{g}\left(\widetilde{\mathbf{Y}}_{1,n}|\mathcal{M}_j\right) p\left(\mathcal{M}_j\right)} \tag{102}$$

with

$$\widehat{g}\left(\widetilde{\mathbf{Y}}_{1,n}|\mathcal{M}_i\right) = \prod_{t=n}^{T-h_k} \widehat{MPL}\left(y_{1,t+h_1}, \ldots, y_{1,t+h_k}|\,\mathbf{Y}_t^*, \mathcal{M}_i\right).$$

The marginalized predictive likelihood procedure is thus in principle applicable to any forecasting model, with any type of prior, as long as the likelihood is normal and it is possible to simulate the posterior distribution of the parameters.

Forecasting performance Andersson and Karlsson (2009) conducted a forecasting exercise with the aim of forecasting US GDP growth and considered VAR models with up to four variables selected from a set of 19 variables (including GDP). Compared to an AR(2) benchmark forecast combination using the predictive weights (102) does better for shorter lead times (up to 4 quarters) but is outperformed for lead times 5 to 8. Selecting a single model based on the predictive weights does slightly worse than the forecast combination for the shorter lead times but performs better and on par with the AR(2) for the longer lead times.

8.2.2 Marginal likelihoods via Bayes factors

Jarociński and Maćkowiak (2011) favours the marginal likelihood as a basis for model comparison and notes that the question of wether a set of variables is useful for forecasting the variables of interest can be addressed in a model containing all entertained variables. To see this write the VAR model in terms of two sets of variables $\mathbf{y}_{1,t}$ and $\mathbf{y}_{2,t}$ where $\mathbf{y}_{1,t}$ contains the variables of interest and, possibly, some additional variables and $\mathbf{y}_{2,t}$ contains the remaining variables,

$$\left(\mathbf{y}_{1,t}', \mathbf{y}_{2,t}'\right) = \sum_{i=1}^{p} \left(\mathbf{y}_{1,t-i}', \mathbf{y}_{2,t-i}'\right) \mathbf{A}_i + \mathbf{x}_t'\mathbf{C} + \mathbf{u}_t'$$

with

$$\mathbf{A}_i = \begin{pmatrix} \mathbf{A}_{i,11} & \mathbf{A}_{i,12} \\ \mathbf{A}_{i,21} & \mathbf{A}_{i,22} \end{pmatrix}$$

partitioned conformably. The notion that $\mathbf{y}_{2,t}$ is not useful for predicting $\mathbf{y}_{1,t}$ (does not Granger-cause $\mathbf{y}_{1,t}$) then corresponds to the block-exogeneity restriction that $\mathbf{A}_{i,21} = 0$, $\forall i$. If the restriction holds $\mathbf{y}_{1,t}$ can be modelled as a function of its own lags and $\mathbf{y}_{2.t}$ is not needed. Each partition of the variables into $\mathbf{y}_{1,t}$ and $\mathbf{y}_{2,t}$ gives rise to a different block-exogeneity restriction and the idea of Jarociński and Maćkowiak (2011) is to compute the marginal likelihood for all the variables under the different restrictions and base model selection or model averaging on these in a standard fashion.

This approach overcomes the main problem with the marginal likelihood of comparing apples with oranges when different sets of left hand variables are considered. Unfortunately, the marginal likelihood under the restrictions is rarely available in closed form. The marginal likelihoods and posterior model probabilities can, however, be computed indirectly by way of the Savage-Dickey density ratio (Dickey (1971)),

$$BF_{R,U} = \frac{m\left(\mathbf{Y}|\mathbf{A}_{i,21} = 0\right)}{m\left(\mathbf{Y}\right)} = \frac{p_{A_{i,21}}\left(\mathbf{A}_{i,21} = 0|\mathbf{Y}\right)}{\pi_{A_{i,21}}\left(\mathbf{A}_{i,21} = 0\right)}, \tag{103}$$

which relates the Bayes factor comparing the restricted model and the unrestricted model to the ratio of the marginal posterior to the marginal prior for the restricted parameters evaluated under the restriction. The second equality in (103) holds under the specific condition,

$$\pi_R\left(\mathbf{\Gamma}_R, \mathbf{\Psi}\right) = \pi_U\left(\mathbf{\Gamma}_R, \mathbf{\Psi}|\mathbf{A}_{i,21} = 0\right), \tag{104}$$

that the prior for the parameters in the restricted model equals the prior for the unrestricted model when conditioning on the restriction.

Jarociński and Maćkowiak (2011) suggests a normal-Wishart prior (section 3.2.1) as a prior for the unrestricted model and constructs the prior for the restricted models by conditioning on the restriction to ensure that condition (104) holds.[25] With the conjugate normal-Wishart prior both the marginal prior and posterior in the unrestricted model will be matricvariate-t distributions. Partition $\mathbf{y}_t$ into the n_1 variables $\mathbf{y}_{1,t}$ and the n_2 variables

[25]The resulting prior for the restricted model is not normal-Wishart and presumably different from what would be used when estimating a model for just $\mathbf{y}_{1,t}$.

$\mathbf{y}_{2,t}$ and let $\mathbf{P}$ and $\mathbf{Q}$ be the matrices of dimension $pn_2 \times k$ and $m \times n_1$ that selects the rows and columns of $\mathbf{\Gamma}$ corresponding to $\mathbf{A}_{i,21}$,

$$\begin{pmatrix} \mathbf{A}_{1,21} \\ \vdots \\ \mathbf{A}_{p,21} \end{pmatrix} = \mathbf{P\Gamma Q}.$$

For the prior we have (see Appendix C) $\mathbf{P\Gamma Q}|\mathbf{\Psi} \sim MN_{pn_2,n_1}\left(\mathbf{P}\underline{\mathbf{\Gamma}}\mathbf{Q}, \mathbf{Q}'\mathbf{\Psi Q}, \mathbf{P}\underline{\mathbf{\Omega}}\mathbf{P}'\right)$ and $\mathbf{P\Gamma Q} \sim Mt_{pn_2,n_1}\left(\mathbf{P}\underline{\mathbf{\Gamma}}\mathbf{Q}, \left(\mathbf{P}\underline{\mathbf{\Omega}}\mathbf{P}'\right)^{-1}, \mathbf{Q}'\underline{\mathbf{S}}\mathbf{Q}, \underline{v} - n_2\right)$ using that $\mathbf{Q}'\mathbf{\Psi Q} \sim iW\left(\mathbf{Q}'\underline{\mathbf{S}}\mathbf{Q}, \underline{v} - n_2\right)$ since $\mathbf{Q}$ is a selection matrix. An equivalent result holds for the posterior and the Bayes factor can be obtained as

$$BF_{R,U} = \prod_{i=1}^{n_1} \frac{\Gamma\left(\left(\underline{v} - n_2 + 1 - i\right)/2\right)\Gamma\left(\left(\overline{v} - n_2 + pn_2 + 1 - i\right)/2\right)}{\Gamma\left(\left(\underline{v} - n_2 + pn_2 + 1 - i\right)/2\right)\Gamma\left(\left(\overline{v} - n_2 + 1 - i\right)/2\right)} \tag{105}$$
$$\times \frac{\left|\mathbf{P}\underline{\mathbf{\Omega}}\mathbf{P}'\right|^{n_1/2}\left|\mathbf{Q}'\underline{\mathbf{S}}\mathbf{Q}\right|^{-(\underline{v}-n_2)/2}\left|\mathbf{Q}'\overline{\mathbf{S}}\mathbf{Q} + \left(\mathbf{P}\overline{\mathbf{\Gamma}}\mathbf{P}'\right)'\left(\mathbf{P}\overline{\mathbf{\Omega}}\mathbf{P}\right)^{-1}\left(\mathbf{P}\overline{\mathbf{\Gamma}}\mathbf{P}'\right)\right|^{-(\overline{v}-n_2+pn_2)/2}}{\left|\mathbf{P}\overline{\mathbf{\Omega}}\mathbf{P}'\right|^{n_1/2}\left|\mathbf{Q}'\overline{\mathbf{S}}\mathbf{Q}\right|^{-(\overline{v}-n_2)/2}\left|\mathbf{Q}'\underline{\mathbf{S}}\mathbf{Q} + \left(\mathbf{P}\underline{\mathbf{\Gamma}}\mathbf{P}'\right)'\left(\mathbf{P}\underline{\mathbf{\Omega}}\mathbf{P}\right)^{-1}\left(\mathbf{P}\underline{\mathbf{\Gamma}}\mathbf{P}'\right)\right|^{-(\underline{v}-n_2+pn_2)/2}}.$$

The posterior model probabilities can be obtained directly from the Bayes factors or via the marginal likelihoods for the restricted and unrestricted models by noting that the marginal likelihood for the unrestricted model is the matricvariate-t distribution given in (20).

Jarociński and Maćkowiak (2011) also considered models defined by multiple block-exogeneity restrictions where the Bayes factor is not available in closed form but can be evaluated using Monte Carlo methods. While the Bayes factor (105) is easy to evaluate the computations can be prohibitive if the number of considered variables is large – especially if multiple block-exogeneity restrictions are considered – and Jarociński and Maćkowiak (2011) proposes a Markov chain Monte Carlo model composition, $(\text{MC})^3$, scheme (Madigan and York (1995)) to identify the most promising models.

9 High Dimensional VARs

Most applications involve relatively small VAR models with up to 5 or 6 variables and occasionally 10 or more variables. There are obvious reasons for this – the number of parameters to estimate grows rapidly with m and p and can exhaust the information in the data while the models get unwieldy and sometimes difficult to interpret. There are, however, occasions that more or less demands that a large number of variables are modelled jointly. The earliest such example is perhaps panel studies, e.g. to forecast regional economic development one could specify a small VAR-model for each region or specify a joint VAR for the panel of regions that allows for interaction between the regions. In this case it is not only the increased size of the model that contributes to the complexity, there is also the need to take account of heterogeneity across regions and, perhaps, time, see Canova and Ciccarelli(2004, 2009).

A second situation is the task of forecasting in a data-rich environment with "wide" data sets that can contain 100 or more variables with potential predictive content for the variables of interest. This has typically been tackled with dynamic factor models (Stock

and Watson (2002), Forni, Hallin, Lippi and Reichlin (2003)) where the information in the data is summarized by a few factors or by combining forecasts from small models with different combinations of predictor variables (see Stock and Watson (2006) for a review). Recent studies do, however, indicate that large Bayesian VAR models can be quite competitive.

VAR models for wide data sets face numerical challenges due to the sheer size of the model, with $m = 100$ variables and $p = 4$ lags there are 40 000 parameters in $\mathbf{\Gamma}$. OLS estimation is still feasible provided that $T > 400$ since this "only" involves inversion of the 400×400 matrix $\mathbf{Z}'\mathbf{Z}$ although estimates will be very imprecise due to the large number of parameters. Similarly, Bayesian analysis with a normal-Wishart prior benefits from the Kronecker structure and is computationally feasible while more general priors such as the normal-diffuse or independent normal Wishart are faced with the inversion of a 40 000×40 000 matrix. The sheer size of the problems makes MCMC exercises impractical and too time consuming with current desktop resources even if one ignores the issue of numerical stability when solving high dimensional equation systems.[26]

In line with the dynamic factor model literature De Mol, Giannone and Reichlin (2008) considers "direct" univariate forecasting models of the form

$$y_{t+h} = \mathbf{x}_t'\boldsymbol{\beta}_h + u_{t+h}$$

where $\mathbf{x}_t$ contains (a large number of) variables believed to be useful when forecasting y_t. Compared to a truly dynamic specification (e.g. a VAR) this has the advantage that there is no need to forecast $\mathbf{x}_t$ for lead times $h > 1$. The disadvantage is that the distribution of the error term is more complicated, in general u_{t+h} follows a MA$(h-1)$ process and that separate equations must be estimated for each lead time. In a small forecasting exercise with $n = 131$ potential predictors in $\mathbf{x}_t$ they demonstrate that principal component regression (that is a dynamic factor model) and Bayesian forecasts based on a normal or double exponential prior for $\boldsymbol{\beta}_h$ are viable methods for dealing with very large data sets. When n is large there is a considerable risk of overfitting and, pragmatically, the success of the Bayesian approach depends on applying an appropriate amount of shrinkage in the prior. De Mol et al. (2008) analyses the behavior of the forecasts as both n and $T \to \infty$ under the assumption that the data can be described by a factor structure, $y_{t+h} = \mathbf{f}_t'\boldsymbol{\gamma} + e_{t+h}$, $\mathbf{x}_t = \mathbf{\Lambda}\mathbf{f}_t + \boldsymbol{\xi}_t$ where $\mathbf{f}_t$ contains the r common factors. They show that the Bayes forecast $y_t(h) = \mathbf{x}_t'\overline{\boldsymbol{\beta}}$ for $\overline{\boldsymbol{\beta}}$ the posterior mean of $\boldsymbol{\beta}$ with a normal prior $\boldsymbol{\beta} \sim N(\mathbf{0}, \underline{\mathbf{\Sigma}})$ converges to the "population forecast" $\mathbf{f}_t'\boldsymbol{\gamma}$ if the variance of $\boldsymbol{\xi}_t$ is small relative to the contribution of the factors to the variance of $\mathbf{x}_t$ and the prior variance for $\boldsymbol{\beta}$ is chosen such that $|\underline{\mathbf{\Sigma}}| = O\left(\frac{1}{nT^{1/2+\delta}}\right)$, $0 < \delta < 1/2$. That is, the degree of shrinkage should increase with both n and T in order to protect against overfitting.

Korobilis (forthcoming*a*) use the same type of univariate direct forecasting model as De Mol et al. (2008) to forecast 129 US marcroeconomic variables using the other 128 variables as explanatory variables. The forecasts are made using 5 different hierarchical

[26]It should be made clear that no actual matrix inverse of this size is needed. The conditional posterior for $\boldsymbol{\gamma}$ has the form $\boldsymbol{\gamma}|\mathbf{Y}_T, \mathbf{\Psi} \sim N\left(\overline{\boldsymbol{\gamma}}, \overline{\mathbf{\Sigma}}_\gamma\right)$ with $\overline{\boldsymbol{\gamma}} = \mathbf{A}^{-1}\mathbf{b}$ and $\overline{\mathbf{\Sigma}}_\gamma = \mathbf{A}^{-1}$. $\overline{\boldsymbol{\gamma}}$ can be calculated by Cholesky decomposing $\mathbf{A} = \mathbf{C}'\mathbf{C}$ and then use forward and back substitution to solve the triangular equations systems $\mathbf{C}'\mathbf{x} = \mathbf{b}$ and $\mathbf{C}\overline{\boldsymbol{\gamma}} = \mathbf{x}$ in turn. A draw from $\boldsymbol{\gamma}|\mathbf{Y}_T, \mathbf{\Psi}$ is obtained by generating a vector of standard normals, $\mathbf{z}$, and computing $\overline{\boldsymbol{\gamma}} + \mathbf{C}^{-1}\mathbf{z} = \overline{\boldsymbol{\gamma}} + \widetilde{\mathbf{z}}$ where $\widetilde{\mathbf{z}}$ is obtained by solving $\mathbf{C}\widetilde{\mathbf{z}} = \mathbf{z}$ by back substitution. This is much more numerically stable than straightforward inversion of the matrices and also faster.

shrinkage priors where the hierarchical structure is used to allow the degree shrinkage to be influenced by the data and specific to each explanatory variable. As a comparison forecasts are also made with a dynamic factor model using the first five principal components as factors. Priors designed to mimic the LASSO and the Elastic Net are found to perform best when the forecasts are compared using the mean absolute error while the dynamic factor model performs best if the mean squared error criterion is used.

9.1 Factor augmented VAR

Bernanke, Boivin and Eliasz (2005) proposed the factor augmented VAR (FAVAR) as a means of incorporating the information from a large number of variables in a VAR in a parsimonious way. There are two basic assumption, that the data admits a factor structure, $\mathbf{x}_t = \mathbf{\Lambda f}_t + \boldsymbol{\xi}_t$ where the information in the n auxiliary variables in $\mathbf{x}_t$ can be represented by the r factors in $\mathbf{f}_t$ with $r \ll n$ and that the variables of interest, $\mathbf{y}_t$, and the factors can be jointly modelled as a VAR

$$\widetilde{\mathbf{y}}_t' = \begin{pmatrix} \mathbf{f}_t \\ \mathbf{y}_t \end{pmatrix}' = \sum_{i=1}^{p} \widetilde{\mathbf{y}}_{t-i}' \mathbf{A}_i + \mathbf{u}_t'. \tag{106}$$

$\mathbf{y}_t$ and $\mathbf{x}_t$ and hence also $\mathbf{f}_t$ are assumed to be stationary and we will work with demeaned data so there is no constant term in the VAR. Bernanke et al. (2005) augments the factor structure by allowing the variables of interest to be directly related to the auxiliary variables

$$\mathbf{x}_t = \mathbf{\Lambda}^f \mathbf{f}_t + \mathbf{\Lambda}^y \mathbf{y}_t + \boldsymbol{\xi}_t \tag{107}$$

instead of only indirectly through the factors $\mathbf{f}_t$. Like any factor model (107) suffers from a fundamental lack of identification since any full rank rotation of the factors will leave the model unaffected, e.g. $\mathbf{\Lambda}^f \mathbf{f}_t = \left(\mathbf{\Lambda}^f \mathbf{P}^{-1}\right)(\mathbf{P}\mathbf{f}_t) = \mathbf{\Lambda}^{f*}\mathbf{f}_t^*$ for $\mathbf{P}$ full rank. Bernanke et al. (2005) show that the restrictions

$$\mathbf{\Lambda}^f = \begin{pmatrix} \mathbf{I}_r \\ \mathbf{\Lambda}_*^f \end{pmatrix}, \ \mathbf{\Lambda}^y = \begin{pmatrix} \mathbf{0}_{r\times m} \\ \mathbf{\Lambda}_*^y \end{pmatrix}$$

together with the exact factor model assumption that $\mathbf{\Xi} = V(\boldsymbol{\xi}_t)$ is diagonal is sufficient for identification. The restriction on $\mathbf{\Lambda}^f$ is just a normalization whereas the restriction on $\mathbf{\Lambda}^y$ is substantial and implies that the first r variables in $\mathbf{x}_t$ does not respond contemporaneously to $\mathbf{y}_t$.

The key to inference in the FAVAR model is to recognize that it is a state space model (see Appendix B) with (107) as the observation equation and (106) as the state equation. The Kalman filter requires that the state equation has the Markov property, i.e. that it is autoregressive of order 1, and we rewrite the state equation with an expanded state vector

$$\begin{aligned} \mathbf{s}_{t+1} &= \begin{pmatrix} \widetilde{\mathbf{y}}_{t+1} \\ \widetilde{\mathbf{y}}_t \\ \vdots \\ \widetilde{\mathbf{y}}_{t-p+2} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_1' & \mathbf{A}_2' & \cdots & \mathbf{A}_{p-1}' & \mathbf{A}_p' \\ \mathbf{I} & \mathbf{0} & & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & & \mathbf{I} & \mathbf{0} \\ \vdots & & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & & \mathbf{I} & \mathbf{0} \end{pmatrix} \mathbf{s}_t + \begin{pmatrix} \mathbf{u}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix} \\ &= \mathbf{T}\mathbf{s}_t + \boldsymbol{\eta}_t \end{aligned} \tag{108}$$

and include the observable $\mathbf{y}_t$ in the observation equation

$$\begin{aligned}\mathbf{w}_t = \begin{pmatrix} \mathbf{x}_t \\ \mathbf{y}_t \end{pmatrix} &= \begin{pmatrix} \begin{pmatrix} \mathbf{I}_r \\ \mathbf{\Lambda}_*^f \end{pmatrix} & \begin{pmatrix} \mathbf{0}_{r\times m} \\ \mathbf{\Lambda}_*^y \end{pmatrix} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_m & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix} \mathbf{s}_t + \begin{pmatrix} \boldsymbol{\xi}_t \\ \mathbf{0} \end{pmatrix} \\ &= \mathbf{Z}\mathbf{s}_t + \boldsymbol{\varepsilon}_t. \end{aligned} \tag{109}$$

We make the usual assumption that the innovations $\mathbf{u}_t$ and $\boldsymbol{\xi}_t$ are iid normal and independent of each other, $\mathbf{u}_t \sim N(\mathbf{0}, \mathbf{\Psi})$ and $\boldsymbol{\xi}_t \sim N(\mathbf{0}, \mathbf{\Xi})$ with $\mathbf{\Xi}$ diagonal.

Conditional on the parameters $\mathbf{T}$, $\mathbf{Z}$, $\mathbf{\Psi}$ and $\mathbf{\Xi}$ we can use the Kalman filter and the simulations smoother to draw the latent factors from the full conditional posterior (see Appendix B). Note that system matrices $\mathbf{T}$, $\mathbf{Z}$, $\mathbf{H} = V(\boldsymbol{\varepsilon}_t)$ and $\mathbf{Q} = V(\boldsymbol{\eta}_t)$ contains a large number of zeros and the computations can be speeded up by taking account of the structure of the matrices. Note that including $\mathbf{y}_t$ as left hand side variables in the observation equation carries $\mathbf{y}_t$ through to the state vector. That is, $\mathbf{s}_{t|t}$ contains $\mathbf{y}_t$ since it is known at time t and $\mathbf{s}_{t+1|t}$ contains the minimum mean squared error prediction, $E_t(\mathbf{y}_{t+1})$, of the unknown $\mathbf{y}_{t+1}$. The Kalman filter recursions need to be started up with a prior for the first state, $\mathbf{s}_1 \sim N(\mathbf{s}_{1|0}, \mathbf{P}_{1|0})$.

Having run the Kalman filter, the last state can be sampled from the full conditional

$$\mathbf{s}_T|\mathbf{y}^T, \mathbf{x}^T, \mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{\Xi}, \mathbf{\Psi} \sim N(\mathbf{s}_{T|T}, \mathbf{P}_{T|T}) \tag{110}$$

where $\mathbf{\Gamma}$ collects the autoregressive parameters, $\mathbf{\Gamma}' = (\mathbf{A}_1', \ldots, \mathbf{A}_p')$ and $\mathbf{\Lambda} = (\mathbf{\Lambda}^f, \mathbf{\Lambda}^y)$. Note that it suffices to draw the factor $\mathbf{f}_T$ since $\mathbf{y}_t$, $t = 1, \ldots, T$ is known. Reflecting this, the variance matrix $\mathbf{P}_{T|T}$ is also singular which would cause numerical problems trying to draw $\mathbf{y}_T$. The remaining factors, $\mathbf{f}_t$, $t = T-1, \ldots, 1$, can be drawn from the conditionals $\mathbf{f}_t|\mathbf{y}^T, \mathbf{x}^T, \mathbf{\Gamma}, \mathbf{\Xi}, \mathbf{\Psi}, \mathbf{f}_{t+1} \sim N(\mathbf{s}_{t|T}, \mathbf{P}_{t|T})$ using the simulation smoother. Algorithm B.3 in Appendix B can, however, not be used directly due to the presence of $\mathbf{y}_t$ and lags in the state vector. To implement the simulation smoother we need the conditional distributions

$$\mathbf{s}_t|\mathbf{y}^t, \mathbf{x}^t, \mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{\Xi}, \mathbf{\Psi}, \mathbf{f}_{t+1} \sim N(\mathbf{s}_{t|t,\mathbf{f}_{t+1}}, \mathbf{P}_{t|t,\mathbf{f}_{t+1}}). \tag{111}$$

Using that $\widetilde{\mathbf{y}}_{t+1|t} = \left(\mathbf{f}_{t+1|t}', \mathbf{y}_{t+1|t}'\right)' \sim N(\mathbf{\Gamma}'\mathbf{s}_{t|t}, \mathbf{\Gamma}'\mathbf{P}_{t|t}\mathbf{\Gamma} + \mathbf{\Psi})$ it is easy to see that the recursions for the parameters of the conditional distributions becomes (see Kim and Nelson (1999, p. 194-196))

$$\begin{aligned}\mathbf{s}_{t|t,\mathbf{f}_{t+1}} &= \mathbf{s}_{t|t} + \mathbf{P}_{t|t}\mathbf{\Gamma}'\left(\mathbf{\Gamma}'\mathbf{P}_{t|t}\mathbf{\Gamma} + \mathbf{\Psi}\right)^{-1}\left(\widetilde{\mathbf{y}}_{t+1} - \mathbf{\Gamma}'\mathbf{s}_{t|t}\right) \\ \mathbf{P}_{t|t,\mathbf{f}_{t+1}} &= \mathbf{P}_{t|t} - \mathbf{P}_{t|t}\mathbf{\Gamma}\left(\mathbf{\Gamma}'\mathbf{P}_{t|t}\mathbf{\Gamma} + \mathbf{\Psi}\right)^{-1}\mathbf{\Gamma}'\mathbf{P}_{t|t}\end{aligned} \tag{112}$$

for $t = T-1, \ldots, 1$. It is again sufficient to only draw the factor $\mathbf{f}_t$ at each iteration.

Inference on the remaining parameters is standard conditional on the factors. The state equation (106) is a standard VAR and draws from the full conditional posterior for $\mathbf{\Psi}$ and $\mathbf{\Gamma}$ can be obtained using the results in section 3.2.1 with a normal-Wishart prior and section 3.2.2 with an independent normal Wishart or normal-diffuse prior.

The observation equation (107) can be analyzed as n univariate regressions when $\mathbf{\Xi}$ is diagonal. The identifying restrictions imply that we have

$$\begin{aligned}x_{it} &= f_{it} + \xi_{it}, \ i = 1, \ldots, r \\ x_{it} &= \widetilde{\mathbf{y}}_t'\boldsymbol{\lambda}_i + \xi_{it}, \ i = r+1, \ldots, n\end{aligned}$$

where $\boldsymbol{\lambda}_i'$ is row i of $\boldsymbol{\Lambda} = \left(\boldsymbol{\Lambda}^f, \boldsymbol{\Lambda}^y\right)$. Let σ_i^2 be the variance ξ_{it}, the conjugate prior is of the normal-Gamma form,

$$\begin{aligned}\sigma_i^2 &\sim iG\left(\underline{a}_i, \underline{b}_i\right)\\ \boldsymbol{\lambda}_i|\sigma_i^2 &\sim N\left(\underline{\boldsymbol{\lambda}}_i, \sigma_i^2\underline{\mathbf{V}}_i\right)\end{aligned}$$

and the full conditional posteriors are given by

$$\begin{aligned}\boldsymbol{\lambda}_i|\mathbf{y}^T, \mathbf{x}^T, \mathbf{f}^T, \sigma_i^2 &\sim N\left(\overline{\boldsymbol{\lambda}}_i, \sigma_i^2\overline{\mathbf{V}}_i\right),\ i = r+1, \ldots, n \qquad (113)\\ \overline{\mathbf{V}}_i &= \left(\underline{\mathbf{V}}_i^{-1} + \widetilde{\mathbf{Y}}'\widetilde{\mathbf{Y}}\right)^{-1}\\ \overline{\boldsymbol{\lambda}}_i &= \overline{\mathbf{V}}_i\left(\underline{\mathbf{V}}_i^{-1}\underline{\boldsymbol{\lambda}}_i + \widetilde{\mathbf{Y}}'\widetilde{\mathbf{Y}}\widehat{\boldsymbol{\lambda}}_i\right)^{-1}\end{aligned}$$

and

$$\begin{aligned}\sigma_i^2|\mathbf{y}^T, \mathbf{x}^T, \mathbf{f}^T &\sim iG\left(\overline{a}_i, \overline{b}_i\right) \qquad (114)\\ \overline{a}_i &= \begin{cases} \underline{a}_i + T/2,\ i = 1, \ldots, r \\ \underline{a}_i + (T - r - m)/2,\ i = r+1, \ldots, n \end{cases}\\ \overline{b}_i &= \begin{cases} \underline{b}_i + \frac{1}{2}\sum_{i=1}^T (x_{it} - f_{it})^2,\ i = 1, \ldots, r \\ \underline{b}_i + \frac{1}{2}\left(\mathbf{x}_i'\mathbf{x}_i + \underline{\boldsymbol{\lambda}}_i'\underline{\mathbf{V}}_i^{-1}\underline{\boldsymbol{\lambda}}_i - \overline{\boldsymbol{\lambda}}_i'\overline{\mathbf{V}}_i\overline{\boldsymbol{\lambda}}_i\right),\ i = r+1, \ldots, n \end{cases}\end{aligned}$$

where $\widetilde{\mathbf{Y}}$ is the matrix of explanatory variables $\widetilde{\mathbf{y}}_t'$ and $\widehat{\boldsymbol{\lambda}}_i$ is the OLS estimate $\widehat{\boldsymbol{\lambda}}_i = \left(\widetilde{\mathbf{Y}}'\widetilde{\mathbf{Y}}\right)^{-1}\widetilde{\mathbf{Y}}'\mathbf{x}_i$.

Specifying the prior It is difficult to have information a priori about the latent factors and the prior for the first state, $\mathbf{s}_1 \sim N\left(\mathbf{s}_{1|0}, \mathbf{P}_{1|0}\right)$, is best taken to be non-informative, for example $\mathbf{s}_{1|0} = \mathbf{0}$ and $\mathbf{P}_{1|0} = 5\mathbf{I}$. It is also difficult to form prior opinions about the factor loadings in $\boldsymbol{\Lambda}^f$ and $\boldsymbol{\Lambda}^y$ and non-informative priors are advisable, Bernanke et al. (2005) sets $\underline{\boldsymbol{\lambda}}_i = \mathbf{0}$, $\underline{\mathbf{V}}_i = \mathbf{I}$, $\underline{a}_i = 0.001$ and $\underline{b}_i = 3$. The prior for $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$ can be based on the same considerations as a standard VAR while taking account of the stationarity assumption.

Sampling from the posterior A Gibbs sampler for the joint posterior distribution of the factors and the parameters can be constructed running the simulation smoother for the factors and sample the parameters from the full conditional posteriors, see Algorithm 15.

Forecasting performance See discussion of Gupta and Kabundi (2010) in section 9.2 and discussion of Korobilis (2008) in section 8.1

9.2 Large BVARs

9.2.1 Reducing parameter uncertainty by shrinkage

Banbura et al. (2010) studies the forecast performance of large BVARs. In an application to forecasting US non-farm employment, CPI and the Federal Funds Rate the performance

Algorithm 15 Gibbs sampler for the FAVAR model

For the FAVAR (106, 107) select starting values $\mathbf{\Gamma}^{(0)}$, $\mathbf{\Psi}^{(0)}$, $\mathbf{\Lambda}^{(0)}$ and $\mathbf{\Xi}^{(0)}$.
For $j = 1, \ldots, B + R$

1. Draw the factor $\mathbf{f}_T^{(j)}$ from the full conditional posterior $\mathbf{s}_T|\mathbf{y}^T, \mathbf{x}^T, \mathbf{\Gamma}^{(j-1)}, \mathbf{\Lambda}^{(j-1)}, \mathbf{\Xi}^{(j-1)}, \mathbf{\Psi}^{(j-1)}$ in (110) obtained by running the Kalman filter (126) in Appendix B. For $t = T-1, \ldots, 1$ draw $\mathbf{f}_t^{(j)}$ from the full condition posterior $\mathbf{s}_t|\mathbf{y}^t, \mathbf{x}^t, \mathbf{\Gamma}^{(j-1)}, \mathbf{\Lambda}^{(j-1)}, \mathbf{\Xi}^{(j-1)}, \mathbf{\Psi}^{(j-1)}, \mathbf{f}_{t+1}^{(j)}$ in (111) obtained by running the simulation smoother (112).

2. Draw $\mathbf{\Psi}^{(j)}$ and $\mathbf{\Gamma}^{(j)}$ from the conditional posteriors $\mathbf{\Psi}|\mathbf{y}^T, \mathbf{x}^T, \mathbf{\Lambda}^{(j-1)}, \mathbf{\Xi}^{(j-1)}, \mathbf{f}^{(j)}$ in (19) and $\mathbf{\Gamma}|\mathbf{y}^T, \mathbf{x}^T, \mathbf{\Lambda}^{(j-1)}, \mathbf{\Xi}^{(j-1)}, \mathbf{\Psi}^{(j)}, \mathbf{f}^{(j)}$ in (18) with a normal-Wishart prior or $\mathbf{\Psi}|\mathbf{y}^T, \mathbf{x}^T, \mathbf{\Gamma}^{(j-1)}, \mathbf{\Lambda}^{(j-1)}, \mathbf{\Xi}^{(j-1)}, \mathbf{f}^{(j)}$ in (25) and $\mathbf{\Gamma}|\mathbf{y}^T, \mathbf{x}^T, \mathbf{\Lambda}^{(j-1)}, \mathbf{\Xi}^{(j-1)}, \mathbf{\Psi}^{(j)}, \mathbf{f}^{(j)}$ in (24) with an independent normal Wishart prior.

3. For $i = 1, \ldots, n$ draw $\sigma_i^{2(j)}$ from the full conditional posterior $\sigma_i^2|\mathbf{y}^T, \mathbf{x}^T, \mathbf{f}^{(j)}$ in (114) and (for $i > r$) $\boldsymbol{\lambda}_i^{(j)}$ from the full conditional posterior $\boldsymbol{\lambda}_i|\mathbf{y}^T, \mathbf{x}^T, \mathbf{f}^{(j)}, \sigma_i^{2(j)}$ in (113).

4. If $j > B$ generate $\mathbf{u}_{T+1}^{(j)}, \ldots, \mathbf{u}_{T+H}^{(j)}$ from $\mathbf{u}_t \sim N\left(0, \mathbf{\Psi}^{(j)}\right)$ and calculate recursively

$$\widetilde{\mathbf{y}}_{T+h}^{*(j)\prime} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}_{T+h-i}^{*(j)\prime} \mathbf{A}_i^{(j)} + \sum_{i=h}^{p} \widetilde{\mathbf{y}}_{T+h-i}^{\prime} \mathbf{A}_i^{(j)} + \mathbf{x}_{T+h}^{\prime} \mathbf{C}^{(j)} + \mathbf{u}_{T+h}^{(j)\prime}.$$

   Discarding the parameters yields $\left\{\widetilde{\mathbf{y}}_{T+1}^{*(j)}, \ldots \widetilde{\mathbf{y}}_{T+H}^{*(j)}\right\}_{j=B+1}^{B+R}$ as a sample from the joint predictive distribution of the factors and the variables of interest $\mathbf{y}$.

---

of 4 different VAR models with 3, 7, 20 and 131 variables, all with 13 lags, is evaluated. Based on the theoretical results of De Mol et al. (2008), Banbura et al. (2010) suggest that the degree of shrinkage applied through the prior should increase with the size of the model. Working with a normal-Wishart prior distribution with Minnesota type prior beliefs, the overall scaling factor π_1 in (21) determines the amount of shrinkage. Banbura et al. sets this to match the in sample fit of the smallest VAR estimated with OLS (i.e. $\pi_1 = \infty$) for the three variables of interest where the fit of model $\mathcal{M}$ with prior scaling π_1 is measured by

$$Fit\left(\pi_1, \mathcal{M}\right) = \frac{1}{3} \sum_{i \in \mathcal{I}} \frac{mse_i^{(\pi_1, \mathcal{M})}}{mse_i^{(0)}}, \tag{115}$$

the average of the in-sample mean square error normalized by the MSE of a pure random walk model. In this particular application this leads to scaling factors ∞, 0.52, 0.33 and 0.19.[27] That is, the small 3-variable VAR is estimated with OLS and the scale factor for the 7 variable VAR is 0.52. The main finding is that, with the increased shrinkage, forecast performance improves with model size but also that most of the gains was achieved with

[27]These numbers differ from the ones reported by Banbura et al. (2010) since they parameterize the prior in terms of the square of π_1.

the 20 variable VAR. A moderately large VAR might thus be sufficient provided that the right variables are selected.

Bloor and Matheson (2010) working in an open economy context with a need to make foreign variables exogenous to domestic variables generalized the approach of Banbura et al. (2010) by considering different amounts of shrinkage for different blocks of equations. This implies that the convenient normal-Wishart prior can not be used and Bloor and Matheson (2010) base their inference on the blocked importance sampler of Zha (1999) embodying the same kind of Minnesota type prior beliefs with the addition of a dummy initial observations prior. To impose different amount of shrinkage the prior hyperparameters are made block specific and chosen using the same strategy as Bloor and Matheson (2010). Bloor and Matheson (2010) forecasts the New Zealand real GDP, tradable CPI, non-tradable CPI, an interest rate and the exchange rate using a range of BVARs with 5, 8, 14 and 94 variables. In addition, forecast results are reported for a univariate AR, a random walk model and frequentist variants of the smallest BVAR. Overall the largest BVAR provides the best forecasts except for the long horizon (4 quarters) forecast of tradable CPI where two 5 variable VARs, the BVAR and a frequentist with lags selected using BIC, gives significantly MSEs.

Gupta and Kabundi (2010) conducts a forecasting exercise where the aim is to forecast the South African per capita growth rate, inflation, the money market rate and the growth rate of the nominal effective exchange rate. The models used are a small (four variable) DSGE model, a dynamic factor model, a FAVAR using the four variables of interest estimated by OLS and a Bayesian variant using a Minnesota type prior, a four variable unrestricted VAR, two BVARs with 4 and 266 variables. The factor models use principal components as factors. For the VAR models the lag length is set to five with quarterly data and the overall scale factor π_1 in the Minnesota prior for the BVARs is set following the approach of Banbura et al. (2010) in addition to common default settings. In addition Gupta and Kabundi (2010) also experiment with the lag decay rate π_3 using settings of 0.5, 1 and 2. For the small VAR the additional shrinkage on lags of other variables is set to $\pi_2 = 0.5$. In the large VAR a tighter specification is used with $\pi_2 = 0.6$ for foreign (world) variables and for domestic variables 0.1 is used in domestic equations and 0.01 in world equations. Overall the large BVAR with the tightest shrinkage, $\pi_1 = 0.01$ and $\pi_3 = 1$, does well and delivers the best forecast for three out of the four variables. The exception being the exchange rate where the DSGE model does best.

9.2.2 Selecting variables - conjugate SSVS

Koop (2010) considers a range of prior specifications for forecasting with large BVARs. This includes the normal-Wishart with Minnesota type prior beliefs used by De Mol et al. (2008), the original Litterman prior with fixed and diagonal error variance matrix (section 3.1) which offers the advantage that different shrinkage can be applied to lags of the dependent variable and lags on other variables, the same Minnesota prior but also allowing for correlation between the variables of interest, the SSVS prior (section 8.1) with two different settings for the prior variances and a new "conjugate" SSVS prior that is less computationally demanding and better suited for large BVARs.

The new SSVS prior takes $\mathbf{\Gamma}$ to be distributed as a matricvariate normal conditionally

on $\mathbf{\Psi}$ and the vector of selection indicators $\boldsymbol{\delta}$,

$$\mathbf{\Gamma}|\mathbf{\Psi},\boldsymbol{\delta} \sim MN_{km}\left(\underline{\mathbf{\Gamma}},\mathbf{\Psi},\underline{\mathbf{\Omega}}_{\delta}\right) \tag{116}$$

with $\underline{\mathbf{\Omega}}_{\delta} = \text{diag}\left(h_1,\ldots,h_k\right)$ for

$$h_i = \begin{cases} \tau_{0,i}^2 \text{ if } \delta_i = 0 \\ \tau_{1,i}^2 \text{ if } \delta_i = 1 \end{cases}.$$

The prior for δ_i is independent Bernoulli distributions with $P\left(\delta_i = 1\right) = p_i$ and for $\mathbf{\Psi}$ an inverse Wishart, $\mathbf{\Psi} \sim iW\left(\underline{\mathbf{S}},\underline{v}\right)$ is used. The prior structure is thus conjugate conditional on $\boldsymbol{\delta}$.

In contrast with the standard SSVS procedure, the conjugate SSVS includes or excludes a variable in all equations at the same time instead of being specific to one variable and equation. While this reduces the flexibility, the Kronecker structure of the prior and posterior variance matrices for $\mathbf{\Gamma}$ makes for much more efficient computations.

The conditional posterior distributions in (18) and (19) still holds but should be interpreted as conditional on $\boldsymbol{\delta}$. The marginal posterior for $\boldsymbol{\delta}$ is obtained up to a proportionality constant by integrating out $\mathbf{\Gamma}$ and $\mathbf{\Psi}$ from the product of the likelihood and the prior yielding the matricvariate-t distribution (20) times the prior for $\boldsymbol{\delta}, \pi\left(\boldsymbol{\delta}\right)$. After some simplifications of the expression for the matricvariate-t density we have

$$p\left(\boldsymbol{\delta}|\mathbf{Y}_T\right) \propto g\left(\boldsymbol{\delta},\mathbf{Y}_T\right) = \left(\left|\underline{\mathbf{\Omega}}_{\delta}\right| / \left|\overline{\mathbf{\Omega}}_{\delta}\right|\right)^{-m/2} \left|\overline{\mathbf{S}}\right|^{-(\underline{v}+T)/2} \pi\left(\boldsymbol{\delta}\right). \tag{117}$$

For k small it is possible to enumerate $p\left(\boldsymbol{\delta}|\mathbf{Y}_T\right)$ but since there are 2^k possible configurations for $\boldsymbol{\delta}$ this quickly becomes infeasible and Koop (2010) suggests a Gibbs sampling approach for sampling from the marginal posterior of $\boldsymbol{\delta}$ originally proposed by Brown, Vanucci and Fearn (1998). This is reproduced as part A of Algorithm 16.

Forecasting performance Koop evaluates the forecasting performance using a data set with 168 US macroeconomic variables and four different VAR models with 3, 20, 40 and 168 variables with four lags and formulated to generate direct forecasts. The variables of interest are real GDP, the CPI and the Federal Funds Rate. For the SSVS priors two different ways of setting the prior variances are used. The "semiautomatic" approach of George et al. (2008) sets $\tau_{0,ij} = \widehat{\sigma}_{\gamma_{ij}}/10$ and $\tau_{1,ij} = 10\widehat{\sigma}_{\gamma_{ij}}$ where $\widehat{\sigma}_{\gamma_{ij}}$ is the standard error of the OLS estimate of γ_{ij} for the standard SSVS. For the conjugate SSVS the maximum of $\widehat{\sigma}_{\gamma_{ij}}$ for a given i is used. The other approach is based on the Minnesota prior and sets $\tau_{1,ij}$ according to (14) and $\tau_{0,ij} = \tau_{1,ij}/10$ for the standard SSVS and mimics the normal-Wishart prior variance for the conjugate SSVS. For priors with Minnesota type prior beliefs the scale factors π_1 and π_2 are set in the same way as in Banbura et al. (2010) using (115).

As a complement forecasts are also calculated using for a number of FAVARs constructed by adding lags of the principal components to the three-variable VAR.

The results of the forecasting exercise is mixed and there is no clear winner but some patterns do emerge. The factor models does not do very well and never performs best. There is a gain from moving to larger models but as in Banbura et al. (2010) the additional gains are small once one moves beyond 20 variables in the VAR. The Minnesota and SSVS type priors have almost the same number of wins and there is no indication that one is better than the other.

Algorithm 16 Gibbs sample for "conjugate" SSVS

With the conditional prior (116) for $\mathbf{\Gamma}$, inverse Wishart prior for $\mathbf{\Psi}$ and independent Bernoulli priors on δ_i part A below samples from the marginal posterior for $\boldsymbol{\delta}$. After convergence this can be complemented with part B to produce draws from the joint posterior for $\boldsymbol{\delta}$, $\mathbf{\Gamma}$ and $\mathbf{\Psi}$.

A. Select starting values $\boldsymbol{\delta}^{(0)}$

   For $j = 1, \ldots, B + R$

   1. Draw $\delta_i^{(j)}$, $i = 1, \ldots, k$, from the full conditional $\delta_i^{(j)} | \mathbf{Y}_T, \delta_1^{(j)}, \ldots, \delta_{i-1}^{(j)}, \delta_{i+1}^{(j-1)}, \ldots, \delta_k^{(j-1)} \sim Ber\left(u_{1i} / \left(u_{0i} + u_{1i}\right)\right)$ where $u_{0i} = g\left(\boldsymbol{\delta}_{-i}, \delta_i = 0, \mathbf{Y}_T\right)$, $u_{1i} = g\left(\boldsymbol{\delta}_{-i}, \delta_i = 1, \mathbf{Y}_T\right)$ and $g(\cdot)$ is given by (117).

B. If $j > B$ and a sample from the joint posterior for $\boldsymbol{\delta}$, $\mathbf{\Gamma}$ and $\mathbf{\Psi}$ or the predictive distribution is desired

   2. Draw $\mathbf{\Psi}^{(j)}$ from the full conditional posterior $\mathbf{\Psi} | \mathbf{Y}_T, \boldsymbol{\delta}^{(j)} \sim iW\left(\overline{\mathbf{S}}, \overline{v}\right)$ in (19).
   3. Draw $\mathbf{\Gamma}^{(j)}$ from the full conditional posterior $\mathbf{\Gamma} | \mathbf{Y}_T, \mathbf{\Psi}^{(j)}, \boldsymbol{\delta}^{(j)} \sim MN_{km}\left(\overline{\mathbf{\Gamma}}, \mathbf{\Psi}, \overline{\mathbf{\Omega}}_\delta\right)$ in (18).
   4. Generate $\mathbf{u}_{T+1}^{(j)}, \ldots, \mathbf{u}_{T+H}^{(j)}$ from $\mathbf{u}_t \sim N\left(0, \mathbf{\Psi}^{(j)}\right)$ and calculate recursively

$$\widetilde{\mathbf{y}}_{T+h}^{(j)\prime} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}_{T+h-i}^{(j)\prime} \mathbf{A}_i^{(j)} + \sum_{i=h}^{p} \mathbf{y}_{T+h-i}^{\prime} \mathbf{A}_i^{(j)} + \mathbf{x}_{T+h}^{\prime} \mathbf{C}^{(j)} + \mathbf{u}_{T+h}^{(j)\prime}.$$

   Discarding the parameters yields $\left\{\widetilde{\mathbf{y}}_{T+1}^{(j)}, \ldots \widetilde{\mathbf{y}}_{T+H}^{(j)}\right\}_{j=B+1}^{B+R}$ as a sample from the joint predictive distribution.

9.3 Reduced rank VAR

Carriero, Kapetanios and Marcellino (2011) use standard VAR models (6) with iterated forecasts for lead times $h > 1$ and propose different ways of overcoming the "curse of dimensionality". In addition to the standard VAR-model with a tight normal-Wishart prior they propose the use of models with reduced rank parameter matrices. Working with data transformed to stationarity and then standardized a VAR without deterministic terms, $\mathbf{y}_t^{\prime} = \sum_{i=1}^{p} \mathbf{y}_{t-i}^{\prime} \mathbf{A}_i + \mathbf{u}_t^{\prime}$, is used and the reduced rank assumption is that the parameter matrices can be written as $\mathbf{A}_i = \boldsymbol{\beta}_i \boldsymbol{\alpha}^{\prime}$ where $\boldsymbol{\beta}_i$ and $\boldsymbol{\alpha}$ are $m \times r$ matrices. In matrix form we can write the VAR as

$$\mathbf{Y} = \mathbf{Z}\mathbf{\Gamma} + \mathbf{u} = \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\alpha}^{\prime} + \mathbf{u} \tag{118}$$

for $\mathbf{\Gamma}^{\prime} = \left(\mathbf{A}_1^{\prime}, \ldots, \mathbf{A}_p^{\prime}\right)$ and $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^{\prime}, \ldots, \boldsymbol{\beta}_p^{\prime}\right)^{\prime}$ a $mp \times r$ matrix with $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ of rank $r < m$. The setup is similar to the cointegrated VECM (44) and the same basic issue of the lack of identification of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ arises here. Following Geweke (1996*a*) Carriero,

Kapetanios and Marcellino use a linear normalization, $\boldsymbol{\alpha}' = (\mathbf{I}_r, \boldsymbol{\alpha}_*')$ and specify a prior of the form

$$\text{vec}(\boldsymbol{\alpha}_*) \sim N(\text{vec}(\underline{\boldsymbol{\alpha}}_*), \underline{\boldsymbol{\Sigma}}_\alpha), \ \boldsymbol{\beta} \sim N(\underline{\boldsymbol{\beta}}, \underline{\boldsymbol{\Sigma}}_\beta), \ \boldsymbol{\Psi} \sim iW(\underline{\mathbf{S}}, \underline{v}). \tag{119}$$

Again, following Geweke, $\underline{\boldsymbol{\alpha}}_*$ and $\underline{\boldsymbol{\beta}}$ are set to zero and $\underline{\boldsymbol{\Sigma}}_\alpha$ and $\underline{\boldsymbol{\Sigma}}_\beta$ are diagonal matrices with diagonal elements $1/\tau^2$ in the application.

The derivation of the full conditional posteriors parallels the one for the VECM in section 5.1 with obvious changes due to the different normalization. The full conditional posterior for $\boldsymbol{\Psi}$ is inverse Wishart,

$$\boldsymbol{\Psi}|\mathbf{Y}_T, \boldsymbol{\beta}, \boldsymbol{\alpha} \sim iW(\overline{\mathbf{S}}, \overline{v}), \ \overline{\mathbf{S}} = \underline{\mathbf{S}} + (\mathbf{Y} - \mathbf{Z}\boldsymbol{\Gamma})'(\mathbf{Y} - \mathbf{Z}\boldsymbol{\Gamma}), \ \overline{v} = \underline{v} + T. \tag{120}$$

For the full conditional posterior for $\boldsymbol{\alpha}_*$ rewrite the model as one set of equations for the r first variables which does not depend on $\boldsymbol{\alpha}_*$

$$\mathbf{Y}_1 = \mathbf{Z}\boldsymbol{\beta} + \mathbf{u}_1$$

and a set of equations for the remaining $m - r$ variables depending on $\boldsymbol{\alpha}_*$

$$\mathbf{Y}_2 = \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\alpha}_*' + \mathbf{u}_2.$$

Using that $\mathbf{u}_2|\mathbf{u}_1 \sim MN_{T,m-r}\left(\mathbf{u}_1\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Psi}_{12}, (\boldsymbol{\Psi}^{22})^{-1}, \mathbf{I}_T\right)$ for

$$\boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Psi}_{11} & \boldsymbol{\Psi}_{12} \\ \boldsymbol{\Psi}_{21} & \boldsymbol{\Psi}_{22} \end{pmatrix}$$

and $\boldsymbol{\Psi}^{22} = \left(\boldsymbol{\Psi}_{22} - \boldsymbol{\Psi}_{21}\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Psi}_{12}\right)^{-1}$ the lower right $(m-r)\times(m-r)$ block of $\boldsymbol{\Psi}^{-1}$ we have

$$\mathbf{Y}_2|\mathbf{Y}_1, \boldsymbol{\beta}, \boldsymbol{\alpha}_* \sim MN_{T,m-r}\left(\left[\mathbf{Y}_1 - \mathbf{Z}\boldsymbol{\beta}\right]\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Psi}_{12} + \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\alpha}_*', \left(\boldsymbol{\Psi}^{22}\right)^{-1}, \mathbf{I}_T\right)$$

or a conditional regression[28]

$$\mathbf{Y}_c = \mathbf{Y}_2 - \left[\mathbf{Y}_1 - \mathbf{Z}\boldsymbol{\beta}\right]\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Psi}_{12} = \mathbf{Z}\boldsymbol{\beta}\boldsymbol{\alpha}_*' + \mathbf{u}_2.$$

It follows that the full conditional posterior for $\boldsymbol{\alpha}_*$ is normal[29]

$$\begin{aligned} \text{vec}(\boldsymbol{\alpha}_*)|\mathbf{Y}_T, \boldsymbol{\beta}, \boldsymbol{\Psi} &\sim N\left(\text{vec}(\overline{\boldsymbol{\alpha}}_*), \overline{\boldsymbol{\Sigma}}_\alpha\right) \\ \overline{\boldsymbol{\Sigma}}_\alpha &= \left(\underline{\boldsymbol{\Sigma}}_\alpha^{-1} + \boldsymbol{\Psi}^{22} \otimes \boldsymbol{\beta}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\beta}\right)^{-1} \\ \text{vec}(\overline{\boldsymbol{\alpha}}_*) &= \overline{\boldsymbol{\Sigma}}_\alpha\left\{\underline{\boldsymbol{\Sigma}}_\alpha^{-1}\text{vec}(\underline{\boldsymbol{\alpha}}_*) + \text{vec}\left(\boldsymbol{\beta}'\mathbf{Z}'\mathbf{Y}_c\boldsymbol{\Psi}^{22}\right)\right\}. \end{aligned} \tag{121}$$

The full conditional posterior for the unrestricted $\boldsymbol{\beta}$ matrix follows immediately after vectorizing the model,

$$\mathbf{y} = (\boldsymbol{\alpha} \otimes \mathbf{Z})\text{vec}(\boldsymbol{\beta}) + \mathbf{u},$$

[28] That is, we factor the likelihood into a marginal distribution for $\mathbf{Y}_1$ which is functionally independent of $\boldsymbol{\alpha}_*$ and a conditional distribution for $\mathbf{Y}_2$ that depends on $\boldsymbol{\alpha}_*$. It is then sufficient to consider the conditional likelihood.

[29] Expressions (14) - (16) in Carriero, Kapetanios and Marcellino (2011), which in turn are based on results in Geweke (1996*a*), are incorrect. See Karlsson (2012) for details.

as a normal distribution,

$$\begin{aligned} \text{vec}(\boldsymbol{\beta}) \,|\mathbf{Y}_T, \boldsymbol{\alpha}, \boldsymbol{\Psi} &\sim N\left(\text{vec}\left(\overline{\boldsymbol{\beta}}\right), \overline{\boldsymbol{\Sigma}}_\beta\right) \\ \overline{\boldsymbol{\Sigma}}_\beta &= \left(\underline{\boldsymbol{\Sigma}}_\beta^{-1} + \boldsymbol{\alpha}'\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha} \otimes \mathbf{Z}'\mathbf{Z}\right)^{-1} \\ \text{vec}\left(\overline{\boldsymbol{\beta}}\right) &= \overline{\boldsymbol{\Sigma}}_\beta \left(\underline{\boldsymbol{\Sigma}}_\beta^{-1} \text{vec}\left(\underline{\boldsymbol{\beta}}\right) + \text{vec}\left(\mathbf{Z}'\mathbf{Y}\boldsymbol{\Psi}^{-1}\boldsymbol{\alpha}\right)\right). \end{aligned} \tag{122}$$

It is thus straightforward to implement a Gibbs sampler for the reduced rank VAR model. Due to the relatively large variance matrix $\overline{\boldsymbol{\Sigma}}_\beta$ the Gibbs sampler can be time consuming and Carriero, Kapetanios and Marcellino (2011) suggests a computationally convenient alternative which they label *reduced rank posterior*. This is based on a reduced rank approximation to the posterior mean of $\boldsymbol{\Gamma}$, $\overline{\boldsymbol{\Gamma}}$, with a tight normal-Wishart prior. Let $\overline{\boldsymbol{\Gamma}} = \mathbf{UDV}'$ be the singular value decomposition of $\overline{\boldsymbol{\Gamma}}$ and collect the r largest singular values and corresponding vectors in the matrices $\mathbf{D}^* = \text{diag}(d_1, d_2, \ldots, d_r)$, $\mathbf{U}^* = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r)$ and $\mathbf{V}^* = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r)$. A rank $r < m$ approximation to $\overline{\boldsymbol{\Gamma}}$ is then given by $\overline{\boldsymbol{\Gamma}}^* = \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*\prime}$.

Forecast performance In a forecasting exercise with 52 macroeconomic variables for the US Carriero, Kapetanios and Marcellino (2011) compare the performance of the Bayesian procedures, VAR with normal-Wishart prior, the reduced rank VAR and the reduced rank posterior, with several alternatives, a reduced rank VAR estimated with maximum likelihood, multivariate boosting, factor models and univariate autoregressions. The reduced rank posterior and the Bayesian reduced rank VAR procedures are found to give the best forecasts, both in terms of forecasting all the 52 variables and when specific variables of interest (industrial production, inflation and the federal funds rate) are singled out.

9.4 Predicting many variables

Carriero, Kapetanios and Marcellino (2009, 2012) takes a slightly different viewpoint and considers the situation where a large number of variables are to be predicted rather than a small set of variables of interest. In an application to forecasting exchange rates Carriero, Kapetanios and Marcellino (2009) use a direct forecast version of a one-lag VAR

$$\mathbf{y}_t' = \mathbf{y}_{t-h}'\boldsymbol{\Phi}_h + \boldsymbol{\phi}_h + \mathbf{e}_{t,h}' \tag{123}$$

with $\mathbf{y}_t$ a $m = 32$ dimensional vector of log exchange rates. Taking $\mathbf{e}_{t,h}$ to be normal, $\mathbf{e}_{t,h} \sim N(\mathbf{0}, \boldsymbol{\Psi}_h)$ they specify a normal-Wishart prior (section 3.2.1) for $\boldsymbol{\Gamma}_h' = (\boldsymbol{\Phi}_h', \boldsymbol{\phi}_h')$ and $\boldsymbol{\Psi}_h$ centered on driftless univariate random walks. To avoid overfitting the prior is very tight with π_1 on the order of 0.01, about 1/10 of the conventional setting for medium sized VARs, and allowed to vary over time and chosen to minimize the sum of the mean square forecast errors for the previous period. In a comparison of the forecast performance with naive random walk forecasts, univariate autoregressions, forecasts from a standard VAR estimated with OLS and factor models with 4 factors, the BVAR is found to perform best with the random walk second.

Carriero, Kapetanios and Marcellino (2012) propose to use the same direct VAR (123) to forecast the term structure of interest rates. In contrast to Carriero et al. (2009) the scaling factor π_1 is chosen in an empirical Bayes fashion by maximizing the marginal

Algorithm 17 Gibbs sampler for the reduced rank VAR model

For the reduced rank VAR (118) and the prior (119) select starting values $\boldsymbol{\alpha}_*^{(0)}$ and $\boldsymbol{\Psi}^{(0)}$. For $j = 1, \ldots, B + R$

1. Generate $\boldsymbol{\beta}^{(j)}$ from the full conditional posterior $\text{vec}(\boldsymbol{\beta}) | \mathbf{Y}_T, \boldsymbol{\alpha}^{(j-1)}, \boldsymbol{\Psi}^{(j-1)} \sim N\left(\text{vec}\left(\overline{\boldsymbol{\beta}}\right), \overline{\boldsymbol{\Sigma}}_\beta\right)$ in (122).

2. Generate $\boldsymbol{\alpha}_*^{(j)}$ from the full conditional posterior $\text{vec}(\boldsymbol{\alpha}_*) | \mathbf{Y}_T, \boldsymbol{\beta}^{(j)}, \boldsymbol{\Psi}^{(j-1)} \sim N\left(\text{vec}\left(\overline{\boldsymbol{\alpha}}_*\right), \overline{\boldsymbol{\Sigma}}_\alpha\right)$ in (121).

3. Generate $\boldsymbol{\Psi}^{(j)}$ from the full conditional posterior $\boldsymbol{\Psi} | \mathbf{Y}_T, \boldsymbol{\beta}^{(j)}, \boldsymbol{\alpha}^{(j)} \sim iW\left(\overline{\mathbf{S}}, \overline{v}\right)$ in (120).

4. If $j > B$ form $\boldsymbol{\Gamma}^{(j)} = \boldsymbol{\beta}^{(j)} \boldsymbol{\alpha}^{(j)'}$, generate $\mathbf{u}_{T+1}^{(j)}, \ldots, \mathbf{u}_{T+H}^{(j)}$ from $\mathbf{u}_t \sim N\left(0, \boldsymbol{\Psi}^{(j)}\right)$ and calculate recursively

$$\widetilde{\mathbf{y}}_{T+h}^{(j)\prime} = \sum_{i=1}^{h-1} \widetilde{\mathbf{y}}_{T+h-i}^{(j)\prime} \mathbf{A}_i^{(j)} + \sum_{i=h}^{p} \mathbf{y}_{T+h-i}^{\prime} \mathbf{A}_i^{(j)} + \mathbf{x}_{T+h}^{\prime} \mathbf{C}^{(j)} + \mathbf{u}_{T+h}^{(j)\prime}.$$

   Discarding the parameters yields $\left\{\widetilde{\mathbf{y}}_{T+1}^{(j)}, \ldots \widetilde{\mathbf{y}}_{T+H}^{(j)}\right\}_{j=B+1}^{B+R}$ as a sample from the joint predictive distribution.

---

likelihood (20) with respect to π_1. In the application to forecasting bond yields for 18 different maturities ranging from 1 to 120 months the marginal likelihood is maximized with π_1 close to 0.003. The forecasting exercise includes, in addition to the direct BVAR, several atheoretic time series models as well as theory based forecasting models. Overall the BVAR performs best when forecast performance is measured by the root mean square error. The picture is less clear when the forecasts are evaluated using economic measures. Carriero et al. (2012) considers two different trading strategies. For the first strategy there is no clear ranking of the models when the different maturities are considered, for the second strategy the BVAR delivers the best result for maturities longer than 21 months.

A Markov chain Monte Carlo Methods

A.1 Gibbs sampler

The Gibbs sampler is particularly well suited to Bayesian computation since it is based on the conditional distributions of subsets of the parameter vector. It is frequently the case that it is easy to generate random numbers from the conditional posteriors even if the joint posterior for all the parameters is non-standard. A case in point is regression models like the VAR-model in section 2.1.1 where the posterior distribution of the regression parameters $\boldsymbol{\gamma}$ conditional on the error variance-covariance is normal and the posterior distribution of the variance-covariance matrix $\boldsymbol{\Psi}$ conditional on $\boldsymbol{\gamma}$ is inverse Wishart. MCMC is not needed in that particular case but we will see that this results carries over to situations where it is not possible to generate random numbers directly from the joint posterior.

The recipe for constructing a Gibbs sampler is as follows.

1. Find a suitable partition of the parameter vector into k subsets $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_k')'$
2. Obtain the set of *full conditional* posterior distributions for the subvectors
$$p\left(\boldsymbol{\theta}_i | \mathbf{Y}_T, \boldsymbol{\theta}_{-i}\right), \ i = 1, \ldots, k$$
3. Draw $j+1$ from the Gibbs sampler is given by generating the subvectors from the full conditional posteriors while updating the conditioning
$$\begin{aligned}
\boldsymbol{\theta}_1^{(j+1)} &\sim p\left(\boldsymbol{\theta}_1 | \mathbf{Y}_T, \boldsymbol{\theta}_2^{(j)}, \ldots \boldsymbol{\theta}_k^{(j)}\right) \\
\boldsymbol{\theta}_2^{(j+1)} &\sim p\left(\boldsymbol{\theta}_2 | \mathbf{Y}_T, \boldsymbol{\theta}_1^{(j+1)}, \boldsymbol{\theta}_3^{(j)}, \ldots \boldsymbol{\theta}_k^{(j)}\right) \\
&\vdots \\
\boldsymbol{\theta}_k^{(j+1)} &\sim p\left(\boldsymbol{\theta}_k | \mathbf{Y}_T, \boldsymbol{\theta}_1^{(j+1)}, \ldots \boldsymbol{\theta}_{k-1}^{(j+1)}\right).
\end{aligned}$$

It is easy to verify that the joint posterior distribution is a stationary distribution of the Gibbs sampler. For the simple case with two subvectors $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ the transition kernel is $f\left(\boldsymbol{\theta}^{(j+1)} | \boldsymbol{\theta}^{(j)}\right) = p\left(\boldsymbol{\theta}_2^{(j+1)} | \mathbf{Y}_T, \boldsymbol{\theta}_1^{(j+1)}\right) p\left(\boldsymbol{\theta}_1^{(j+1)} | \mathbf{Y}_T, \boldsymbol{\theta}^{(j)}\right)$. If $\boldsymbol{\theta}^{(j)}$ is a draw from the posterior $p\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{Y}_T\right) = p\left(\boldsymbol{\theta}_1 | \mathbf{Y}_T, \boldsymbol{\theta}_2\right) p\left(\boldsymbol{\theta}_2 | \mathbf{Y}_T\right)$ marginalizing out $\boldsymbol{\theta}^{(j)}$ from the joint distribution of $\boldsymbol{\theta}^{(j+1)}$ and $\boldsymbol{\theta}^{(j)}$ yields the posterior distribution

$$\begin{aligned}
&\int\int p\left(\boldsymbol{\theta}_2^{(j+1)} | \mathbf{Y}_T, \boldsymbol{\theta}_1^{(j+1)}\right) p\left(\boldsymbol{\theta}_1^{(j+1)} | \mathbf{Y}_T, \boldsymbol{\theta}_2^{(j)}\right) p\left(\boldsymbol{\theta}_1^{(j)} | \mathbf{Y}_T, \boldsymbol{\theta}_2^{(j)}\right) p\left(\boldsymbol{\theta}_2^{(j)} | \mathbf{Y}_T\right) d\boldsymbol{\theta}_1^{(j)} d\boldsymbol{\theta}_2^{(j)} \\
&= \int p\left(\boldsymbol{\theta}_2^{(j+1)} | \mathbf{Y}_T, \boldsymbol{\theta}_1^{(j+1)}\right) p\left(\boldsymbol{\theta}_1^{(j+1)} | \mathbf{Y}_T, \boldsymbol{\theta}_2^{(j)}\right) p\left(\boldsymbol{\theta}_2^{(j)} | \mathbf{Y}_T\right) d\boldsymbol{\theta}_2^{(j)} \\
&= p\left(\boldsymbol{\theta}_2^{(j+1)} | \mathbf{Y}_T, \boldsymbol{\theta}_1^{(j+1)}\right) p\left(\boldsymbol{\theta}_1^{(j+1)} | \mathbf{Y}_T\right) = p\left(\boldsymbol{\theta}_1^{(j+1)}, \boldsymbol{\theta}_2^{(j+1)} | \mathbf{Y}_T\right).
\end{aligned}$$

The Gibbs sampler is thus quite straightforward to implement and the form of the sampler follows directly from the model when the full conditionals are well known distributions. This makes it very appealing. There is, however, no guarantee that a naively

implemented Gibbs sampler will perform well or even that it is convergent. Reparameterizing the model or modifying the blocks (the partition into subvectors) to put highly correlated parameters into the same block can often improve the performance and speed of convergence dramatically.

A.2 Metropolis-Hastings

The Metropolis-Hastings algorithm is a more general method that does not rely on the availability of tractable full conditionals. The basic idea is similar to acceptance-rejectance sampling and importance sampling in that an approximation to the desired distribution is used to generate a proposal for the next draw from the chain. The proposal is accepted or rejected based on how well it agrees with the desired distribution and by a judicious choice of the acceptance probability on can obtain a Markov chain with the desired distribution as it's stationary distribution.

In Metropolis-Hastings the proposal distribution itself is allowed to be a Markov chain and the proposed value for $\boldsymbol{\theta}^{(j+1)}$ can depend on the current value $\boldsymbol{\theta}^{(j)}$ through the conditional distribution $q\left(\mathbf{x}|\boldsymbol{\theta}^{(j)}\right)$. The algorithm is as follows

1. Draw a proposal $\mathbf{x}$ from the conditional distribution $q\left(\mathbf{x}|\boldsymbol{\theta}^{(j)}\right)$.

2. Set $\boldsymbol{\theta}^{(j+1)} = \mathbf{x}$ with probability

$$\alpha\left(\boldsymbol{\theta}^{(j)}, \mathbf{x}\right) = \min\left(1, \frac{p\left(\mathbf{x}|\mathbf{Y}_T\right)/q\left(\mathbf{x}|\boldsymbol{\theta}^{(j)}\right)}{p\left(\boldsymbol{\theta}^{(j)}|\mathbf{Y}_T\right)/q\left(\boldsymbol{\theta}^{(j)}|\mathbf{x}\right)}\right) \tag{124}$$

   and keep the old value otherwise, $\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)}$. That is, generate u from a uniform distribution on $(0,1)$ and set $\boldsymbol{\theta}^{(j+1)} = \mathbf{x}$ if $u \leq \alpha$ and $\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)}$ otherwise.

The transition kernel of the resulting Markov chain is given by the conditional distribution

$$f\left(\boldsymbol{\theta}^{(j+1)}|\boldsymbol{\theta}^{(j)}\right) = \begin{cases} q\left(\boldsymbol{\theta}^{(j+1)}|\boldsymbol{\theta}^{(j)}\right)\alpha\left(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^{(j+1)}\right), & \boldsymbol{\theta}^{(j+1)} \neq \boldsymbol{\theta}^{(j)} \\ q\left(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(j)}\right) + \int_{\mathbf{x}\neq\boldsymbol{\theta}^{(j)}} q\left(\mathbf{x}|\boldsymbol{\theta}^{(j)}\right)\left(1-\alpha\left(\boldsymbol{\theta}^{(j)}, \mathbf{x}\right)\right)d\mathbf{x}, & \boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} \end{cases}.$$

That the Markov chain has the posterior as a stationary distribution can be checked by verifying that the detailed balance condition

$$f\left(\boldsymbol{\theta}^{(j)}\right)p\left(\boldsymbol{\theta}^{(j+1)}|\boldsymbol{\theta}^{(j)}\right) = f\left(\boldsymbol{\theta}^{(j+1)}\right)p\left(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(j+1)}\right)$$

holds for $f(\cdot)$ the posterior distribution. Note that the detailed balance condition implies a form of symmetry, the chain moves from $\mathbf{x}$ to $\mathbf{y}$ at the same rate as it moves from $\mathbf{y}$ to $\mathbf{x}$.

Note that any constants cancel from the acceptance probability α and it can be written in terms of the product of the likelihood and the prior instead of the, typically, unknown joint posterior. That is

$$\alpha\left(\boldsymbol{\theta}^{(j)}, \mathbf{x}\right) = \min\left(1, \frac{L\left(\mathbf{Y}_T|\mathbf{x}\right)\pi\left(\mathbf{x}\right)/q\left(\mathbf{x}|\boldsymbol{\theta}^{(j)}\right)}{L\left(\mathbf{Y}_T|\boldsymbol{\theta}^{(j)}\right)\pi\left(\boldsymbol{\theta}^{(j)}\right)/q\left(\boldsymbol{\theta}^{(j)}|\mathbf{x}\right)}\right)$$

The choice of the proposal distribution q is crucial for the performance of the Markov chain and it is important that it is well tailored to the posterior distribution. Examples of common types of proposal chains are

- Independence chain: The proposal steps are drawn from a fix density, $q\left(\mathbf{x}|\boldsymbol{\theta}^{(j)}\right) = f\left(\mathbf{x}\right)$. It is important for the performance of the Markov chain that the proposal distribution is well tailored to the posterior over the whole parameter space which can be difficult with high dimensional parameter vectors. There are, on the other hand, theoretical advantages, the resulting Metropolis chain is uniformly ergodic if the weights $p\left(\mathbf{x}|\mathbf{Y}_T\right)/q\left(\mathbf{x}|\boldsymbol{\theta}^{(j)}\right)$ are bounded on the parameter space.

- Random walk chain: The proposal steps follow a random walk, $\mathbf{x} = \boldsymbol{\theta}^{(j)} + \mathbf{e}$, $q\left(\mathbf{x}|\boldsymbol{\theta}^{(j)}\right) = f\left(\mathbf{x} - \boldsymbol{\theta}^{(j)}\right)$ where f is the density of $\mathbf{e}$. The random walk chain makes it possible to construct a proposal density that matches the posterior well locally, but the proposal should not be to local or the chain will move very slowly through the parameter space.

It is possible to divide the parameter vector into blocks, just as with the Gibbs sampler, and update one block at a time with different proposal distributions for each block of parameters. This makes it easier to adapt the proposal to the posterior and can make for a better performing Markov chain. With a partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_k')$ and updating in order, the update for block m is analogous to the update for the full parameter vector.

1. Propose $\mathbf{x}_m$ from the proposal density $q_m\left(\mathbf{x}_m, \boldsymbol{\theta}_1^{(j+1)}, \ldots, \boldsymbol{\theta}_{m-1}^{(j+1)}, \boldsymbol{\theta}_m^{(j)}, \ldots, \boldsymbol{\theta}_k^{(j)}\right)$.

2. Accept the proposal and set $\boldsymbol{\theta}_m^{(j+1)} = \mathbf{x}_k$ with probability α given by (124) otherwise set $\boldsymbol{\theta}_m^{(j+1)} = \boldsymbol{\theta}_m^{(j)}$.

Note that the acceptance probability simplifies and can be written in terms of the full conditional posterior for $\boldsymbol{\theta}_m$ if this is available,

$$\alpha\left(\boldsymbol{\theta}^{(j)}, \mathbf{x}_m\right) = \min\left(1, \frac{p\left(\mathbf{x}_m|\mathbf{Y}_T, \boldsymbol{\theta}_1^{(j+1)}, \ldots, \boldsymbol{\theta}_{m-1}^{(j+1)}, \boldsymbol{\theta}_{m+1}^{(j)}, \ldots, \boldsymbol{\theta}_k^{(j)}\right)}{q_m\left(\mathbf{x}_m|\boldsymbol{\theta}_1^{(j+1)}, \ldots, \boldsymbol{\theta}_{m-1}^{(j+1)}, \boldsymbol{\theta}_m^{(j)}, \ldots, \boldsymbol{\theta}_k^{(j)}\right)} \Bigg/ \frac{p\left(\boldsymbol{\theta}^{(j)}|\mathbf{Y}_T, \boldsymbol{\theta}_1^{(j+1)}, \ldots, \boldsymbol{\theta}_{m-1}^{(j+1)}, \boldsymbol{\theta}_{m+1}^{(j)}, \ldots, \boldsymbol{\theta}_k^{(j)}\right)}{q_m\left(\boldsymbol{\theta}_m^{(j)}|\boldsymbol{\theta}_1^{(j+1)}, \ldots, \boldsymbol{\theta}_{m-1}^{(j+1)}, \mathbf{x}, \boldsymbol{\theta}_{m+1}^{(j)}, \ldots, \boldsymbol{\theta}_k^{(j)}\right)}\right).$$

In this case the full conditional posterior is an excellent proposal density and with this choice of q_m the acceptance ratio simplifies to one.

The Gibbs sampler is thus a special case of the Metropolis-Hastings algorithm and we can use a mix of Metropolis-Hastings updates and Gibbs updates in the Markov chain. Gibbs updates for the components with convenient full conditional posteriors and Metropolis-Hastings for the other components. Although somewhat of a misnomer, this is commonly known as a Metropolis-Hastings within Gibbs chain. In this context it is useful to note that it is sufficient for uniform ergodicity that one of the Metropolis-Hasting steps uses an independence proposal with bounded weights

$$p\left(\mathbf{x}_m|\mathbf{Y}_T, \boldsymbol{\theta}_1^{(j+1)}, \ldots, \boldsymbol{\theta}_{m-1}^{(j+1)}, \boldsymbol{\theta}_{m+1}^{(j)}, \ldots, \boldsymbol{\theta}_k^{(j)}\right)/q_m\left(\mathbf{x}_m\right).$$

A.3 Autocorrelation in the Markov chain

The output from a Markov chain is by construction autocorrelated and this affects the precision of estimates of posterior quantities sometimes to the point where they are close to being unusable. Ideally one would go back to the drawing board and construct a Markov chain with lower autocorrelation that mixes well. This is, however, not always possible and one must then be particularly careful in the choice of burn-in and make sure that the Markov chain runs long enough to explore the full parameter space.

A common strategy in these situations is to *thin the chain,* i.e. to retain only every m^{th} draw from the chain where m is chosen to make the autocorrelation between $\boldsymbol{\theta}^{(j)}$ and $\boldsymbol{\theta}^{(j+m)}$ negligible. Based on a sample of size R/m after discarding the burn-in we can then estimate the posterior mean of a function $g(\cdot)$ of the parameters as

$$\overline{g}_{R/m} = \frac{m}{R}\sum_{i=1}^{R/m} g\left(\boldsymbol{\theta}^{([i-1]m+1)}\right)$$

and an estimate of the numerical standard error is given by the square root of

$$\widehat{V}\left(\overline{g}_{R/m}\right) = \frac{\widehat{V}\left(g\left(\boldsymbol{\theta}\right)|\mathbf{Y}_T\right)}{R/m}.$$

This is a statistically inefficient procedure and it can be shown that $V\left(\overline{g}_{R/m}\right) \geq V\left(\overline{g}_R\right)$. On the other hand it might reduce the storage and memory requirements considerably.

If the chain is not thinned or when the thinning leaves some autocorrelation this must be accounted for when estimating the numerical standard errors. Two common methods is the *batched mean* method and the time series based *spectral estimate.* The batched mean method divides the data into R/m batches, each containing m consecutive draws from the Markov chain, and calculate the batch means

$$\overline{g}_{m,j} = \frac{1}{m}\sum_{i=(j-1)m+1}^{jm} g\left(\boldsymbol{\theta}^{(i)}\right).$$

For sufficiently large m the batch means will be essentially uncorrelated and we can estimate the variance of the batch means as

$$\widehat{V}\left(\overline{g}_m\right) = \frac{1}{R/m-1}\sum_{j=1}^{R/m}\left(\overline{g}_{m,j}-\overline{g}_R\right)^2.$$

An estimate of the variance of $\overline{g}_R$ is then given by

$$\widehat{V}\left(\overline{g}_R\right) = \frac{m}{R}\widehat{V}\left(\overline{g}_m\right) = \frac{m}{R\left(R/m-1\right)}\sum_{j=1}^{R/m}\left(\overline{g}_{m,j}-\overline{g}_R\right)^2.$$

The so called spectral estimate is simply the Newey and West (1987) autocorrelation consistent estimator of the asymptotic variance (12). A common implementation is the estimator

$$\widehat{V}\left(\overline{g}_R\right) = \frac{1}{R}\sum_{j=-m}^{m}\left(1-\frac{|j|}{m+1}\right)\widehat{\gamma}_j$$

with truncation at lag m and the autocovariances,

$$\widehat{\gamma}_j = \frac{1}{R}\sum_{i=1}^{R-j}\left[g\left(\boldsymbol{\theta}^{(i)}\right) - \overline{g}_R\right]\left[g\left(\boldsymbol{\theta}^{(i+j)}\right) - \overline{g}_R\right],$$

at larger lags are downweighted using a Bartlett kernel. For consistency the truncation should go to infinity with R, $m = o\left(R^{1/4}\right)$.

The autocorrelation will in general lead to a loss of efficiency compared to the case when we can generate iid draws from the posterior. It is common to measure the loss with the relative numerical efficiency (RNE)

$$RNE = \frac{\widehat{V}\left(g\left(\boldsymbol{\theta}\right)\right)/R}{\widehat{V}\left(\overline{g}\right)}.$$

An alternative measure is the *effective sample size*, the number of iid draws that would give the same numerical standard error as the R draws we have from the sampler. This is simply R times the RNE.

A.4 Assessing convergence

It should be clear from the discussion above that it can not be taken for granted that a Gibbs or Metropolis-Hastings sampler converges to the desired posterior distribution. Nor that, if the sampler is convergent, it does converge in a reasonable number of steps and that the output can be used to compute reliable estimates of posterior quantities. Trying to assess if the sampler fails to converge or not and the approximate number of steps required to be "close enough" to convergence is thus important. Even if convergence can be proved for a particular sampler there is very little information about how quickly it converges and an empirical assessment of the amount of burn-in needed must be made. Unfortunately, the output from the chain only constitutes a sample and can not be used to prove convergence – all we can do is to look for signs of lack of convergence or slow convergence.

Some of the most powerful diagnostics or indicators of problems are quite simple in nature. High and persistent autocorrelation in the chain indicates slow mixing and slow convergence to the posterior distribution. Is the posterior multimodal? If so, the chain might get stuck at one of the modes if the probability mass connecting the modes is small. Simple plots of the output, trace plots of the parameters, $\theta_i^{(j)}$, or some function of the parameters, $g\left(\boldsymbol{\theta}^{(j)}\right)$, plots of running means, $\overline{g}_t = \frac{1}{t}\sum_{j=1}^{t} g\left(\boldsymbol{\theta}^{(j)}\right)$, or CUSUMs, $S_t = \sum_{j=1}^{t}\left[g\left(\boldsymbol{\theta}^{(j)}\right) - \overline{g}_R\right]$, can also be informative. If the trace plot or the running means settle down after a number of steps this can indicate a suitable amount of burn-in. Similarly for the CUSUM plots. In addition Yu and Mykland (1998) argue that the CUSUM plot can be informative about how well the sampler mixes, "a good sampler should have an oscillatory path plot and small excursions; or a bad sampler should have a smooth path plot and large excursions".

Brooks (1998) proposed a formal test for deviations from the ideal case of iid output from the sampler based on the CUSUM. First determine a suitable amount of burn-in, B, based on preliminary plots of the output and calculate $\widehat{\mu} = (R-B)^{-1}\sum_{j=B+1}^{R} g\left(\boldsymbol{\theta}^{(j)}\right)$

and $S_t = \sum_{j=B+1}^{t} \left[g\left(\boldsymbol{\theta}^{(j)}\right) - \widehat{\mu} \right]$ for $t = B+1, \ldots, R$ and $g(\cdot)$ some function of the parameters. Next define

$$d_j = \begin{cases} 1 \text{ if } S_{j-1} > S_j \text{ and } S_j < S_{j+1} \text{ or } S_{j-1} < S_j \text{ and } S_j > S_{j+1} \\ 0 \text{ otherwise} \end{cases}$$

where $d_j = 1$ indicates non-smoothness or "hairiness" of the plot and $d_j = 0$ indicates smoothness. The running means $D_t = (t - B - 1)^{-1} \sum_{j=B+1}^{t-1} d_j$ for $t = B+2, \ldots, R$ lies between 0 and 1 and captures the overall behavior of the Markov chain. If we, in addition to the iid assumption, assume that $g\left(\boldsymbol{\theta}^{(j)}\right)$ is symmetric around the mean we have $P(d_j = 1) = 1/2$ and D_t is Binomially distributed. We can then plot D_t against the bounds $\pm Z_{\alpha/2} \sqrt{\frac{1}{4(t-B-1)}}$ and diagnose non-convergence if D_T fails to lie within the bounds $100(1-\alpha)\%$ of the time.

Geweke (1992) proposed monitoring convergence by the statistic

$$z_G = \frac{\overline{g}_a - \overline{g}_b}{\sqrt{V(\overline{g}_a) + V(\overline{g}_b)}}$$

where $g(\cdot)$ is some function of the output of the chain and

$$\overline{g}_a = \frac{1}{n_a} \sum_{j=m+1}^{m+n_a} g\left(\boldsymbol{\theta}^{(j)}\right), \; \overline{g}_b = \frac{1}{n_b} \sum_{j=R-n_b+1}^{R} g\left(\boldsymbol{\theta}^{(j)}\right)$$

for a chain that is run R steps with $R > n_a + n_b + m$ and the distance between the estimates such that they can be taken to be uncorrelated. The variances are estimated taking account of the autocorrelation structure in the chain, for example by the spectral estimate above. If the chain has converged after m steps the distribution of the draws $m+1, \ldots, m+n_a$ is the same as the distribution of the draws at the end of the chain and z_G approximately standard normal. Calculating z_G for a range of values of m and comparing to critical values from a standard normal will thus give an indication of the burn in needed for the chain.

Gelman and Rubin (1992), proposed running several shorter chains started at points that are overdispersed compared to the posterior. Let $\boldsymbol{\theta}_i^{(j)}$ denote the output from chain i for m chains run $n = R - B$ steps from burn-in and define the between and within chain variances as

$$B = \frac{n}{m-1} \sum_{i=1}^{m} (\overline{g}_i - \overline{g})^2, \quad W = \frac{1}{m(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(g\left(\boldsymbol{\theta}_i^{(j)}\right) - \overline{g}_i \right)^2.$$

$$\overline{g}_i = \frac{1}{n} \sum_{j=1}^{n} g\left(\boldsymbol{\theta}_i^{(j)}\right), \quad \overline{g} = \frac{1}{m} \sum_{i=1}^{m} \overline{g}_i$$

Convergence failure or convergence on different stationary distributions after the selected burn-in is indicated by the between chain variation, B, being larger than the within chain variation, W. If the chains have converged after B draws we have two unbiased estimates of the variance, $V = (1 - 1/n) W + B/n$ and W. The first tends to overestimate the variance if convergence has not been achieved (the between chain variation is large) and

the latter tends to underestimate the variance (the chains have not had time to explore the full parameter space). The convergence diagnostic is $r = \sqrt{V/W}$ or a version including a "degree of freedom" correction. Gelman (1996) suggested the rule of thumb to accept convergence if $r < 1.2$ for all monitored quantities.

The Brooks and Geweke diagnostics and Gelman-Rubin diagnostics are quite different in nature. The Brooks and Geweke diagnostics are based on a single long chain and will fail to detect convergence failures caused by the chain being stuck at one of the modes of a multimodal posterior. The Gelman-Rubin statistic, on the other hand, is more likely to detect this type of problem but is much less informative about the amount of burn-in needed.

B State space models

Consider the linear state space model for the m observed variables in $\mathbf{y}_t$, $t = 1, \ldots, T$,

$$\begin{aligned} \mathbf{y}_t &= \mathbf{Z}_t \mathbf{s}_t + \boldsymbol{\varepsilon}_t, \ \boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \mathbf{H}_t) \\ \mathbf{s}_{t+1} &= \mathbf{d}_t + \mathbf{T}_t \mathbf{s}_t + \boldsymbol{\eta}_t, \ \boldsymbol{\eta}_t \sim N(\mathbf{0}, \mathbf{Q}_t) \end{aligned} \tag{125}$$

with the initial condition or prior on the first state, $\mathbf{s}_1 \sim N\left(\mathbf{s}_{1|0}, \mathbf{P}_{1|0}\right)$. The n dimensional state vectors $\mathbf{s}_t$ are unobserved and the matrices $\mathbf{Z}_t$, $\mathbf{H}_t$, $\mathbf{T}_t$ and $\mathbf{Q}_t$ are assumed known for the purpose of the discussion here (they are in general functions of the data, unknown parameters or simply known constants). The subscript $t|s$ indicates a time t property conditional on information up to time s, a superscript t indicates a sequence running from 1 to t, e.g. $\mathbf{s}_{i|j} = E(\mathbf{s}_i|\mathbf{y}^j) = E(\mathbf{s}_i|\mathbf{y}_1, \ldots, \mathbf{y}_j)$.

General references on state space models include Harvey (1989) and Durbin and Koopman (2001), West and Harrison (1997) and Kim and Nelson (1999) provides a Bayesian treatment and Giordini, Pitt and Kohn (2011) reviews Bayesian inference in general state space models. The Kalman filter and smoothing algorithms given below are standard. The version of the simulation smoother is due to Carter and Kohn (1994). There are many variations on these algorithms, the ones given here are straightforward and intuitive but not the most computationally efficient versions.

B.1 Kalman filter

The Kalman filter runs forward through the data and returns the means and variances of the conditional distributions $\mathbf{s}_t|\mathbf{y}^t \sim N\left(\mathbf{s}_{t|t}, \mathbf{P}_{t|t}\right)$ and $\mathbf{s}_{t+1}|\mathbf{y}^t \sim N\left(\mathbf{s}_{t+1|t}, \mathbf{P}_{t+1|t}\right)$,

$$\begin{array}{ll} \mathbf{v}_t = \mathbf{y}_t - \mathbf{Z}_t \mathbf{s}_{t|t-1} & \mathbf{F}_t = \mathbf{Z}_t \mathbf{P}_{t|t-1} \mathbf{Z}_t' + \mathbf{H}_t \\ \mathbf{K}_t = \mathbf{P}_{t|t-1} \mathbf{Z}_t' \mathbf{F}_t & \mathbf{s}_{t|t} = \mathbf{s}_{t|t-1} + \mathbf{K}_t \mathbf{v}_t \\ \mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{K}_t \mathbf{Z}_t \mathbf{P}_{t|t-1} & \\ \mathbf{s}_{t+1|t} = \mathbf{d}_t + \mathbf{T}_t \mathbf{s}_{t|t} & \mathbf{P}_{t+1|t} = \mathbf{T}_t \mathbf{P}_{t|t} \mathbf{T}_t' + \mathbf{Q}_t, \end{array} \tag{126}$$

for $t = 1, \ldots, T$.

B.2 Smoothing

At the end of the filtering run we have the distribution of the last state, $\mathbf{s}_T|\mathbf{y}^T$, conditional on all the data but for the earlier states we only have the distribution conditional on a

Algorithm 18 Simulation Smoother

1. Generate $\mathbf{s}_T$ from the conditional distribution, $\mathbf{s}_T|\mathbf{y}^T \sim N\left(\mathbf{s}_{T|T}, \mathbf{P}_{T|T}\right)$
2. For $t = T-1, \ldots, 1$
   (a) Calculate
   $$\begin{aligned}\mathbf{s}_{t|t,\mathbf{s}_{t+1}} &= \mathbf{s}_{t|t} + \mathbf{P}_{t|t}\mathbf{T}_t'\mathbf{P}_{t+1|t}^{-1}\left(\mathbf{s}_{t+1} - \mathbf{s}_{t+1|t}\right)\\ \mathbf{P}_{t|t,\mathbf{s}_{t+1}} &= \mathbf{P}_{t|t} - \mathbf{P}_{t|t}\mathbf{T}_t'\mathbf{P}_{t+1|t}^{-1}\mathbf{T}_t\mathbf{P}_{t|t}\end{aligned}$$
   (b) Generate $\mathbf{s}_t$ from the conditional distribution $\mathbf{s}_t|\mathbf{y}^t, \mathbf{s}_{t+1} = \mathbf{s}_t|\mathbf{y}^T, \mathbf{s}_{t+1} \sim N\left(\mathbf{s}_{t|t,\mathbf{s}_{t+1}}, \mathbf{P}_{t|t,\mathbf{s}_{t+1}}\right)$.

subset of the data and all information has not been used. The fixed-interval smoother runs backwards through the data and returns the means and variances of the conditional distributions $\mathbf{s}_t|\mathbf{y}^T \sim N\left(\mathbf{s}_{t|T}, \mathbf{P}_{t|T}\right)$,

$$\begin{aligned}\mathbf{s}_{t|T} &= \mathbf{s}_{t|t} + \mathbf{P}_{t|t}\mathbf{T}_t'\mathbf{P}_{t+1|t}^{-1}\left(\mathbf{s}_{t+1|T} - \mathbf{s}_{t+1|t}\right)\\ \mathbf{P}_{t|T} &= \mathbf{P}_{t|t} - \mathbf{P}_{t|t}\mathbf{T}_t'\mathbf{P}_{t+1|t}^{-1}\left(\mathbf{P}_{t+1|T} - \mathbf{P}_{t+1|t}\right)\mathbf{P}_{t+1|t}^{-1}\mathbf{T}_t\mathbf{P}_{t|t}\end{aligned} \tag{127}$$

for $t = T-1, \ldots, 1$.

B.3 Simulation smoother

The simulation smoother is a device for generating random numbers from the joint distribution of the states conditional on the data, $\mathbf{s}^T|\mathbf{y}^T$. The output from the fixed-interval smoother can not be used for this since it carries no information about the dependence between the states at different time points. The simulation smoother is based on the partition

$$p\left(\mathbf{s}_1, \ldots, \mathbf{s}_T|\mathbf{y}^T\right) = p\left(\mathbf{s}_T|\mathbf{y}^T\right)\prod_{t=1}^{t-1} p\left(\mathbf{s}_t|\mathbf{y}^T, \mathbf{s}_{t+1}\right)$$

and generates a draw from the joint distribution by working backwards through the data and generating $\mathbf{s}_t$ from the conditional distributions.

C Distributions

Definition 1 (Gamma) *x is Gamma distributed with shape parameter α and inverse scale parameter β, $x \sim G(\alpha, \beta)$ if the density is*

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x).$$

We have $E(x) = \alpha/\beta$ and $V(x) = \alpha/\beta^2$.

Definition 2 (Inverse Gamma) *$y = x^{-1}$ is inverse Gamma distributed, $y \sim iG(\alpha, \beta)$ if x is Gamma distributed $x \sim G(\alpha, \beta)$. The density of y is*

$$f(y) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{-(\alpha+1)} \exp\left(-\frac{\beta}{y}\right)$$

with moments $E(y) = \beta/(\alpha - 1)$ and $V(y) = \beta/\left[(\alpha-1)^2(\alpha-2)\right]$.

Definition 3 (Matricvariate normal) *The $p \times q$ matrix $\mathbf{X}$ is said to have a matricvariate normal distribution*

$$\mathbf{X} \sim MN_{pq}(\mathbf{M}, \mathbf{Q}, \mathbf{P})$$

where $\mathbf{M}$ is $p \times q$ and $\mathbf{P}$ and $\mathbf{Q}$ are positive definite symmetric matrices of dimensions $p \times p$ and $q \times q$ if $\operatorname{vec}(\mathbf{X})$ is multivariate normal

$$\operatorname{vec}(\mathbf{X}) \sim N(\operatorname{vec}(\mathbf{M}), \mathbf{Q} \otimes \mathbf{P}).$$

The density of $\mathbf{X}$ is

$$\begin{aligned}
&MN_{pq}(\mathbf{X}; \mathbf{M}, \mathbf{Q}, \mathbf{P}) \\
&= (2\pi)^{-pq/2} |\mathbf{Q} \otimes \mathbf{P}|^{-1/2} \\
&\times \exp\left[-\frac{1}{2}(\operatorname{vec}(\mathbf{X}) - \operatorname{vec}(\mathbf{M}))' \left(\mathbf{Q}^{-1} \otimes \mathbf{P}^{-1}\right)(\operatorname{vec}(\mathbf{X}) - \operatorname{vec}(\mathbf{M}))\right] \\
&= (2\pi)^{-pq/2} |\mathbf{Q}|^{-p/2} |\mathbf{P}|^{-q/2} \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\mathbf{Q}^{-1}(\mathbf{X} - \mathbf{M})' \mathbf{P}^{-1}(\mathbf{X} - \mathbf{M})\right]\right\}.
\end{aligned}$$

Remark 1 *The defining feature of the matricvariate normal is the Kronecker structure for the variance-covariance matrix. $\mathbf{Q}$ is proportional to the variance matrix of the rows of. $\mathbf{X}$ and $\mathbf{P}$ is proportional to the variance matrix of the columns of $\mathbf{X}$. The elements in row i are correlated with the elements in row j if $p_{ij} \neq 0$ and the elements in column i are correlated with the elements in column j if $q_{ij} \neq 0$.*

Remark 2 *Suppose that $\mathbf{X} \sim MN_{pq}(\mathbf{M}, \mathbf{Q}, \mathbf{P})$*

1. $\mathbf{X}' \sim MN_{qp}(\mathbf{M}', \mathbf{P}, \mathbf{Q})$.
2. $\mathbf{AXB} \sim MN_{kl}(\mathbf{AMB}, \mathbf{B}'\mathbf{QB}, \mathbf{APA}')$ *for* $\mathbf{A}$ $k \times p$ *and* $\mathbf{B}$ $q \times l$.

Algorithm 19 Matricvariate normal random number generator

To generate $\mathbf{X} \sim MN_{pq}(\mathbf{M}, \mathbf{Q}, \mathbf{P})$ calculate the Cholesky factors of $\mathbf{Q}$ and $\mathbf{P}$, $\mathbf{Q} = \mathbf{LL}'$, $\mathbf{P} = \mathbf{CC}'$, generate $\mathbf{Y}$ as a $p \times q$ matrix of standard normals and calculate $\mathbf{X} = \mathbf{M} + \mathbf{CYL}' \sim MN_{pq}(\mathbf{M}, \mathbf{Q}, \mathbf{P})$.

Definition 4 (Wishart) *A $q \times q$ positive semi definite symmetric matrix $\mathbf{A}$ is said to have a Wishart distribution, $\mathbf{A} \sim W_q(\mathbf{B},v)$ if it's density is given by*

$$W_q(\mathbf{A};\mathbf{B},v) = k^{-1}|\mathbf{B}|^{-v/2}|\mathbf{A}|^{(v-q-1)/2}\exp\left[-\frac{1}{2}\operatorname{tr}\mathbf{A}\mathbf{B}^{-1}\right]$$

for $\mathbf{B}$ a positive definite symmetric matrix and $v \geq q$ degrees of freedom and

$$k = 2^{vq/2}\pi^{q(q-1)/4}\prod_{i=1}^{q}\Gamma\left(\left(v+1-i\right)/2\right).$$

$$\begin{aligned} E(\mathbf{A}) &= v\mathbf{B} \\ V(a_{ij}) &= v\left(b_{ij}^2 + b_{ii}b_{jj}\right) \end{aligned}$$

for a_{ij} one of the $q(q+1)/2$ distinct elements of $\mathbf{A}$.

Remark 3 *The Wishart distribution is a matricvariate generalization of the χ^2 distribution and arises frequently in multivariate analysis. If $\mathbf{x}_i$ are iid. $N(\mu,\mathbf{\Sigma})$ q-dimensional random vectors then $\sum_{i=1}^{n}(\mathbf{x}_i-\mu)(\mathbf{x}_i-\mu)' \sim W_q(\mathbf{\Sigma},n)$ and $\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})' \sim W(\mathbf{\Sigma},n-1)$. If $\mathbf{A} \sim W_q(\mathbf{B},v)$ then $\mathbf{PAP}' \sim W_p(\mathbf{PBP}',v)$ for $\mathbf{P}$ a $p \times q$ matrix of rank p $(p \leq q)$.*

Algorithm 20 Wishart random number generator

Wishart distributed matrices, $\mathbf{A} \sim W_q(\mathbf{B},v)$ can be generated by brute force by first generating v vectors $\mathbf{x}_i \sim N(\mathbf{0},\mathbf{B})$ and forming $\mathbf{A} = \sum_{i=1}^{v}\mathbf{x}_i\mathbf{x}_i'$. A more efficient algorithm is based on the Bartlett decomposition of a Wishart matrix (Anderson (1984)) has been proposed by Smith and Hocking (1972) and Geweke (1988). Let $\mathbf{P}$ be a $q \times q$ lower triangular matrix where $p_{ii}^2 \sim \chi_{v-i+1}^2$ (i.e. p_{ii} is the square root of the χ^2) and $p_{ij} \sim N(0,1)$, $i<j$, then $\mathbf{PP}' \sim W_q(\mathbf{I},v)$. In addition let $\mathbf{L}$ be the lower triangular Cholesky factor of $\mathbf{B} = \mathbf{LL}'$, then $\mathbf{A} = (\mathbf{LP})(\mathbf{LP})' \sim W_q(\mathbf{B},v)$. Note that in many cases it is more convenient to work directly with the lower triangular matrix $\mathbf{C} = \mathbf{LP}$ than $\mathbf{A}$, e.g. when the ultimate objective is to generate random numbers $\mathbf{z} \sim N(\boldsymbol{\mu},\mathbf{A})$. First generate $x_i \sim N(0,1)$, $i = 1,\ldots,q$ and form $\mathbf{z} = \boldsymbol{\mu} + \mathbf{Cx}$.
In some cases it is more convenient with an upper triangular decomposition $\mathbf{QQ}' \sim W_q(\mathbf{I},v)$. For this let $q_{ii}^2 \sim \chi_{v-q+i}^2$ and $q_{ij} \sim N(0,1)$, $i>j$.

Definition 5 (Inverse Wishart) *The $q \times q$ matrix $\mathbf{A}$ is said to have an inverse Wishart distribution,*

$$\mathbf{A} \sim iW_q(\mathbf{B},v)$$

if $\mathbf{A}^{-1} \sim W(\mathbf{B}^{-1},v)$. The density of $\mathbf{A}$ is given by

$$iW_q(\mathbf{A};\mathbf{B},v) = k^{-1}|\mathbf{B}|^{v/2}|\mathbf{A}|^{-(v+q+1)/2}\exp\left[-\frac{1}{2}\operatorname{tr}\mathbf{A}^{-1}\mathbf{B}\right]$$

Algorithm 21 Inverse Wishart random number generator

To generate $\mathbf{A} \sim iW_q(\mathbf{B},v)$, first generate the upper triangular Bartlett decomposition matrix $\mathbf{Q}$ of a Wishart distributed, $W_q(\mathbf{I},v)$, matrix. Second calculate the lower triangular Cholesky decomposition, $\mathbf{LL}' = \mathbf{B}$, we then have $\mathbf{L}^{-\mathsf{T}}\mathbf{L}^{-1} = \mathbf{B}^{-1}$ and $\mathbf{L}^{-\mathsf{T}}\mathbf{QQ}'\mathbf{L}^{-1} \sim W_q(\mathbf{B}^{-1},v)$. Let $\mathbf{C} = \mathbf{LQ}^{-\mathsf{T}}$ and we have $\mathbf{A} = \mathbf{CC}' \sim iW_q(\mathbf{B},v)$ for $\mathbf{C}$ lower triangular. Sometimes $\mathbf{A} \sim iW_q(\mathbf{D}^{-1},v)$ is needed. The inversion of $\mathbf{D}$ can be avoided by letting $\mathbf{L}$ be the Cholesky decomposition of $\mathbf{D}$, $\mathbf{LL}' = \mathbf{D}$, generate the lower triangular Bartlett decomposition matrix $\mathbf{P}$ and let $\mathbf{C}$ be the upper triangular matrix $\mathbf{C} = (\mathbf{LP})^{-\mathsf{T}}$ for $\mathbf{A} = \mathbf{CC}' \sim iW_q(\mathbf{D}^{-1},v)$

with k as for the Wishart distribution.

$$E(\mathbf{A}) = \frac{1}{v-q-1}\mathbf{B},\ v > q+1$$
$$V(a_{ij}) = \frac{(v-q-1)\,b_{ii}b_{jj} + (v-q+1)\,b_{ij}^2}{(v-q-1)^2\,(v-q)\,(v-q-3)},\ v > q+3$$

Definition 6 (normal-Wishart) *If $\mathbf{X}|\mathbf{\Sigma} \sim MN_{pq}(\mathbf{M},\mathbf{\Sigma},\mathbf{P})$ and $\mathbf{\Sigma} \sim iW_q(\mathbf{Q},v)$ then the joint distribution of $\mathbf{X}$ and $\mathbf{\Sigma}$ is said to be normal-Wishart with kernel*

$$p(\mathbf{X},\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-(v+p+q+1)/2} \exp\left\{-\frac{1}{2}\operatorname{tr}\left[\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})'\mathbf{P}^{-1}(\mathbf{X}-\mathbf{M})\right]\right\} \exp\left[-\frac{1}{2}\operatorname{tr}\mathbf{\Sigma}^{-1}\mathbf{B}\right]$$

Algorithm 22 Normal-Wishart random number generator

To generate $\mathbf{X}|\mathbf{\Sigma} \sim MN_{pq}(\mathbf{M},\mathbf{\Sigma},\mathbf{A})$ and $\mathbf{\Sigma} \sim iW_q(\mathbf{B},v)$ first generate the triangular factor $\mathbf{C}$ of an inverse Wishart and calculate $\mathbf{\Sigma} = \mathbf{CC}'$ (if needed). Second generate $\mathbf{Y}$ as a $p \times q$ matrix of standard normals and form $\mathbf{X} = \mathbf{M} + \mathbf{LYC}' \sim MN_{pq}(\mathbf{M},\mathbf{\Sigma},\mathbf{A})$ for $\mathbf{L}$ the Cholesky factor of $\mathbf{A} = \mathbf{LL}'$.

Definition 7 (Matricvariate t) *A random $p \times q$ matrix $\mathbf{X}$ is said to have a matricvariate t distribution if the density is given by*

$$Mt_{pq}(\mathbf{X};\mathbf{M},\mathbf{P},\mathbf{Q},v) = k^{-1}\left|\mathbf{Q} + (\mathbf{X}-\mathbf{M})'\mathbf{P}(\mathbf{X}-\mathbf{M})\right|^{-(v+p)/2}$$

for $\mathbf{M}$ a $p \times q$ mean matrix, $\mathbf{Q}$ and $\mathbf{P}$, $q \times q$ and $p \times p$ positive definite symmetric scale matrices and $v \geq q$ degrees of freedom. The integrating constant is given by

$$k = \pi^{pq/2}|\mathbf{P}|^{-q/2}|\mathbf{Q}|^{-v/2}\prod_{i=1}^{q}\frac{\Gamma((v+1-i)/2)}{\Gamma((v+p+1-i)/2)}.$$

We have

$$E(\mathbf{X}) = \mathbf{M}, v > q$$
$$V(\operatorname{vec}\mathbf{X}) = \frac{1}{v-q-1}\mathbf{Q}\otimes\mathbf{P}^{-1}, v > q+1.$$

Remark 4 *If* $\mathbf{X}|\mathbf{\Sigma} \sim MN_{pq}(\mathbf{M}, \mathbf{\Sigma}, \mathbf{P})$ *and* $\mathbf{\Sigma} \sim iW_q(\mathbf{Q}, v)$ *(normal-Wishart) then the marginal distribution of* $\mathbf{X}$ *is matricvariate* t*.*

$$\mathbf{X} \sim Mt_{pq}\left(\mathbf{M}, \mathbf{P^{-1}}, \mathbf{Q}, v\right).$$

It follows that Algorithm 22 also is a matricvariate t *random number generator where the draws of* $\mathbf{C}$ *or* $\mathbf{\Sigma}$ *are simply discarded.*
In addition, the distribution of $\mathbf{\Sigma}$ *conditional on* $\mathbf{X}$ *is inverse Wishart*

$$\mathbf{\Sigma}|\mathbf{X} \sim iW_q\left(\mathbf{Q} + (\mathbf{X} - \mathbf{M})' \mathbf{P}^{-1} (\mathbf{X} - \mathbf{M}), v + p\right).$$

References

Adolfson, M., Andersson, M. K., Linde, J., Villani, M. and Vredin, A. (2007), 'Modern forecasting models in action: Improving macroeconomic analyses at central banks', *International Journal of Central Banking* **3**, 111–144.

Adolfson, M., Lasen, S., Lind, J. and Villani, M. (2008), 'Evaluating an estimated new keynesian small open economy model', *Journal of Economic Dynamics and Control* **32**(8), 2690–2721.

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis*, 2nd edn, John Wiley & Sons, New York.

Andersson, M. K., Palmqvist, S. and Waggoner, D. F. (2010), Density conditional forecasts in dynamic multivariate models, Working Paper Series 247, Sveriges Riksbank.

Andersson, M. and Karlsson, S. (2009), Bayesian forecast combination for VAR models, *in* Chib, Koop and Griffiths (2008), pp. 501–524.

Banbura, M., Giannone, D. and Reichlin, L. (2010), 'Large bayesian vector auto regressions', *Journal of Applied Econometrics* **25**, 71–92.

Barbieri, M. M. and Berger, J. O. (2004), 'Optimal predictive model selection', *The Annals of Statistics* **32**(3), 870–897.

Bauwens, L. and Lubrano, M. (1996), Identification restrictions and posterior densities in cointegrated gaussian var systems, *in* T. B. Fomby and R. C. Hill, eds, 'Advances in Econometrics', Vol. 11B, JAI Press.

Beechey, M. and Österholm, P. (2010), 'Forecasting inflation in an inflation-targeting regime: A role for informative steady-state priors', *International Journal of Forecasting* **26**(2), 248–264.

Bernanke, B., Boivin, J. and Eliasz, P. (2005), 'Measuring the effect of monetary policy: a factor augmented vector autoregressive (FAVAR) approach', *Quarterly Journal of Economics* **120**, 387–422.

Bloor, C. and Matheson, T. (2010), 'Analysing shock transmission in a data-rich environment: a large bvar for new zealand', *Empirical Economics* **39**, 537–558.

Bloor, C. and Matheson, T. D. (2011), 'Real-time conditional forecasts with bayesian vars: An application to new zealand', *The North American Journal of Economics and Finance* **22**(1), 26–42.

Brooks, S. P. (1998), 'Quantitative convergence assessment for markov chain monte carlo via cusums', *Statistics and Computing* **8**, 267–274.

Brown, P. J., Vanucci, M. and Fearn, T. (1998), 'Multivariate bayesian variable selection and prediction', *Journal of the Royal Statistical Society, Ser. B* **60**, 627–641.

Canova, F. (2007), 'G-7 inflation forecasts: Random walk, phillips curve or what else?', *Macroeconomic Dynamics* **11**(01), 1–30.

Canova, F. and Ciccarelli, M. (2004), 'Forecasting and turning point predictions in a bayesian panel var model', *Journal of Econometrics* **120**, 327–359.

Canova, F. and Ciccarelli, M. (2009), 'Estimating multicountry var models', *International Economic Review* **50**(3), 929–959.

Carriero, A., Clark, T. and Marcellino, M. (2011), Bayesian vars: specification choices and forecast accuracy, Working Paper 1112, Federal Reserve Bank of Cleveland.

Carriero, A., Kapetanios, G. and Marcellino, M. (2009), 'Forecasting exchange rates with a large bayesian var', *International Journal of Forecasting* **25**(2), 400–417.

Carriero, A., Kapetanios, G. and Marcellino, M. (2011), 'Forecasting large datasets with bayesian reduced rank multivariate models', *Journal of Applied Econometrics* **26**, 735–761.

Carriero, A., Kapetanios, G. and Marcellino, M. (2012), 'Forecasting government bond yields with large bayesian vector autoregressions', *Journal of Banking & Finance* **36**(7), 2026 – 2047.

Carter, C. K. and Kohn, R. (1994), 'On gibbs sampling for state space models', *Biometrika* **81**(3), pp. 541–553.

Chib, S. (1995), 'Marginal likelihood from the Gibbs output', *Journal of the American Statistical Association* **90**, 1313–1321.

Chib, S. (1998), 'Estimation and comparison of multiple change point models', *Journal of Econometrics* **86**, 221–241.

Chib, S. and Greenberg, E. (1995), 'Understanding the Metropolis-Hastings algorithm', *American Statistician* **40**, 327–335.

Chib, S., Koop, G. and Griffiths, B., eds (2008), *Bayesian Econometrics*, Vol. 23 of *Advances in Econometrics*, Emerald.

Clark, T. E. (2011), 'Real-time density forecasts from bayesian vector autoregressions with stochastic volatility', *Journal of Business & Economic Statistics* **29**(3), 327–341.

Clark, T. E. and McCracken, M. W. (2010), 'Averaging forecasts from VARs with uncertain instabilities', *Journal of Applied Econometrics* **25**, 5–29.

Clements, M. P. and Hendry, D. F., eds (2011), *The Oxford Handbook of Economic Forecasting*, Oxford University Press.

Cogley, T., Morozov, S. and Sargent, T. J. (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**(11), 1893 – 1925.

Cogley, T. and Sargent, T. J. (2002), Evolving post-world war ii u.s. inflation dynamics, *in* B. S. Bernanke and K. S. Rogoff, eds, 'NBER Macroeconomics Annual 2001, Volume 16', National Bureau of Economic Research, Inc, pp. 331–388.

Cogley, T. and Sargent, T. J. (2005), 'Drifts and volatilities: monetary policies and outcomes in the post wwii us', *Review of Economic Dynamics* **8**, 262–302.

Corradi, V. and Swanson, N. R. (2006), Predictive density evaluation, *in* Elliott, Granger and Timmermann (2006), pp. 197 – 284.

D'Agostino, A., Gambetti, L. and Giannone, D. (forthcoming), 'Macroeconomic forecasting and structural change', *Journal of Applied Econometrics* .

De Mol, C., Giannone, D. and Reichlin, L. (2008), 'Forecasting using a large number of predictors: is bayesian regression a valid alternative to principal components?', *Journal of Econometrics* **146**, 318–328.

DeJong, D. N. (1992), 'Co-integration and trend-stationarity in macroeconomic time series: Evidence from the likelihood function', *Journal of Econometrics* **52**(3), 347–370.

DelNegro, M. and Schorfheide, F. (2011), Bayesian methods in microeconometrics, *in* Clements and Hendry (2011), chapter 7, pp. 293–389.

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

Doan, T., Litterman, R. B. and Sims, C. (1984), 'Forecasting and conditional projection using realistic prior distributions', *Econometric Reviews* **3**, 1–144.

Dorfman, J. H. (1995), 'A numerical bayesian test for cointegration of ar processes', *Journal of Econometrics* **66**, 289–324.

Drèze, J. H. and Morales, J. A. (1976), 'Bayesian full information analysis of simultaneous equations', *Journal of the American Statistical Association* **71**, 919–923.

Durbin, J. and Koopman, S. J. (2001), *Time Series Analysis by State Space Methods*, Oxford University Press.

Elliott, G., Granger, C. W. J. and Timmermann, A., eds (2006), *Handbook of Economic Forecasting*, Vol. 1, Elsevier.

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2003), 'Do financial variables help forecasting inflation and real activity in the euro area?', *Journal of Monetary Economics* **50**, 1243–1255.

Gamerman, D. (1997), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Chapman & Hall.

Gelman, A. (1996), Inference and monitoring convergence, *in* W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds, 'Markov Chain Monte Carlo in Practice', Chapman and Hall, chapter 8, pp. 131–143.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003), *Bayesian Data Analysis*, 2 edn, Chapman and Hall/CRC.

Gelman, A. and Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical Science* **7**, 457–511. with discussion.

George, E. I. and McCulloch, R. E. (1993), 'Variable selection via gibbs sampling', *Journal of the American Statistical Association* **88**, 881–889.

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

Geweke, J. (1988), 'Antithetic acceleration of monte carlo integration in bayesian inference', *Journal of Econometrics* **38**, 73–89.

Geweke, J. (1989), 'Bayesian inference in econometric models using monte carlo integration', *Econometrica* **57**, 1317–1339.

Geweke, J. (1992), Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, *in* J. M. Bernardo, J. O. Berger, A. P. David and A. F. M. Smith, eds, 'Bayesian Statistics 4', Clarendon Press, pp. 169–193.

Geweke, J. (1996*a*), 'Bayesian reduced rank regression in econometrics', *Journal of Econometrics* **75**, 121–146.

Geweke, J. (1996*b*), Variable selection and model comparison in regression, *in* J. M. Bernardo, J. O. Berger, A. P. David and A. F. M. Smith, eds, 'Bayesian Statistics', Vol. 5, Oxford University Press, pp. 609–620.

Geweke, J. (1999), 'Using simulation methods for bayesian econometric models: Inference, development and communication', *Econometric Reviews* **18**, 1–126. with discussion.

Geweke, J. (2005), *Contemporary Bayesian Econometrics and Statistics*, Wiley-Interscience.

Geweke, J. and Whiteman, C. H. (2006), Bayesian forecasting, *in* Elliott et al. (2006), chapter 1, pp. 3–80.

Giannone, D., Lenza, M. and Primiceri, G. E. (2012), Prior selection for vector autoregressions, Working Papers ECARES ECARES 2012-002, ULB – Universite Libre de Bruxelles.

Giordini, P., Pitt, M. and Kohn, R. (2011), Bayesian inference for time series state space models, *in* Clements and Hendry (2011), chapter 3, pp. 61–124.

Gupta, R. and Kabundi, A. (2010), 'Forecasting macroeconomic variables in a small open economy: A comparison between small- and large-scale model', *Journal of Forecasting* **29**, 168–186.

Harvey, A. C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

Highfield, R. A. (1987), Forecasting with Bayesian State Space Models, PhD thesis, Graduate School of Business, University of Chicago.

Jarociński, M. (2010), 'Conditional forecasts and uncertainty about forecast revisions in vector autoregressions', *Economics Letters* **108**(3), 257 – 259.

Jarociński, M. and Maćkowiak, B. (2011), Choice of variables in vector autoregressions. Manuscript.

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

Kadiyala, K. R. and Karlsson, S. (1993), 'Forecasting with generalized bayesian vector autoregressions', *Journal of Forecasting* **12**, 365–378.

Kadiyala, K. R. and Karlsson, S. (1997), 'Numerical methods for estimation and inference in bayesian var-models', *Journal of Applied Econometrics* **12**, 99–132.

Karlsson, S. (2012), Conditional posteriors for the reduced rank regression model, Working Papers 2012:11, Örebro University Business School.

Kim, C. and Nelson, C. R. (1999), *State Space Models with Regime Switching*, MIT Press.

Kim, S., Shephard, N. and Chib, S. (1998), 'Stochastic volatility: Likelihood inference and comparison with arch models', *The Review of Economic Studies* **65**(3), 361–393.

Kleibergen, F. and van Dijk, H. K. (1994), 'On the shape of the likelihood/posterior in cointegration models', *Econometric Theory* **1**, 514–551.

Kloek, T. and van Dijk, H. K. (1978), 'Bayesian estimates of equation system parameters: An application of integration by monte carlo', *Econometrica* **46**, 1–19.

Koop, G. (2003), *Bayesian Econometrics*, John Wiley & Sons, Chichester.

Koop, G. (2010), Forecasting with medium and large bayesian vars, Technical Report WP 10-43, The Rimini Centre for Economic Analysis.

Koop, G. and Korobilis, D. (2009), 'Bayesian multivariate time series methods for empirical macroeconomics', *Foundations and Trends in Econometrics* **3**, 267–358.

Koop, G., León-González, R. and Strachan, R. W. (2010), 'Efficient posterior simulation for cointegrated models with priors on the cointegration space', *Econometric Reviews* **29**, 224–242.

Koop, G. and Potter, S. (2007), 'Estimation and forecasting in models with multiple breaks', *Review of Economic Studies* **74**, 763–789.

Koop, G., Strachan, R. W., van Dijk, H. K. and Villani, M. (2006), Bayesian approaches to cointegration, *in* T. C. Mills and P. K., eds, 'The Palgrave Handbook of Theoretical Econometrics', Vol. 1, Palgrave McMillan, chapter 25.

Korobilis, D. (2008), Forecasting in vector autoregressions with many predictors, *in* Chib et al. (2008), pp. 403–431.

Korobilis, D. (forthcoming*a*), 'Hierarchical shrinkage priors for dynamic regressions with many predictors', *International Journal of Forecasting* .

Korobilis, D. (forthcoming*b*), 'Var forecasting using bayesian variable selection', *Journal of Applied Econometrics* .

Litterman, R. B. (1979), Techniques of forecasting using vector autoregressions, Working Paper 115, Federal Reserve Bank of Minneapolis.

Litterman, R. B. (1980), A bayesian procedure for forecasting with vector autoregressions, mimeo, Massachusetts Institute of Technology.

Litterman, R. B. (1986), 'Forecasting with bayesian vector autoregressions - five years of experience', *Journal of Business & Economic Statistics* **4**, 25–38.

Lütkepohl, H. (2006), Forecasting with VARMA models, *in* Elliott et al. (2006), chapter 6, pp. 287–325.

Madigan, D. and York, J. (1995), 'Bayesian graphical models for discrete data', *International Statistical Review* **63**, 215–232.

McNees, S. K. (1986), 'Forecasting accuracy of alternative techniques: A comparison of u.s. macroeconomic forecasts', *Journal of Business & Economic Statistics* **4**, 5–15.

Newey, W. K. and West, K. D. (1987), 'A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix', *Econometrica* **55**, 703–708.

Österholm, P. (2008*a*), 'Can forecasting performance be improved by considering the steady state? an application to swedish inflation and interest rate', *Journal of Forecasting* **27**(1), 41–51.

Österholm, P. (2008*b*), 'A structural bayesian var for model-based fan charts', *Applied Economics* **40**(12), 1557–1569.

Pesaran, M. H., Petenuzzo, D. and Timmermann, A. (2006), 'Forecasting time series subject to multiple structural breaks', *Review of Economic Studies* **73**, 1057–1084.

Peters, G. W., Kannan, B., Lassock, B. and Mellen, C. (2010), 'Model selection and adaptive markov chain monte carlo for bayesian cointegrated var-models', *Bayesian Analysis* **5**, 465–492.

Primiceri, G. E. (2005), 'Time varying structural vector autoregressions and monetary policy', *The Review of Economic Studies* **72**(3), 821–852.

Robert, C. P. and Casella, G. (1999), *Monte Carlo Statistical Methods*, Springer Verlag.

Robertson, J. C., Tallman, E. W. and Whiteman, C. H. (2005), 'Forecasting using relative entropy', *Journal of Money, Credit and Banking* **37**(3), 383–401.

Rothenberg, T. J. (1971), 'Identification in parametric models', *Econometrica* **39**, 577–599.

Rubio-Ramirez, J. F., Waggoner, D. F. and Zha, T. (2010), 'Structural vector autoregressions: Theory of identification and algorithms for inference', *The Review of Economic Studies* **77**, 665–696.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**, 461–464.

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

Sims, C. A. (1993), A nine-variable probabalistic macroeconomic forecasting model, *in* J. H. Stock and M. W. Watson, eds, 'Business Cycles, Indicators and Forecasting', University of Chicago Press, pp. 179–204.

Sims, C. A. and Zha, T. (1998), 'Bayesian methods for dynamic multivariate models', *International Econom Review* **39**, 949–968.

Smith, W. B. and Hocking, R. R. (1972), 'Algorithm as 53: Wishart variate generator', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **21**, 341–345.

Stock, J. H. and Watson, M. W. (2002), 'Macroeconomic forecasting using diffusion indexes', *Journal of Business & Economic Statistics* **20**, 147–162.

Stock, J. H. and Watson, M. W. (2006), Forecasting with many predictors, *in* Elliott et al. (2006), chapter 10.

Strachan, R. (2003), 'Valid bayesian estimation of the cointegrating error correction model', *Journal of Business & Economic Statistics* **21**(1), 185–95.

Strachan, R. and Inder, B. (2004), 'Bayesian analysis of the error correction model', *Journal of Econometrics* **123**(2), 307–325.

Sugita, K. (2002), Testing for cointegration rank using bayes factors, Warwick Economic Research Papers 654, University of Warwick.

Sugita, K. (2009), 'A monte carlo comparison of bayesian testing for cointegration rank', *Economics Bulletin* **29**(3), 2145–2151.

Theil, H. and Goldberger, A. S. (1960), 'On pure and mixed statistical estimation in economics', *International Economic Review* **2**, 65–78.

Tierny, L. (1994), 'Markov chains for exploring posterior distributions', *The Annals of Statistics* **22**, 1701–1762. with discussion.

Timmermann, A. (2006), Forecast combinations, *in* Elliott et al. (2006), chapter 4.

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), pp. 614–618.

Villani, M. (2000), Aspects of Bayesian Cointegration, PhD thesis, Stockholm University.

Villani, M. (2001), 'Bayesian prediction with cointegrated vector autoregressions', *International Journal of Forecasting* **17**, 585–605.

Villani, M. (2005), 'Bayesian reference analysis of cointegration', *Econom̃tric Theory* **21**, 326–357.

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

Waggoner, D. F. and Zha, T. (1999), 'Conditional forecasts in dynamic multivariate models', *The Review of Economics and Statistics* **81**, 639–651.

Waggoner, D. F. and Zha, T. (2003*a*), 'A Gibbs sampler for structural vector autoregressions', *Journal of Economic Dynamics & Control* **28**, 349–366.

Waggoner, D. F. and Zha, T. (2003*b*), 'Likelihood preserving normalization in multiple equation models', *Journal of Econometrics* **114**(2), 329–347.

West, M. and Harrison, P. (1997), *Bayesian Forecasting and Dynamic Models*, 2nd ed. edn, Springer.

Wright, J. H. (2010), Evaluating real-time var forecasts with an informative democratic prior, Working Papers 10-19, Federal Reserve Bank of Philadelphia.

Yu, B. and Mykland, P. (1998), 'Looking at Markov samplers through cusum path plots. a simple diagnostic idea', *Statistics and Computing* **8**, 275–286.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, John Wiley & Sons.

Zha, T. (1999), 'Block recursion and structural vector autoregressions', *Journal of Econometrics* **90**(2), 291–316.