Using VARs and TVP-VARs with Many Macroeconomic Variables

Gary Koop

University of Strathclyde

January 14, 2013

Abstract

This paper discusses the challenges faced by the empirical macroeconomist and methods for surmounting them. These challenges arise due to the fact that macroeconometric models potentially include a large number of variables and allow for time variation in parameters. These considerations lead to models which have a large number of parameters to estimate relative to the number of observations. A wide range of approaches are surveyed which aim to overcome the resulting problems. We stress the related themes of prior shrinkage, model averaging and model selection. Subsequently, we consider a particular modelling approach in detail. This involves the use of dynamic model selection methods with large TVP-VARs. A forecasting exercise involving a large US macroeconomic data set illustrates the practicality and empirical success of our approach.

Keywords: Bayesian VAR; forecasting; time-varying coefficients; state-space model

JEL Classification: C11, C52, E27, E37

Acknowledgements: The author is a Fellow of the Rimini Centre for Economic Analysis. I would like to thank the Economic and Social Research Council for financial support under Grant RES-062-23-2646. Helpful comments from conference participants at Macromodels 2012 are gratefully acknowledged. The empirical illustration in this paper is based on joint work with Dimitris Korobilis and I would like to thank him for allowing its inclusion here and for other helpful comments.

1 Introduction

Macroeconomists build econometric models for several reasons. They may be interested in estimating a theoretical model such as a dynamic stochastic general equilibrium (DSGE) model. Or they may wish to calculate impulse responses or variance decompositions. Or they may wish to forecast. All such activities require the development of a statistical model for some time series variables. A wide range of models, with well developed econometric methods, exist for such purposes. Since the pioneering work of Sims (1980), one of the most popular is the vector autoregressive (VAR) model. However, in recent years a number of challenges have arisen for the researcher working with VARs or similar models. This paper discusses these challenges and methods which can be used to overcome then.

The challenges arise since the macroeconomist often has to estimate many parameters with a relatively small amount of data. This can occur for several reasons. For instance, the macroeconomist wishing to incorporate all the information in the hundreds of macroeconomic variables collected by government statistical agencies may end up with a model a large number of parameters. The presence of structural breaks or other forms of parameter change increases the number of parameters that must be estimated. And, in many countries, the available time span of the data can be quite short.

These issues have been influential in inspiring a resurgence of Bayesian methods in empirical macroeconomics. If data information is weak relative to the number of parameters being estimated, prior information can be very useful in obtaining sensible results. This prior information can be of many sorts. In the case of DSGE models which involve parameters with a structural interpretation, the researcher often has strong prior information about them. Or prior information about such structural parameters can be based on previous estimates obtained using microeconomic data (see, e.g., Del Negro and Schorfheide, 2008). In the case of VARs or extensions of them, it is common to choose priors based on a training sample of data set aside to calibrate the prior (see, e.g., Primiceri, 2005). For VARs, the Minnesota prior is another popular conventional choice (see, e.g., Doan, Litterman and Sims, 1984). There has also been recent interest in giving Minnesota prior hyperparameters a hierarchical treatment which allows them to be estimated from the data (see, e.g., Giannone, Lenza and Primiceri, 2012). Model averaging or model selection methods can also be useful. That is, instead of working with one parameter rich model, the researcher may wish to average over several parsimonious models or select a single parsimonious model. When faced with the large model spaces that often result (i.e. when the researcher has many models under consideration), Bayesian methods have also been found attractive.

However, the use of large parameter or model spaces can lead to computational problems. That is, Bayesians typically estimate their models using computationally-intensive posterior simulation methods (e.g. Markov chain Monte Carlo, MCMC, methods are very popular). Even with a single small model, computational costs can be non-trivial. With larger models or large model spaces, the computational burden of a Bayesian analysis can become prohibitive.

These considerations motivate the present paper. We begin by explaining the challenges faced by the empirical macroeconomist and briefly outline some conventional treatments. We then argue that dynamic model averaging (DMA) or dynamic model selection (DMS) methods may be an attractive way of surmounting these challenges. DMS is dynamic in the sense that it allows for a different model to be selected at each point in time. DMA does model averaging with weights which can vary over time. We discuss how doing DMA or DMS using a fully specified Bayesian model is computationally prohibitive. However, we show how approximate methods, using so-called forgetting factors, developed in Raftery, Karny and Ettler (2010) can be used to implement DMA and DMS even in very large models and/or with large model spaces. We illustrate the success of these methods using a forecasting example adapted from Koop and Korobilis (2012).

2 Statistical Challenges in Modern Macroeconomics

2.1 Overview

The macroeconomist will often be interested in building econometric models for variables which we denote by y_t. There are many types of models that the macroeconomist might want to build (e.g. DSGE models), but we can illustrate the general nature of the challenges facing the macroeconomist in terms of a regression:

$$y_t = x_t' \beta + \varepsilon_t, \tag{1}$$

where ε_t are independent $N\left(0, \sigma^2\right)$ random variables. Much recent literature (including the present paper) has focussed on three main challenges which arise in many macroeconomic applications. These are:

1. x_t might contain a large number of explanatory variables (most of which are probably unimportant).
2. y_t might contain a large number of dependent variables.
3. β might not be constant over time.

In this section we will discuss why we call these "challenges" and offer a brief overview of existing treatments of them. But the general point is that all of these may lead to models which have a large number of parameters relative to the number of observations. This can lead to over-parameterization and over-fitting problems. The econometrician fitting an econometric model is faced with a situation where, with so many dimensions to fit, somewhere the noise in the data will be fit rather than the true pattern which exists. A common consequence of this is that the econometrician can find apparently good in-sample fit, but poor forecasting performance out of sample. Some of the patterns apparently found can be spurious and, when these patterns are used out of sample, can lead to poor forecasts. A second problem caused by large numbers of parameters relative to sample size is imprecise estimates. These considerations lead to a search for more parsimonious models. They also lead to a desire for more information to combine with weak data information to lead to more precise estimates. This information can be in the form of a Bayesian prior or classical shrinkage of coefficients.

Furthermore, the combination of these three challenges leads to a fourth challenge: computation. Apart from a few simple special cases (e.g. the homoskedastic VAR with natural conjugate or Minnesota prior) analytical formulae for posteriors and predictives are not available for the models macroeconomists work with and simulation methods are required. We will discuss this computational challenge at the end of this section.

2.2 The Challenge Posed by Many Explanatory Variables

To illustrate the problems posed by many explanatory variables, we begin with the conventional regression case (where y_t is a scalar) and let k be the number of potential explanatory variables. As an example, consider the cross-country growth regression literature (see, among many others, Sala-i-Martin et al, 2004 or Fernandez, Ley and Steel, 2001). The dependent variable is real GDP growth per capita and there are dozens of potential explanatory variables. Most of these explanatory variables are likely to be unimportant but the researcher does not know, a priori, which ones are important and which ones are not. For instance, Fernandez, Ley and Steel (2001) has 41 explanatory variables. With only a moderately large sample size of 140 countries, conventional econometric methods are unsatisfactory. That is, one might be tempted to use hypothesis testing and delete unimportant variables. However, such a strategy would involve so many hypothesis tests that problems associated with pre-testing would be enormous. With k potential explanatory variables, there are up to 2^k possible restricted regressions which use as explanatory variables a subset of the variables in x_t.

Most of the cross-country growth regression literature uses cross-sectional data sets, but in applications involving time series data it is also possible that an explanatory variable is important at some points in time, but not others. We will consider parameter change in a subsequent sub-section, but note here that in such a context it is often not possible to test whether an explanatory variable should simply be excluded or included, since the preferred model may include it at some points in time but not others.

This has led to a growing literature which avoids the use of hypothesis testing procedures and adopts other, closely related, econometric methods. These can be labelled: i) model averaging, ii) model selection and iii) prior shrinkage.

Model averaging involves working with all 2^k models and presenting results (e.g. forecasts) which are averaged across them. For the Bayesian, the weights used in this average are the marginal likelihoods, although classical model averaging methods can also be done using weights proportional to an information criterion (see, e.g., Sala-i-Martin et al, 2004). What often happens with such methods is that empirical results end up being dominated by several parsimonious models. For instance, forecasts will be based on several models with a small number of explanatory variables, with very little weight attached to forecasts from large models with many explanatory variables.

Model selection methods involve selecting a single preferred model out of the 2^k models using some criteria (e.g. the marginal likelihood). Note that the term model selection, when used in this context, is also sometimes called variable selection. Intermediate cases are also possible where the researcher initially selects a small number of models (in the spirit of model selection methods), but then does model averaging over this set.

Prior shrinkage can be done in many ways. The general idea is that, by combining prior information with weak data information, more precise inferences can be obtained and overfitting can be minimized. In the case where the researcher has many potential explanatory variables, most of which are probably unimportant (i.e. have coefficients near zero) it is common to use a prior with mean zero and thus, shrink coefficients towards zero. De Mol, Giannone and Reichlin (2008) is a prominent recent paper in this literature.

The close relation between model averaging, model shrinkage and prior shrinkage can be seen by using a mixtures of normal prior for the elements of $\beta = (\beta_1, .., \beta_k)'$:

$$\beta_i | \tau_i^2 \sim N\left(0, \sigma\tau_i^2\right).$$

A conventional prior would simply assume a particular value for τ_i^2 with small values implying a tight prior with a great deal of shrinkage towards the prior mean of zero. Large values imply a relatively noninformative prior with little shrinkage. However, a wide range of different shrinkage, selection or averaging methods arise if τ_i^2 is given different treatments using a hierarchical prior. For instance, as shown in Park and Casella (2008) an exponential mixing density:

$$\tau_i^2 | \lambda \sim Exp\left(\frac{\sigma^2\lambda^2}{2}\right)$$

leads to the Bayesian Lasso which is equivalent to prior shrinkage using a Laplace prior as in De Mol, Giannone and Reichlin (2008). Lasso stands for "least absolute shrinkage and selection operator" which, as its name suggests, can also be used for model selection.

Other alternatives (see, e.g., Chipman, George and McCulloch, 2001 for a classic survey or Frühwirth-Schnatter and Wagner, 2010, for a recent time series implementation) would add a layer to the prior hierarchy and let $\tau_i^2 \subseteq \{v_s, v_l\}$ with $\Pr\left(\tau_i^2 = v_s\right) = q_i$ and $\Pr\left(\tau_i^2 = v_l\right) = 1 - q_i$. The prior variances, v_s and v_l are set to "small" and "large" values. Exact details of implementation vary and some researchers set $v_s = 0$. The idea is that each coefficient is either shrunk towards zero (with probability q_i) or estimated in a data based fashion (with probability $1 - q_i$). Note that, again, this can be interpreted as a prior providing shrinkage, a model selection advice (i.e. if variables are excluded when q_i is estimated to be sufficiently small) or a model averaging device (i.e. if the researcher presents results from an MCMC algorithm which averages over models with $\tau_i^2 = v_s$ and $\tau_i^2 = v_l$ with probability q_i and $1 - q_i$, respectively).

The preceding discussion is only a brief illustration of the general ideas that are being used in a burgeoning field which includes many extensions and alternatives to the basic methods described here. The point worth noting here is that such methods are increasingly used with VARs. To formalize our discussion of VARs, we begin with definitions and notation.

Each equation of a VAR can be written as a regression:

$$y_{mt} = x'_{mt}\widetilde{\beta}_m + \varepsilon_{mt,}$$

with $t = 1, .., T$ observations for $m = 1, .., M$ variables. y_{mt} is the t^{th} observation on the m^{th} variable, x_{mt} is a k_m-vector containing the t^{th} observation of the vector of explanatory variables relevant for the m^{th} variable and $\widetilde{\beta}_m$ is the accompanying k_m-vector of regression coefficients. The conventional VAR involves setting $x_{mt} = \left(1, y'_{t-1}, .., y'_{t-p}\right)'$ for $m = 1, .., M$. However, by allowing for x_{mt} to vary across equations we are allowing for the possibility of a restricted VAR (i.e. it allows for some of the coefficients on the lagged dependent variables to be restricted to zero).

If we stack all equations into vectors/matrices as $y_t = (y_{1t}, .., y_{Mt})'$, $\varepsilon_t = (\varepsilon_{1t}, .., \varepsilon_{Mt})'$,

$$\beta = \begin{pmatrix} \widetilde{\beta}_1 \\ \vdots \\ \widetilde{\beta}_M \end{pmatrix},$$

$$x_t = \begin{pmatrix} x'_{1t} & 0 & \cdots & 0 \\ 0 & x'_{2t} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & x'_{Mt} \end{pmatrix},$$

we can write the VAR in the regression form given in (1) except that now y_t is a vector and ε_t is i.i.d. $N(0, \Sigma)$.

Note that, in the VAR, we are likely to have many explanatory variables in each equation. That is, each equation will have p lags of each of M dependent variables (plus, possibly, an intercept, other deterministic terms and exogenous variables). If M and/or p is at all large, the problem of having many explanatory variables (most of which are probably unimportant) that we saw with cross-country growth regressions will also arise here. Thus, the need for prior shrinkage, variable selection and/or model averaging becomes important. There is a large literature on prior shrinkage in VARs, beginning with the Minnesota prior (see, e.g., Doan, Litterman and Sims, 1984) which is still popularly used in recent papers such as Banbura, Giannone and Reichlin (2010), Carriero, Clark and Marcellino (2011) and Giannone, Lenza and Primiceri (2012), among many others. Moving slightly beyond the Minnesota prior, there are many other conventional priors used for VARs and Kadiyala and Karlsson (1997) offers a discussion of many of these.

However, just as with regressions, there is a growing literature which does model averaging or variable selection or uses hierarchical priors with VARs. Andersson and Karlsson (2009) and Ding and Karlsson (2012) are two recent examples in the model averaging and model selection vein of the literature. An important pioneering paper in the use of hierarchical methods involving mixture of normals priors is George, Sun and Ni (2008). Subsequently, there have been a wide range of VAR papers adopting approaches which are similar in spirit, but differ in details. Recent examples include Gefang (2012), Koop (2012) and Korobilis (2012, 2013), but many others exist. What papers like these have in common is that they use hierarchical priors to obtain more parsimonious VARs and can be used for prior shrinkage, variable selection or model averaging. Details are not provided here, but the interested reader is referred to the monograph by Koop and Korobilis (2009) for a survey of some of these methods.

2.3 The Challenge Posed by Many Dependent Variables

Conventionally, researchers have used VARs with only a few (e.g. usually two or three but at most ten) dependent variables. However, there has recently been an increasing interest in large VARs where the number of dependent variables is much larger than this. An early paper was Banbura, Giannone and Reichlin (2010) which considered VARs of over a hundred dependent variables. This interest in large VARs was motivated by the fact that macroeconomic data bases typically contain hundreds of variables. Since more variables can potentially provide the researcher with more information, it might pay to use them. After all, more information is better than less. Historically, researchers with such large data sets have used factor methods which attempt to extract the information in large numbers of variables into a much smaller number of factors (see, e.g., Forni, Hallin. Lippi and Reichlin, 2000, and Stock and Watson, 2002a,b). However, in a forecasting comparison exercise using a large set of US macroeconomic variables, Banbura et al (2009) found large VARs to forecast better than factor methods. This has led to a growing literature using large VARs in a wide variety of empirical work (see, among others, Giannone, Lenza, Momferatou and Onorante, 2010, Carriero, Kapetanios and Marcellino, 2009, Carriero, Clark and Marcellino, 2011, Koop, 2011, and Gefang, 2012).

In a sense, the challenges posed by large VARs are the same as those imposed by having a large number of explanatory variables. That is, the number of parameters relative to the number of explanatory variables can be very large with these models. Thus, some combination of prior shrinkage, variable selection or model averaging is required to obtain sensible results. For instance, Banbura et al (2010) worked with monthly data and included 12 lags in their large VAR. This lead to tens of thousands of coefficients and their data set only began in 1959. Even equation-by-equation OLS estimation of the resulting VAR is impossible since the number of explanatory variables in each equation is more than the number of observations.

In this literature, various approaches have been used. But they all fall into the same categories as described in the preceding sub-section. Some authors (e.g. Banbura et al, 2010) use conventional Minnesota priors. Other authors (e.g. Koop, 2011 or Gefang, 2012) use hierarchical priors which can be interpreted as variable selection or model averaging methods. Others (e.g. Koop, 2012) use alternative model averaging approaches which involve averaging over different smaller dimensional VARs.

2.4 The Challenge Posed by Parameter Change

Thus far, we have only discussed regressions and VARs with constant coefficients. However, there is a large amount of evidence for parameter change in most macroeconomic variables (see, among many others, Stock and Watson, 1998) suggesting that the assumption of constant coefficients may be a poor one. Furthermore, there is a large amount of evidence that error covariance matrices in VARs should not be constant (see, among many others, Sims and Zha, 2006).

There are a huge number of econometric models which allow for time variation in parameters (e.g. Markov switching, structural break models, threshold models, etc.). However, time-varying parameter (TVP) models are becoming an increasingly popular choice (e.g. Cogley and Sargent, 2001, 2005, Primiceri, 2005 and many others). The TVP-VAR extends the VAR defined in above as:

$$y_t = x_t \beta_t + \varepsilon_t, \tag{2}$$

where

$$\beta_{t+1} = \beta_t + u_t \tag{3}$$

where u_t is i.i.d. $N(0,Q)$. Thus, the VAR coefficients are allowed to vary gradually over time. Furthermore, previously we assumed ε_t to be i.i.d. $N(0,\Sigma)$ and, thus, the model was homoskedastic. In general, it may be desirable to assume ε_t to be i.i.d. $N(0,\Sigma_t)$ so as to allow for heteroskedasticity. There are many forms that Σ_t could take (e.g. various types of multivariate GARCH or regime switching behavior are possible). However, many papers have found the form of multivariate stochastic volatility used in Primiceri (2005) to be attractive. For the sake of brevity, complete details are not provided here but the interested reader can find them in the monograph, Koop and Korobilis (2009).

Allowing for parameter change greatly increases over-parameterization concerns. Estimating one high-dimensional vector β can be daunting enough, estimating $t = 1,..,T$ vectors β_t is even more daunting. The equation modelling the evolution of β_t, (3), can be interpreted as a hierarchical prior for β_t and this is a big help in reducing over-parameterization concerns. Nevertheless, the TVP-VAR contains more parameters than the VAR (i.e. the VAR can be interpreted as a restricted special case of the TVP-VAR with $Q = 0$). Any concerns about over-parameterization in a VAR will become even greater in the comparable TVP-VAR.

Most of the existing TVP-VAR literature (e.g. Cogley and Sargent, 2001, 2005 and Primiceri, 2005) works with small dimensional models and use of three dependent variables is common. With models of this dimension, over-parameterization concerns are less and the need for prior shrinkage, variable selection or model averaging is less. Thus, the existing literature typically works with a single model and no model selection or averaging is done. Fairly tight priors are used, though, particularly on Q. Q controls the evolution of the VAR coefficients and, without a tight prior specifying that Q is small, it is possible to over-fit and find too much change in the VAR coefficients. In this case, β_t can wander widely over time, including into counter-intuitive and empirically damaging explosive regions of the parameter space.

In the subsequent sections, we will discuss new methods, based on Koop and Korobilis (2012), for shrinkage, model selection and model averaging in larger TVP-VARs and illustrate their success in an empirical example. However, before doing this, we discuss the computational challenge of VARs and TVP-VARs (and model averaging) as this motivates the need for the approximate approach used with these new methods.

2.5 The Computational Challenge

Even in the regression model with which we began this section, doing model averaging can be computationally demanding. In the beginning of Section 2.2, we showed how model averaging can involve working with 2^k models. Even with the Normal linear regression model with natural conjugate prior (for which analytical posterior and predictive results exist), exhaustively evaluating each and every model can take days or weeks (or more) of computer time unless $k < 25$ or so. This has led to the use of simulation algorithms (e.g. the Markov Chain Monte Carlo Model Composition or MC^3 algorithm of Madigan and York,1995). Methods for doing model selection or averaging using hierarchical priors such as those discussed in Section 2.2 (e.g. George, Ni and Sun, 2008 or Bayesian Lasso methods of Park and Casella, 2008) also require the use of MCMC methods.

With homoskedastic VARs, the use of a natural conjugate or a Minnesota prior leads to an analytical posterior and predictive density and MCMC methods are not required. However, as discussed in Kadiyala and Karlsson (1997), these priors can have some unattractive properties. Freeing up these priors to avoid these properties requires the use of posterior simulation methods. Restricted VARs, which do not have exactly the same set of lagged dependent variables in each equation, also require the use of MCMC methods. Freeing up the homoskedasticity assumption to allow for errors which exhibit multivariate stochastic volatility or GARCH effects also leads to MCMC methods. In sum, MCMC methods are required to do Bayesian inference in VARs as soon as we move beyond the simplest setup in more empirically-realistic directions.

TVP-VARs are state space models and one of their advantages is that well-known statistical methods for state space models (e.g. based on the Kalman filter) are available. But, full Bayesian analysis still involves MCMC methods. Exact details of the MCMC methods required for many of these VARs and TVP-VARs is provided in Koop and Korobilis (2009) and the website associated with this monograph provides computer code for implementation. With small models, such MCMC methods work very well, but they often become very demanding in high-dimensional models.

To provide a rough idea of approximate computer time, consider the three variable TVP-

VAR of Primiceri (2005). Taking 10,000 MCMC draws (which may not be enough to ensure convergence of the algorithm) takes approximately 1 hour on a good quality personal computer. This is fine if the researcher wishes to estimate one small model. But doing a recursive forecasting exercise may involve repeatedly doing MCMC methods on an expanding window of data. If the researcher forecasts at 100 points in time, the computation burden then becomes about 100 hours. Furthermore, this only involves one small model. The researcher doing model averaging or selection may have many models and computation time will increase linearly with the number of models. Finally, these computation times are all for a small model. Increasing the dimension of the model size vastly increases computation. Remember that the standard TVP-VAR with M dependent variables defined in Section 2.2 implies β_t will contain $M \times (Mp + 1)$ elements. Thus, the dimension of β_t (and thus computation time) does not increase linearly in M, but at a much faster rate.

In our experience, doing MCMC estimation and forecasting with TVP-VARs is computationally infeasible unless the researcher is working with a small number of low dimensional models. With VARs which require use of MCMC methods, it is possible to work with somewhat higher dimensional models (e.g. Koop, 2011, used the methods of George, Sun and Ni, 2008 in a VAR forecasting exercise using 20 variables), but the computational burden still seems insurmountable with larger VARs.[1] For this reason, when we develop our Bayesian methods for forecasting using model averaging with TVP-VARs an approximate approach which does not require the use of MCMC methods will be used.

3 Surmounting these Challenges in a Potentially Large TVP-VAR

3.1 Overview

In the preceding section, we described the challenges that must be overcome for the macroeconomist to build a sensible econometric model such as a TVP-VAR. Such a model will po-

[1] There are some potentially promising computational developments in particle filtering and sequential importance sampling (possibly involving massively parallel computation) which may allow for estimation of larger VARs and TVP-VARs without use of MCMC methods in the future. But such methods are in their infancy. See, for instance, Andrieu, Doucet and Holenstein (2010) and Durham and Geweke (2012).

tentially contain many variables. However, we emphasize the word “potentially” here. We are not necessarily saying that a large TVP-VAR will always be necessary. It is possible that a smaller, more parsimonious model, is sufficient. Furthermore, it is possible that time variation in parameters is not required and that a constant coefficient VAR is sufficient. However, we do not want to simply impose such choices. We want an econometric methodology which chooses automatically the dimension of the VAR and the degree of coefficient change. This is what we discuss in this half of the paper. Based on Koop and Korobilis (2012), we describe an econometric methodology for estimating potentially large TVP-VARs and automatically making many of the necessary specification choices such as the dimensionality of the TVP-VAR.

To do this, we first describe an approximate method involving the use of a forgetting factor which allows for computationally simple inference in large TVP-VARs without the need for MCMC methods. Next, we describe how over-parameterization can be avoided through the use of DMA or DMS. In the preceding section we discussed model averaging and model selection methods. DMA and DMS share the same general motivation as these methods, but are dynamic. That is, they allow for changes over time in how the model averaging or variable selection is done. For instance, DMS allows for the possibility of switching between a small and large TVP-VAR so that a more parsimonious representation is used much of the time, with the large TVP-VAR only used occasionally when needed.

3.2 Easy Estimation of TVP-VARs using Forgetting Factors

In this sub-section we will describe the main ideas involved in estimating TVP-VARs using forgetting factors. Complete details are available in Koop and Korobilis (2012).

The TVP-VAR is defined in (2) and (3) with ε_t being i.i.d. $N(0, \Sigma_t)$ and u_t being i.i.d. $N(0, Q_t)$. The basic idea of our algorithm is that, if Σ_t and Q_t were known, then computation vastly simplified and MCMC methods would not be required. That is, the TVP-VAR is a state space model and standard algorithms can be used for estimating the states (in our case β_t for $t = 1, .., T$). A full Bayesian approach such as that used by Primiceri (2005) would involve the use of MCMC methods for drawing Σ_t and Q_t and, at each draw, state space algorithms could be used to provide draws of β_t. Since many MCMC draws are necessary, this can be computationally very demanding. In the past, before the explosion of computer power after the 1980s, it was common to use various approximations to avoid such a computational burden. We adapt one of these methods (see Raftery, Karny and Ettler, 2010, for a discussion of the method used in this paper or West and Harrison, 1997, for a general overview of such methods).

To describe the basics of forgetting factor methods, we first introduce notation where $y^s = (y_1, .., y_s)'$ denotes observations through time s. The Kalman filter is a standard tool for estimating state space models such as TVP-VAR. Kalman filtering involves an updating formula

$$\beta_{t-1}|y^{t-1} \sim N\left(\beta_{t-1|t-1}, V_{t-1|t-1}\right) \tag{4}$$

where formulae for $\beta_{t-1|t-1}$ and $V_{t-1|t-1}$ are given in textbook sources (e.g. Durbin and Koopman, 2001).

Kalman filtering then proceeds using a prediction equation:

$$\beta_t|y^{t-1} \sim N\left(\beta_{t|t-1}, V_{t|t-1}\right) \tag{5}$$

where

$$V_{t|t-1} = V_{t-1|t-1} + Q_t \tag{6}$$

Kalman filtering requires the initialization of the Kalman filter, i.e. requires a prior density for $\beta_0|y^0$ which involves choice of $\beta_{0|0}$ and $V_{0|0}$. Then it proceeds by sequentially using the formulae for $\beta_1|y^0$, $\beta_1|y^1, \beta_2|y^1$, etc. from (5) and (4). The Kalman filter also provides a predictive density $p\left(y_t|y^{t-1}\right)$ which can be used for forecasting y_t given data through time $t-1$.

Equation (6) is the only place where Q_t appears in the Kalman filtering formulae. It can be removed if we replace (6) by:

$$V_{t|t-1} = \frac{1}{\lambda} V_{t-1|t-1}.$$

The scalar λ is called a forgetting factor with $0 < \lambda \leq 1$. The desirable properties of forgetting factor estimates are discussed in papers such as Raftery, Karny and Ettler (2010). For instance, it can be shown that they imply that observations j periods in the past have weight λ^j in the estimation of β_t. λ can be set to a number slightly less than one. For quarterly macroeconomic data, $\lambda = 0.99$ implies observations five years ago receive approximately 80% as much weight as last period's observation. This is a value suitable for fairly stable models where coefficient change is gradual. The choice $\lambda = 1$ yields the constant coefficient VAR.

To estimate Σ_t we use a similar approach involving a decay factor, κ. In particular, we use an Exponentially Weighted Moving Average (EWMA) estimator:

$$\widehat{\Sigma}_t = \kappa \widehat{\Sigma}_{t-1} + (1-\kappa)\widehat{\varepsilon}_t \widehat{\varepsilon}_t', \tag{7}$$

where $\widehat{\varepsilon}_t = y_t - \beta_{t|t} x_t$ is produced by the Kalman filter.

Using these approximations for Σ_t and Q_t along with Kalman filtering methods for inference on β_t for $t = 1, .., T$ results in fast and easy computation of even large TVP-VARs without the need for MCMC methods.

As noted above, Kalman filtering requires a prior for the initial condition. In our empirical work, we use a version of the Minnesota prior with

$$\beta_0 \sim N\left(\beta_{0|0}, V_{0|0}\right).$$

This Minnesota prior shrinks coefficients towards a prior mean (in our case $\beta_{0|0}$ is zero) with the prior variance, $V_{0|0}$, controlling the amount of shrinkage. If $\underline{V}_i$ contains prior variances for coefficients in equation i, then we make the Minnesota prior choices of

$$\underline{V}_i = \begin{cases} \frac{\gamma}{r^2} \text{ for coefficients on lag } r \text{ for } r = 1, .., p \\ \underline{a} \text{ for the intercepts} \end{cases}.$$

We set $\underline{a}$ set to large number so as to be noninformative. We will return to the issue of the selection of γ in the empirical section.

3.3 Model Averaging and Model Switching in TVP-VARs

Through the use of forgetting factors we have surmounted the computational challenge associated with large TVP-VARs. But TVP-VARs can still be over-parameterized. The use of the Minnesota prior on the initial condition is some help in overcoming this problem, but we go further by using model averaging and selection methods. In this sub-section we describe how these are implemented.

So far have discussed one single TVP-VAR. With many TVP regression models, Raftery et al (2010) develop methods for doing DMA or DMS which involves the use of an additional forgetting factor. The basic idea can be described quite simply. Suppose the researcher has $j = 1, .., J$ models. DMA or DMS involve calculating the probability that model j should be used for forecasting at time t, given information through time $t-1$. We denote this probability by $\pi_{t|t-1,j}$.

Once we have calculated $\pi_{t|t-1,j}$ for $j = 1, .., J$, we can use these probabilities for forecasting y_t given information available at time $t-1$ as follows: DMS involves forecasting using the single model with the highest value for $\pi_{t|t-1,j}$. DMA involves forecasting using a weighed average of forecasts from all J models where the weights are given by $\pi_{t|t-1,j}$. We stress that these weights are changing over time, so that DMS allows for model switching (i.e. using a different model to forecast at different points in time) and DMA uses different weights for model averaging at each point in time.

Raftery et al (2010) develop a fast recursive algorithm, similar to the Kalman filter, using a forgetting factor approximation for obtaining $\pi_{t|t-1,j}$. Their algorithm starts with an initial condition $\pi_{0|0,j}$ (which we set so that, a priori, each model is equally likely). They then use a forgetting factor α and model prediction equation:

$$\pi_{t|t-1,j} = \frac{\pi^{\alpha}_{t-1|t-1,j}}{\sum_{l=1}^{J} \pi^{\alpha}_{t-1|t-1,l}} \tag{8}$$

and a model updating equation:

$$\pi_{t|t,j} = \frac{\pi_{t|t-1,j} p_j \left(y_t | y^{t-1}\right)}{\sum_{l=1}^{J} \pi_{t|t-1,l} p_l \left(y_t | y^{t-1}\right)} \tag{9}$$

to complete their filtering algorithm. Note that all that is required is $p_j \left(y_t | y^{t-1}\right)$, which is the predictive likelihood for model j (i.e. the predictive density for model j evaluated at y_t). This is produced by the Kalman filter run for model j. Thus, in its entirety, the DMA algorithm of Raftery et al (2010) requires the researcher to run J Kalman filters as well as the filtering algorithm given in (8) and (9). This is much faster than using MCMC methods and, thus, large model spaces can be used.

The properties of this approach are discussed in Raftery et al (2010). Here we provide some intuition for why this approach has attractive properties. The predictive likelihood is a measure of forecast performance and the forgetting factor, α, has a similar interpretation as the other forgetting factor, λ. With these facts, it can be seen that the approach implies that models that have forecast well in the recent past have higher $\pi_{t|t-1,j}$ and, thus, are more likely to be used to forecast now. That is, it is easy to show that:

$$\pi_{t|t-1,j} = \prod_{i=1}^{t-1} \left[p_j \left(y_{t-i} | y^{t-i-1}\right)\right]^{\alpha^i}$$

Thus, model j will receive more weight at time t if it has forecast well in the recent past where the interpretation of the "recent past" is controlled by the forgetting factor, α. With $\alpha = 0.99$ forecast performance five years ago receives 80% as much weight as forecast performance last period. With $\alpha = 1$ it can be shown that $\pi_{t|t-1,k}$ is proportional to the marginal likelihood using data through time $t-1$. Thus, setting $\alpha = 1$ produces conventional Bayesian model averaging done in a recursive fashion.

Belmonte and Koop (2013) provide evidence that the approach of Raftery et al (2010) yields results which are similar to those produced by doing DMA or DMS using a fully specified Bayesian model estimated using MCMC methods for an empirical example involving only a small number of models (MCMC methods are not possible with larger model spaces).

DMA and DMS using the algorithm of Raftery et al (2010) can be done with any set of models. Raftery et al (2010) use a set of TVP regression models. In the empirical application in the next section, we use it for a different set of models. For the Bayesian, different priors lead to different models. Motivated by this consideration, we use different models defined by different degrees of prior shrinkage in the Minnesota prior. That is, we consider seven different values for γ, $\left[10^{-10}, 10^{-5}, 0.001, 0.005, 0.01, 0.05, 0.1\right]$, leading to seven different degrees of prior shrinkage of the initial VAR coefficients. In this way, the algorithm can choose the degree of prior shrinkage that leads to the best forecast performance.

Our empirical application also includes TVP-VARs of different dimension in its model space. In particular, we define small, medium and large TVP-VARs. The small TVP-VAR contains the variables we want to forecast (GDP growth, inflation and interest rates). The medium contains the variables in small model plus four others suggested by the DSGE literature. The large contains the variables in the medium model plus 18 others often used to forecast inflation or output growth. The Data Appendix provides a complete listing and breakdown of all variables used in this paper. Thus, our model space is composed of a range of TVP-VARs of differerent dimension and different degree of prior shrinkage.

The incorporation of TVP-VARs of different dimension raises one issue. Remember that the predictive likelihood, $p_j\left(y_{t-i}|y^{t-i-1}\right)$, plays the key role in DMS. Our small, medium and large TVP-VARs will use different definitions of y_t and thus are not comparable. Accordingly, we use the predictive likelihood for the variables in the small TVP-VAR. These are common to all models.

4 Empirical Illustration

This empirical illustration is adapted from the working paper version of Koop and Korobilis (2012) and differs somewhat from the results in the final version of the paper due to some differences in specification and details of implementation.

4.1 Data

Our data set comprises 25 major quarterly US macroeconomic variables and runs from 1959:Q1 to 2010:Q2. Following, e.g., Stock and Watson (2008) and recommendations in Carriero, Clark and Marcellino (2011) we transform all variables to stationarity, as described in the Data Appendix. We investigate the performance of our approach in forecasting inflation, GDP and the interest rate which are the variables in our small TVP-VAR. The transformations are such that the dependent variables are the percentage change in inflation (the second log difference of CPI), GDP growth (the log difference of real GDP) and the change in the interest rate (the difference of the Fed funds rate). We also standardize all variables by subtracting off a mean and dividing by a standard deviation. We calculate this mean and standard deviation for each variable using data from 1959Q1 through 1969Q4 (i.e. data before our forecast evaluation period). For the sake of brevity, we focus on DMS instead of DMA.

4.2 Other Modelling Choices and Models for Comparison

We use a lag length of 4. This is consistent with quarterly data. Our approach, which we label TVP-VAR-DMS, requires selection of forgetting and decay factors. In this paper, we simply set them to standard values (e.g. as recommended by Raftery et al, 2010) of $\alpha = \lambda = 0.99$ and $\kappa = 0.96$. Koop and Korobilis (2012) estimate the forgetting and decay factors and the interested reader is referred that paper for details about methods for their estimation.

We compare the performance of TVP-VAR-DMS to many special cases. Unless otherwise noted, these special cases are restricted versions of TVP-VAR-DMS and, thus (where relevant) have exactly the same modelling choices, priors and select the prior shrinkage parameter in the same way. They include:

- TVP-VARs of each dimension, with no DMS being done over TVP-VAR dimension.
- VARs of each dimension, obtained by setting $\lambda = 1$.
- Homoskedastic versions of each VAR.[2]

[2] When forecasting y_t given information through $t-1$, Σ is estimated as $\frac{1}{t-1}\sum_{i=1}^{t-1}\widehat{\varepsilon}_i\widehat{\varepsilon}_i'$.

We also present random walk forecasts (labelled RW) and forecasts from a homoskedastic small VAR estimated using OLS methods (labelled Small VAR OLS).

4.3 Estimation Results

In this empirical illustration, we focus on forecasting. But it is worthwhile briefly noting two aspects relating to estimation. Figure 1 plots the value of γ, the shrinkage parameter in the Minnesota prior, selected at each point in time for TVP-VARs of different dimension. Note that, as expected, we are finding that the necessary degree of shrinkage increases as the dimension of the TVP-VAR increases.

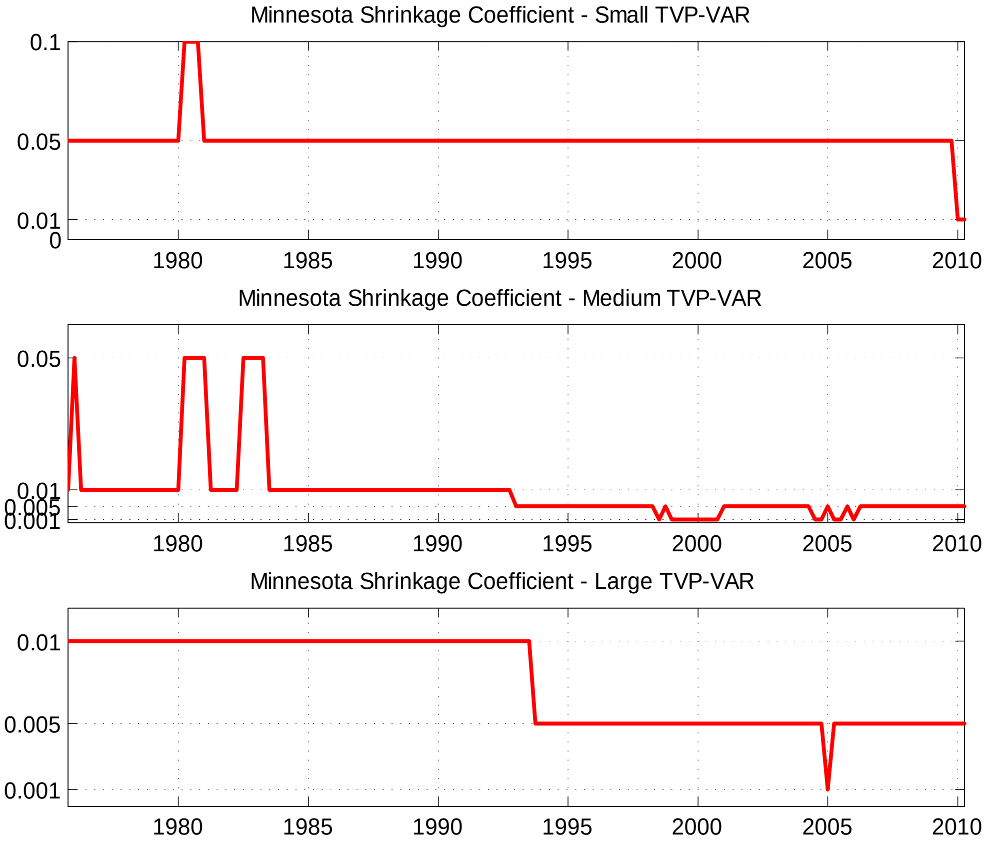

Figure 2 plots the time-varying probabilities associated with the TVP-VAR of each dimension. Note that, for each dimension of TVP-VAR, the optimum value for the Minnesota prior shrinkage parameter, γ, is chosen and the probability plotted in Figure 2 is for this optimum value. Remember that TVP-VAR-DMS will forecast with the TVP-VAR of dimension with highest probability. It can be seen that this means TVP-VAR-DMS will involve a great deal of switching between TVP-VARs of different dimension. In the relatively stable period from 1990 through 2007, the small TVP-VAR is being used to forecast. For most of the remaining time the large TVP-VAR is selected, although there are some exceptions to this (e.g. the medium TVP-VAR is selected for most of the 1967-1973 period). Thus, TVP-VAR-DMS is achieving parsimony, at least at many points in time, by avoiding the large TVP-VAR and selecting one of a smaller dimension.

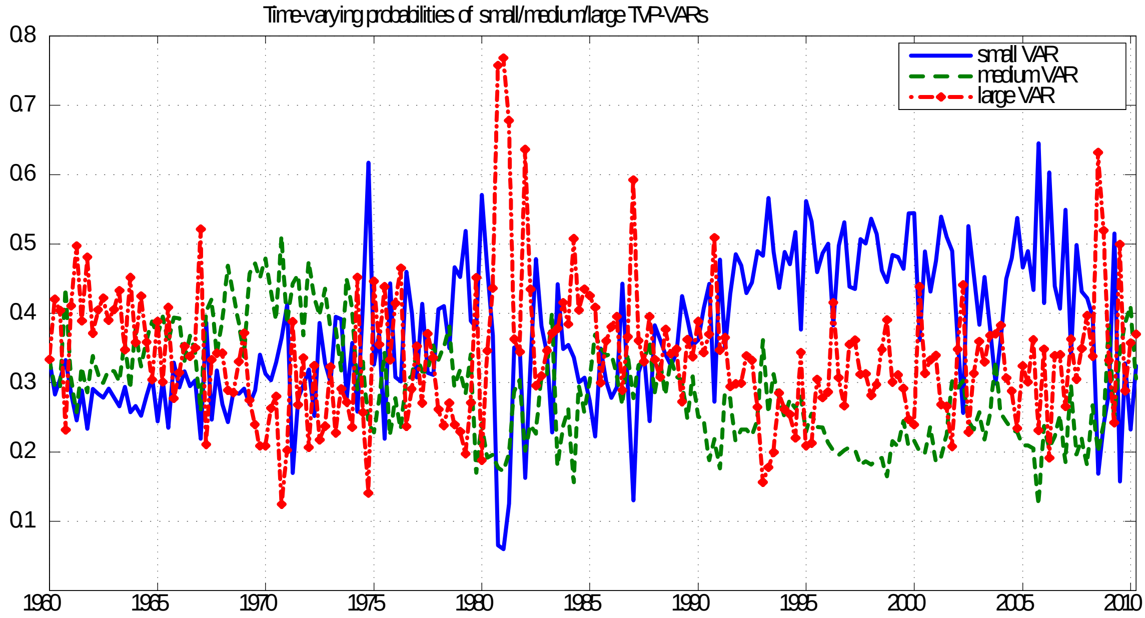

Figure 2

4.4 Forecast Comparison

We present iterated forecasts for horizons of up to two years ($h = 1, .., 8$) with a forecast evaluation period of 1970Q1 through 2010Q2. The use of iterated forecasts does increase the computational burden since predictive simulation is required (i.e. when $h > 1$ an analytical formula for the predictive density does not exist). We do predictive simulation in two different ways. The first (simpler) way uses the VAR coefficients which hold at time T to forecast variables at time $T + h$. This is labelled $\beta_{T+h} = \beta_T$ in the tables below and assumes no VAR coefficient change between T and $T + h$. The second way, labelled $\beta_{T+h} \sim RW$ in the tables, does allow for coefficient change out-of-sample and simulates from the random walk state equation (3) to produce draws of β_{T+h}. Both ways provide us with β_{T+h} and we simulate draws of $y_{\tau+h}$ conditional on β_{T+h} to approximate the predictive density.[3]

As measures of forecast performance, we use mean squared forecast errors (MSFEs) and predictive likelihoods. It is natural to use the joint predictive density for our three variables of interest (i.e. inflation, GDP and the interest rate) as an overall measure of forecast performance. Thus, Tables 1 through 3 present MSFEs for each of our three variables of interest separately. Table 4 presents sums of log predictive likelihoods using the joint predictive likelihood for these three variables.

MSFEs are presented relative to the TVP-VAR-DMS approach which simulates β_{T+h} from the random walk state equation. Tables 1 through 3 are mostly filled with numbers greater than one, indicating TVP-VAR-DMS is forecasting better than other forecasting approaches. This is particularly true for inflation and GDP. For the interest rate, TVP-VAR-DMS forecasts best at several forecast horizons but there are some forecast horizons (especially $h = 7, 8$) where large TVP-VARs are forecasting best. Nevertheless, overall MSFEs indicate TVP-VAR-DMS is the best forecasting approach among the approaches we consider. Note, too, that TVP-VAR-DMS is forecasting much better than our most simple benchmarks: random walk forecasts and forecasts from a small VAR estimated using OLS methods.

Next consider results for TVP-VARs of a fixed dimension. Overall, we are finding that large TVP-VARs tend to forecast better than small or medium ones, although there are many exceptions to this. For instance, large TVP-VARs tend to do well when forecasting interest rates and inflation, but when forecasting GDP the small TVP-VAR tends to do better. Such findings highlight that there may often be uncertainty about TVP-VAR dimensionality suggesting the usefulness of TVP-VAR-DMS. In general, though, MSFEs indicate that heteroskedastic VARs

[3] For longer-term forecasting, this has the slight drawback that our approach is based on the model updating equation which uses one-step ahead predictive likelihoods (which may not be ideal when forecasting $h > 1$ periods ahead).

tend to forecast about as well as TVP-VARs of the same dimension suggesting that, with this data set, allowing for time-variation in VAR coefficients is less important than allowing for DDS.

With regards to predictive simulation, MSFEs suggest that simulating β_{T+h} from the random walk state equation yields only modest forecast improvements over the simpler strategy of assuming no change in VAR coefficients over the horizon that the forecast is being made.

Table 1: Relative Mean Squared Forecast Errors, GDP

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
				FULL MODEL				
TVP-VAR-DMS, $\lambda = 0.99, \beta_{T+h} = \beta_T$	1.00	1.02	1.02	1.03	1.02	1.00	1.01	0.99
TVP-VAR-DMS, $\lambda = 0.99, \beta_{T+h} \sim RW$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
				SMALL VAR				
TVP-VAR, $\lambda = 0.99, \beta_{T+h} = \beta_T$	1.04	0.95	1.08	1.00	1.04	1.08	1.01	1.02
TVP-VAR, $\lambda = 0.99, \beta_{T+h} \sim RW$	1.05	0.95	1.08	1.03	1.03	1.06	0.99	1.02
VAR, heteroskedastic	1.04	0.94	1.06	1.03	1.04	1.06	1.02	1.04
VAR, homoskedastic	1.09	1.01	1.04	1.01	1.06	1.08	1.02	1.04
				MEDIUM VAR				
TVP-VAR, $\lambda = 0.99, \beta_{T+h} = \beta_T$	1.09	0.99	1.04	1.04	1.06	1.05	1.02	1.07
TVP-VAR, $\lambda = 0.99, \beta_{T+h} \sim RW$	1.10	1.00	1.04	1.07	1.06	1.05	1.03	1.05
VAR, heteroskedastic	1.10	1.00	1.02	1.05	1.09	1.02	1.02	1.10
VAR, homoskedastic	1.08	1.02	1.04	1.08	1.09	1.03	1.00	1.08
				LARGE VAR				
TVP-VAR, $\lambda = 0.99, \beta_{T+h} = \beta_T$	1.03	1.04	1.02	1.06	1.07	1.07	1.06	1.10
TVP-VAR, $\lambda = 0.99, \beta_{T+h} \sim RW$	1.02	1.05	1.03	1.06	1.06	1.08	1.07	1.09
VAR, heteroskedastic	1.09	1.12	1.08	1.11	1.09	1.10	1.10	1.13
VAR, homoskedastic	1.02	1.05	1.04	1.04	1.03	1.03	1.03	1.05
				BENCHMARK MODELS				
RW	1.59	1.71	1.81	1.97	1.96	1.88	1.96	2.22
Small VAR OLS	1.19	1.13	1.53	1.29	1.31	1.36	1.27	1.29

Table 2: Relative Mean Squared Forecast Errors, Inflation

	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$	$h = 7$	$h = 8$
				FULL MODEL				
TVP-VAR-DMS, $\lambda = 0.99, \beta_{T+h} = \beta_T$	1.02	0.99	1.00	1.00	1.00	1.01	0.99	1.00
TVP-VAR-DMS, $\lambda = 0.99, \beta_{T+h} \sim RW$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
				SMALL VAR				
TVP-VAR, $\lambda = 0.99, \beta_{T+h} = \beta_T$	1.04	1.05	1.07	1.06	1.06	1.06	1.00	1.04
TVP-VAR, $\lambda = 0.99, \beta_{T+h} \sim RW$	1.03	1.06	1.07	1.06	1.05	1.04	1.01	1.04
VAR, heteroskedastic	1.02	1.04	1.03	1.01	1.02	1.02	0.98	1.05
VAR, homoskedastic	1.05	1.08	1.05	1.02	1.02	1.03	0.98	1.06
				MEDIUM VAR				
TVP-VAR, $\lambda = 0.99, \beta_{T+h} = \beta_T$	1.08	1.06	1.07	1.01	1.00	1.04	0.99	1.05
TVP-VAR, $\lambda = 0.99, \beta_{T+h} \sim RW$	1.08	1.05	1.05	1.01	1.00	1.05	0.99	1.04
VAR, heteroskedastic	1.07	1.06	1.02	1.00	1.02	1.02	0.96	1.07
VAR, homoskedastic	1.11	1.10	1.11	1.03	1.03	1.04	0.96	1.09
				LARGE VAR				
TVP-VAR, $\lambda = 0.99, \beta_{T+h} = \beta_T$	1.01	1.02	1.02	0.95	0.99	1.04	0.97	1.04
TVP-VAR, $\lambda = 0.99, \beta_{T+h} \sim RW$	1.01	1.03	1.03	0.95	1.01	1.04	0.97	1.02
VAR, heteroskedastic	1.05	1.03	1.03	0.95	1.01	1.03	0.96	1.04
VAR, homoskedastic	1.05	1.05	1.04	0.96	1.01	1.05	0.97	1.07
				BENCHMARK MODELS				
RW	3.26	2.71	1.69	2.07	2.11	1.73	1.65	1.74
Small VAR OLS	1.09	1.23	1.12	1.14	1.16	1.05	1.02	1.18

Table 3: Relative Mean Squared Forecast Errors, Interest Rates

	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
FULL MODEL								
TVP-VAR-DMS, $\lambda = 0.99, \beta_{T+h} = \beta_T$	1.03	1.00	1.02	1.00	0.99	0.99	1.00	0.99
TVP-VAR-DMS, $\lambda = 0.99, \beta_{T+h} \sim RW$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SMALL VAR								
TVP-VAR, $\lambda = 0.99, \beta_{T+h} = \beta_T$	1.16	1.02	1.14	1.19	1.01	0.99	1.16	1.11
TVP-VAR, $\lambda = 0.99, \beta_{T+h} \sim RW$	1.16	1.00	1.16	1.20	1.02	1.01	1.14	1.11
VAR, heteroskedastic	1.19	1.00	1.12	1.09	1.00	0.96	1.05	1.01
VAR, homoskedastic	1.25	1.10	1.15	1.11	0.99	0.96	1.04	1.03
MEDIUM VAR								
TVP-VAR, $\lambda = 0.99, \beta_{T+h} = \beta_T$	1.18	1.01	1.10	1.06	0.98	0.99	0.98	0.97
TVP-VAR, $\lambda = 0.99, \beta_{T+h} \sim RW$	1.20	1.01	1.12	1.06	0.97	1.00	0.98	0.98
VAR, heteroskedastic	1.17	0.97	1.05	1.02	0.97	1.00	0.98	0.96
VAR, homoskedastic	1.25	1.06	1.11	1.03	1.00	1.01	0.96	0.98
LARGE VAR								
TVP-VAR, $\lambda = 0.99, \beta_{T+h} = \beta_T$	1.07	0.94	1.06	0.96	0.98	1.00	0.91	0.92
TVP-VAR, $\lambda = 0.99, \beta_{T+h} \sim RW$	1.05	0.94	1.05	0.97	0.98	1.00	0.92	0.91
VAR, heteroskedastic	1.07	0.95	1.06	0.97	0.99	0.99	0.92	0.91
VAR, homoskedastic	1.13	0.98	1.06	0.99	1.01	1.02	0.92	0.92
BENCHMARK MODELS								
RW	1.91	2.16	1.92	1.87	1.64	1.98	2.37	1.93
Small VAR OLS	1.76	1.47	1.59	2.11	1.78	1.69	2.23	2.03

Predictive likelihoods are presented in Table 4, relative to TVP-VAR-DMS. To be precise, the numbers in Table 4 are the sum of log predictive likelihoods for a specific model minus the sum of log predictive likelihoods for TVP-VAR-DMS. The fact that almost all of these numbers are negative supports the main story told by the MSFEs: TVP-VAR-DMS is forecasting well at most forecast horizons. At $h = 1$, TVP-VAR-DMS forecasts best by a considerable margin and at other forecast horizons it beats other TVP-VAR approaches. However, there are some important differences between predictive likelihood and MSFE results that are worth noting.

The importance of allowing for heteroskedastic errors in getting the shape of the predictive density correct is clearly shown by the poor performance of homoskedastic models in Table 4. In fact, the heteroskedastic VAR exhibits the best forecast performance at many horizons. However, the dimensionality of this best forecasting model differs across horizons. For instance, at $h = 2$ the small model forecasts best, but at $h = 3$ the medium model wins and at $h = 4$ it is the large heteroskedastic VAR. This suggests, even when the researcher is using a VAR (instead of a TVP-VAR), doing DMS over VARs of different dimension still might be a useful as a conservative forecasting device which can forecast well in a context where there is uncertainty over the dimension of the VAR.

Table 4: Relative Predictive Likelihoods, Total (all 3 variables)

	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$	$h = 7$	$h = 8$
				FULL MODEL				
TVP-VAR-DMS, $\lambda = 0.99, \beta_{T+h} = \beta_T$	0.84	0.91	2.47	4.03	4.76	3.30	6.69	4.11
TVP-VAR-DMS, $\lambda = 0.99, \beta_{T+h} \sim RW$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
				SMALL VAR				
TVP-VAR, $\lambda = 0.99, \beta_{T+h} = \beta_T$	-6.71	4.62	-3.70	-2.72	2.73	1.93	-0.32	0.68
TVP-VAR, $\lambda = 0.99, \beta_{T+h} \sim RW$	-5.95	4.84	-1.95	-2.56	2.20	-0.92	-1.04	-3.32
VAR, heteroskedastic	-6.18	6.86	-1.39	1.57	12.00	6.24	5.87	9.11
VAR, homoskedastic	-47.44	-29.97	-27.74	-22.87	-15.96	-18.50	-18.92	-15.93
				MEDIUM VAR				
TVP-VAR, $\lambda = 0.99, \beta_{T+h} = \beta_T$	-23.55	0.79	-1.58	2.84	11.31	5.85	7.69	9.27
TVP-VAR, $\lambda = 0.99, \beta_{T+h} \sim RW$	-23.22	-0.09	-3.16	-0.54	11.33	5.07	8.13	9.80
VAR, heteroskedastic	-20.89	1.08	5.07	8.39	15.12	14.02	14.79	14.52
VAR, homoskedastic	-58.28	-31.86	-29.35	-21.09	-10.14	-13.94	-7.38	-10.65
				LARGE VAR				
TVP-VAR, $\lambda = 0.99, \beta_{T+h} = \beta_T$	-18.16	-7.81	-6.85	-1.32	3.03	-3.69	1.46	8.33
TVP-VAR, $\lambda = 0.99, \beta_{T+h} \sim RW$	-16.14	-8.25	-9.70	-2.45	-0.24	-7.56	-1.48	2.93
VAR, heteroskedastic	-17.30	-1.63	-1.76	8.46	12.46	6.03	10.36	13.24
VAR, homoskedastic	-50.33	-37.35	-35.31	-28.60	-17.52	-29.13	-22.05	-20.50
				BENCHMARK MODELS				
RW	-	-	-	-	-	-	-	-
Small VAR OLS	-52.94	-40.42	-49.99	-52.48	-45.69	-36.48	-37.92	-49.35

5 Conclusion

Our forecasting exercise, which involves using a large TVP-VAR and doing DMS over prior shrinkage and TVP-VAR dimension, indicates that our methods are a feasible and attractive way of addressing the challenges of working with a parameter-rich TVP-VAR. In it, we worked with a relatively small model space. Many extensions of the model space are possible without altering the algorithm presented in this paper. For instance, the model space could be augmented with VARs as well as TVP-VARs. Or models with different lag lengths could be included. Homoskedastic versions of each model could also be included. Other extensions would require slight modifications. For instance, our TVP-VAR estimation method involves a single forgetting factor, λ, controlling the degree of evolution of VAR coefficients. Allowing for different forgetting factors for each equation or even each coefficient would also be possible (see, e.g., the component discounting approach described in West and Harrison, 1997, pages 196-198). All in all, the approaches described in this paper offer a promising groundwork for the development of many approaches to addressing the challenges of modern empirical macroeconomics.

References

Andersson, M. and Karlsson, S. (2009). “Bayesian forecast combination for VAR models,” pages 501-524 in S. Chib, W. Griffiths, G. Koop and D. Terrell (eds.), *Advances in Econometrics, Volume 23*, Emerald Group: Bingley, UK.

Andrieu, C., Doucet, A. and Holenstein, R. (2010). “Particle Markov chain Monte Carlo methods,” *Journal of the Royal Statistical Society: Series B*, 72, 269-342.

Banbura, M., Giannone, D. and Reichlin, L. (2010). “Large Bayesian Vector Auto Regressions,” *Journal of Applied Econometrics*, 25, 71-92.

Belmonte, M. and Koop, G. (2013). “Model switching and model averaging in time-varying parameter regression models,” manuscript.

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, 400-417.

Chipman, H., George, E. and McCulloch, R. (2001). “The practical implementation of Bayesian model selection,” pages 65-134 in Institute of Mathematical Statistics Lecture Notes - Monograph Series, Volume 38, edited by P. Lahiri.

Cogley, T. and Sargent, T. (2001). “Evolving post-World War II inflation dynamics,” *NBER Macroeconomic Annual*, 16, 331-373.

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

De Mol, C., Giannone, D. and Reichlin, L. (2008). “Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?” *Journal of Econometrics*, 146, 318-328.

Del Negro, M. and F. Schorfheide (2008). “Forming priors for DSGE models (and how it affects the assessment of nominal rigidities),” *Journal of Monetary Economics*, 55, 1191-1208.

Ding, S. and Karlsson, S. (2012). “Model averaging and variable selection in VAR models,” manuscript.

Doan, T., Litterman, R. and Sims, C. (1984). “Forecasting and conditional projection using realistic prior distributions," *Econometric Reviews*, 3, 1-144.

Durbin, J. and Koopman, S. (2001). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Durham, G. and Geweke, J. (2012). "Adaptive sequential posterior simulators for massively parallel computing environments," manuscript.

Fernandez, C., Ley, E. and Steel, M. (2001). "Model uncertainty in cross-country growth regressions," *Journal of Applied Econometrics,* 16, 563-576.

Forni M., Hallin M., Lippi, M. and Reichlin, L. (2000). "The generalized dynamic factor model: identification and estimation" *Review of Economics and Statistics* 82: 540–554.

Gefang, D. (2012). "Bayesian doubly adaptive elastic-net Lasso for VAR shrinkage," manuscript.

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

Giannone, D., Lenza, M., Momferatou, D. and Onorante, L. (2010). "Short-term inflation projections: a Bayesian vector autoregressive approach," ECARES working paper 2010-011, Universite Libre de Bruxelles.

Giannone, D., Lenza, M. and Primiceri, G. (2012). "Prior selection for vector autoregressions," *Centre for Economic Policy Research,* working paper 8755.

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

Koop, G. (2011). "Forecasting with medium and large Bayesian VARs," *Journal of Applied Econometrics,* first published online 2011: DOI: 10.1002/jae.1270.

Koop, G. (2012). "Forecasting with dimension switching VARs," manuscript.

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

Koop, G. and Korobilis, D. (2012). "Large time-varying parameter VARs," *Journal of Econometrics*, forthcoming.

Korobilis, D. (2012). "VAR forecasting using Bayesian variable selection," *Journal of Applied Econometrics*, forthcoming.

Korobilis, D. (2013). “Bayesian forecasting with highly correlated predictors,” *Economics Letters*, forthcoming.

Madigan, D. and York, J. (1995). ”Bayesian Graphical Models for Discrete Data,” *International Statistical Review*, 63, 215-232.

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

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

Raftery, A., Karny, M. and Ettler, P. (2010). “Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill,” *Technometrics*, 52, 52-66.

Sala-i-Martin, X., Doppelhofer, G. and Miller, R. (2004). “Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach,” *American Economic Review*, 94, 813-835.

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

Sims, C. and Zha, T. (2006). “Were there regime switches in macroeconomic policy?” *American Economic Review*, 96, 54-81.

Stock J. and Watson, M. (2002a). “Forecasting using principal components from a large number of predictors,” *Journal of the American Statistical Association*, 97, 1167-1179.

Stock J. and Watson, M. (2002b). “Macroeconomic forecasting using diffusion indexes,” *Journal of Business and Economics Statistics*, 20, 147-162.

Stock, J. and Watson, M. (2008). “Forecasting in dynamic factor models subject to structural instability,” in *The Methodology and Practice of Econometrics, A Festschrift in Honour of Professor David F. Hendry*, edited by J. Castle and N. Shephard, Oxford: Oxford University Press.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models (second edition)*. New York: Springer.

A Data Appendix

All series were downloaded from St. Louis' FRED database and cover the quarters 1959:Q1 to 2010:Q2. Some series in the database were observed only on a monthly basis and quarterly values were computed by averaging the monthly values over the quarter. All variables are transformed to be approximately stationary following Stock and Watson (2008). In particular, if $z_{i,t}$ is the original untransformed series, the transformation codes are (column Tcode below): 1 - no transformation (levels), $x_{i,t} = z_{i,t}$; 2 - first difference, $x_{i,t} = z_{i,t} - z_{i,t-1}$; 3 - second difference, $x_{i,t} = z_{i,t} - z_{i,t-2}$; 4 - logarithm, $x_{i,t} = \log z_{i,t}$; 5 - first difference of logarithm, $x_{i,t} = \ln z_{i,t} - \ln z_{i,t-1}$; 6 - second difference of logarithm, $x_{i,t} = \ln z_{i,t} - \ln z_{i,t-2}$.

Table A1: Series used in the Small VAR with $n = 3$

Series ID	Tcode	Description
GDPC96	5	Real Gross Domestic Product
CPIAUCSL	6	Consumer Price Index: All Items
FEDFUNDS	2	Effective Federal Funds Rate

Table A2: Additional series used in the Medium VAR with $n = 7$

Series ID	Tcode	Description
PMCP	1	NAPM Commodity Prices Index
BORROW	6	Borrowings of Depository Institutions from the Fed
SP500	5	S&P 500 Index
M2SL	6	M2 Money Stock

Table A3: Additional Series used in the Large VAR with $n = 25$

Series ID	Tcode	Description
PINCOME	6	Personal Income
PCECC96	5	Real Personal Consumption Expenditures
INDPRO	5	Industrial Production Index
UTL11	1	Capacity Utilization: Manufacturing
UNRATE	2	Civilian Unemployment Rate
HOUST	4	Housing Starts: Total: New Privately Owned Housing Units
PPIFCG	6	Producer Price Index: All Commodities
PCECTPI	5	Personal Consumption Expenditures: Chain-type Price Index
AHEMAN	6	Average Hourly Earnings: Manufacturing
M1SL	6	M1 Money Stock
OILPRICE	5	Spot Oil Price: West Texas Intermediate
GS10	2	10-Year Treasury Constant Maturity Rate
EXUSUK	5	U.S. / U.K Foreign Exchange Rate
GPDIC96	5	Real Gross Private Domestic Investment
PAYEMS	5	Total Nonfarm Payrolls: All Employees
PMI	1	ISM Manufacturing: PMI Composite Index
NAPMNOI	1	ISM Manufacturing: New Orders Index
OPHPBS	5	Business Sector: Output Per Hour of All Persons