University of Strathclyde

Department of Economics

Three Essays in Macroeconomic Forecasting Using Bayesian Model Selection

by

Dimitris Korompilis - Magkas

A thesis presented in fulfilment of the requirements for the degree of Doctor of Philosophy

2010

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed:..... Date:.....

.

῾Εν οἶδα ὅτι οὐδὲν οἶδα

"I know one thing, that I know nothing", (Socrates, paraphrased from Plato's Apology)

Acknowledgements

I would like to thank my advisor, Gary Koop, for all his support and guidance through my PhD years. Many thanks go to Luc Bauwens, John Geweke, Roberto Leon-Gonzalez, John Maheu, Gael Martin, Gianluca Moretti, Theodore Panagiotidis, Rodney Strachan, and participants at various conferences for stimulating discussions.

I would also like to thank Professor Rod Cross and Professor Domenico Giannone for their excellent comments on my thesis.

This thesis was supported financially by the Department of Economics, University of Strathclyde which I gratefully acknowledge.

Finally, the writing of this thesis would not have been possible without the support of my friends and family whose encouragement, patience and understanding have kept me going throughout the past three years.

Contents

Contents i

List of Figures ii

List of Tables iii

Abstract v

1 Introduction 1

1.1 General background 1

1.2 Why is (Bayesian) model selection important? 6

1.3 The contribution of this thesis 9

2 Forecasting using Dynamic Model Averaging 14

2.1 Introduction 14

2.2 Forecasting Inflation 17

2.3 Empirical Work 30

2.4 Conclusions 46

3 Forecasting with many predictors 49

3.1 Introduction 49

3.2 Methodology 52

3.3 Bayesian model selection and averaging 56

3.4 Empirical Application 63

3.5 Conclusions 75

4 Forecasting using Bayesian variable selection **77**
4.1 Introduction 77
4.2 Variable selection in vector autoregressions 80
4.3 Simulated numerical examples 89
4.4 Macroeconomic Forecasting with VARs 96
4.5 Concluding remarks 109

5 Conclusion **113**
5.1 Summary & policy implications 113
5.2 Further research 115

Bibliography **117**

Appendices **129**
A Data Appendix (Chapter 2) 129
B Technical Appendix (Chapter 3) 130
C Data Appendix (Chapter 3) 134
D Technical Appendix (Chapter 4) 139

List of Figures

2.1 Expected Number of Predictors in Each Forecasting Exercise . . 33
2.2 Posterior Probability of Inclusion of Main Predictors (CPI inflation, $h = 1$) 35
2.3 Posterior Probability of Inclusion of Main Predictors (CPI inflation, $h = 4$) 36
2.4 Posterior Probability of Inclusion of Main Predictors (CPI inflation, $h = 8$) 36

2.5 Posterior Probability of Inclusion of Main Predictors (GDP deflator inflation, $h = 1$) 37
2.6 Posterior Probability of Inclusion of Main Predictors (GDP deflator inflation, $h = 4$) 37
2.7 Posterior Probability of Inclusion of Main Predictors (GDP deflator inflation, $h = 8$) 38
2.8 Posterior Means of Coefficients on Main Predictors (CPI inflation, $h = 1$) 39

4.1 Graph of the data for UK inflation, unemployment, GDP, and interest rate. 98

List of Tables

2.1 Comparing Different Forecasting Methods: CPI inflation 43
2.2 Comparing Different Forecasting Methods: GDP Deflator inflation 44
2.3 Sensitivity Analysis: CPI inflation 46
2.4 Sensitivity Analysis: GDP Deflator inflation 47

3.1 Average Posterior Probabilities of Explanatory Variables in the 3-variable VAR 69
3.2 Average Posterior Probabilities of autoregressive lags in the 3-variable VAR 70
3.3 Forecast Comparison - relative RMSFE 74

4.1 Average of posterior of restrictions γ for $h = 1, 4, 8$ (VAR model) 105
4.2 Average of posterior of restrictions γ for $h = 1, 4, 8$ (TVP-VAR model) 106
4.3 Forecast evaluation, $h = 1$ 110

4.4 Forecast evaluation, $h = 4$ 110
4.5 Forecast evaluation, $h = 8$ 111

Abstract

This thesis explores several aspects of Bayesian model selection in time series forecasting of macroeconomic variables. The contribution is provided in three essays.

In the first essay (Chapter 2) I forecast quarterly US inflation based on the generalized Phillips curve using econometric methods which incorporate dynamic model averaging. These methods not only allow for coefficients to change over time, but also for the entire forecasting model to change over time. I find that dynamic model averaging leads to substantial forecasting improvements over simple benchmark regressions and more sophisticated approaches such as those using time varying coefficient models. I also provide evidence on which sets of predictors are relevant for forecasting in each period.

In the second essay (Chapter 3) I address the issue of improving the forecasting performance of vector autoregressions (VARs) when the set of available predictors is inconveniently large to handle with methods and diagnostics used in traditional small-scale models. First, I summarize available information from a large dataset into a considerably smaller set of variables through factors estimated using standard principal components. However, even in the case of reducing the dimension of the data the true number of factors may still be large. For that reason I introduce in my analysis simple and efficient Bayesian model selection methods. I conduct model estimation and selection of predictors automatically through a stochastic search variable selection (SSVS) algorithm which requires minimal input by the user. I apply these methods to forecast 8 main U.S. macroeconomic variables using 124 potential predictors. I find improved out of sample

fit in high dimensional specifications that would otherwise suffer from the proliferation of parameters.

Finally, in the third essay (Chapter 4) I develop methods for automatic selection of variables in forecasting Bayesian vector autoregressions (VARs) using the Gibbs sampler. In particular, I extend the algorithms of Chapter 3 and provide computationally efficient algorithms for stochastic variable selection in generic (linear and nonlinear) VARs. The performance of the proposed variable selection method is assessed in a small Monte Carlo experiment, and in forecasting four short macroeconomic series for the UK using time-varying parameters vector autoregressions (TVP-VARs). I find that restricted models consistently improve upon their unrestricted counterparts in forecasting, showing the merits of variable selection in selecting parsimonious models.

Chapter 1

Introduction

1.1 General background

What is the probability that it is going to rain tomorrow? People try to make optimal decisions under uncertainty in their everyday life. However, decisions are costly. Taking an umbrella on a day that turns out to be sunny entails the cost of carrying the umbrella. Not carrying an umbrella on a rainy day can also prove to be very frustrating. Common sense suggest that accurate predictions of future events is a very important, yet difficult task.

In macroeconomics, and economics and finance in general, the cost of wrong decisions can far exceed the cost of getting wet. Consequently, reliable forecasts of future uncertainty are very important during decision making. For instance, many decisions are based on expectations about the level of price inflation in future months or years. The scope of many Central Banks is to maintain inflation at a natural target level, and accurate

predictions about future deviations of prices from this target are important in driving the right decisions. In accordance, with these predictions investors in financial markets hedge the risk of nominal assets; firms make investment and price-setting decisions; employees and employers negotiate nominal wages; households make choices over consumption today or consumption in the future (saving). The influence of such predictions on agents' decisions is thus widespread across the economy.

Statisticians have long ago developed methods to infer relationships between observed data. Thus, based on historical evidence, economists assign probabilities to events of interest (like future inflation). In this introduction I explain why a quite natural way to assign probabilities to events is through Bayesian inference. Bayesians assign a numerical value (the degree of belief) on a specific hypothesis, prior to collecting any evidence supporting or rejecting this hypothesis. Then they collect evidence (data) that is meant to be consistent or inconsistent with a given hypothesis (model and parameters) - a procedure known as the *scientific method*. Their prior beliefs are updated from the evidence, which provides a-posteriori a probability that the hypothesis is supported by this collected evidence. This procedure can be summarized in an elementary probability theorem of statistics, called Bayes theorem:

$$p\left(H|E\right) = \frac{p\left(E|H\right)p\left(H\right)}{p\left(E\right)} \tag{1.1}$$

where $p\left(H\right)$ is our prior belief about a hypothesis H (aka *prior*), $p\left(E|H\right)$ is the conditional probability of observing an event E given that the hy-

pothesis H is true (aka *likelihood*)[1], $p(H|E)$ is essentially our prior belief updated by the support from the evidence (aka *posterior*), and $p(E)$ is the marginal probability of observing the event E under all possible hypotheses (aka *marginal likelihood*).

There are some features which are important to be discussed here. The economist will spend some time thinking of a proper model to implement forecasts, that is her hypothesis H above. She has the freedom to assign any prior belief on the hypothesis, representing the degree of belief about specific features of this hypothesis (model specification, and parameters). This proves to be a specific advantage in macroeconomic forecasting. It is the case that macroeconomists often use priors in order to shrink the parameter space of models, priors based on theory (Dynamic Stochastic General Equilibrium, DSGE, models), priors based on their own experience of using a certain class of models again and again, or priors declaring prior ignorance about the degree of support of hypothesis H (non-informative priors). Every agent in an economy (households, firms, investors, decision makers) has their own expectations about the evolution of macroeconomic figures, without having to look at models. Subsequently when doing formal statistical forecasting it is a blessing that one can incorporate information to a model using their prior beliefs.

There are other features that differentiate the Bayesian paradigm from classical statistics. The aim of this thesis is *not* to contribute to the ongoing philosophical dispute between "Bayesians" and "Frequentists", i.e. as

[1] Actually it is the case that $p(E|H)$ is not equal to the likelihood function $L(H;E)$ but *proportional* to it.

to which method is the most appropriate to use. For this, the reader is referred to the excellent exposition in Robert (2001). Rather, this thesis uses Bayesian methods due to their attractive properties from an empirical point of view. In the remainder of this introduction I explain why this is the case.

The motivation for this thesis is to built methods for selection of variables in macroeconometric forecasting models. Reducing uncertainty about the "true" model specification is of paramount interest if the researcher wants to reduce uncertainty about the final prediction. Model selection is implemented through the comparison of the marginal likelihoods $p(E)$. Assuming that we have two candidate models, M_1 and M_2, then the ratio of marginal likelihoods of each distinct model gives a probability (in the "Bayesian sense", described earlier) that model M_1 has to be preferred over model M_2. This ratio can be defined as

$$BF = \frac{p(E|M_1)}{p(E|M_2)}$$

and it is called the Bayes factor. In fact the Bayes factor obeys the *law of likelihood*: the extent to which the evidence supports one parameter value or hypothesis against another is equal to the ratio of their likelihoods. However instead of using the complete data likelihood functions, $p(E|H, M_1)$ and $p(E|H, M_2)$, the use of marginal likelihoods takes into account *model complexity*. This has the implication that more parsimonious models are selected using Bayesian methods, a result which complies with Occam's

razor[2].

As mentioned above, another aspect of Bayesian model selection is the role of the prior. Uninformative priors do not make sense in this context, since the marginal likelihood does not exist for flat prior densities $p(H)$. Some of the model selection methods described in this thesis involve well defined and informative priors. In fact, as shown in Chapter 3, model selection can be implemented solely through the prior. However, as I explain in this thesis, the use of proper priors does not mean that model selection results will be different among researchers. Default, non-informative choices exist and are discussed extensively in the next chapters. Data-based model selection approaches are easily defined, where the researcher can let the data speak. Nevertheless, there is the additional freedom that restrictions do not only have to come from the data, but the researcher is allowed to restrict the model space accordingly. Additional restrictions from economic theory, experience or external information can be incorporated as a complement to data-based model selection.

The above discussion pertains to the theoretical aspects of Bayesian methods, i.e. model selection based on a conditional probability measure (not a frequency) and the use of priors. The reader will realize in this thesis that there is an additional attractive aspect of Bayesian inference which is purely computational. Using modern posterior simulation methods - i.e. computational Monte Carlo methods to compute the posterior $p(H|E)$ when this is analytically intractable - it is possible to do model selection

[2]Occam's razor is a principle that states that "entities must not be multiplied beyond necessity" (entia non sunt multiplicanda praeter necessitatem). Subsequently among two equally performing models, the more parsimonious one is to be preferred.

over a large number of models, without having to calculate the marginal likelihood for each and every model. The problems that will be the focus of this thesis can all be cast in linear or nonlinear regression form. This implies that when using n exogenous variables for forecasting (also called *predictor variables* or simply *predictors* in this context) the number of all possible models is 2^n. Even with modern computing power, it is too difficult to enumerate each model when $n > 20$ or 30 in most contexts. Furthermore, as it will be clear in the next chapter, *variable selection* in regression is equivalent to (nested) *model selection*, so these two terms will be used interchangeably as equivalent through this thesis.

1.2 Why is (Bayesian) model selection important?

> *"[...] the significance level a has to be determined. It has become conventional to use a = .05 or .01 based on Sir Ronald Fischer's experience with relatively small agricultural experiments (on the order of 30 to 200 plots). Subsequent advice has emphasized the need to take into account the power of the test against H_1 when setting a, and to balance power and significance in some appropriate way. However a precise way of doing this is lacking, and this advice seems to boil down to a vague suggestion that a be lower for large sample sizes, a suggestion that is mostly ignored in practice."* - Raftery (1995, p.114)

The answer to this question can hopefully prove to be quite obvious. From a theoretical point of view, using the famous words of George Box "all models are false but some are useful". The task of model selection is exactly this one: to discriminate good models from bad ones that can lead to losses during decision making. More practically, and connected with the specific features of macroeconomic datasets, model selection is important for several reasons. First, sometimes data collection can be quite costly. This statement is more relevant for biological (gene expression) or financial (at the micro level) datasets, however monitoring macroeconomic variables involves a large effort from national statistical agencies. Related to this issue is the problem of measurement error and data revisions evident in macroeconomic variables, which proliferates in large dimensions (i.e. when many variables are used). Second, model selection can help preserve parsimony and save valuable degrees of freedom in small (in terms of observations) datasets. Most macroeconomic variables are observed monthly, quarterly or annually[3]. This problem can be more evident when one wants to consider the maximum possible information (variables) available, or when using traditional multivariate models (see Chapters 3 and 4). Bayesian model selection can be used when the number of predictor variables is larger than the number of observations. Therefore, model selection has been used in very demanding problems, most notably in genomics. In gene mapping studies only a tiny number of genes is assumed to have a large effect on a trait[4].

[3]The exception is some interest and exchange rates which are determined in the financial markets and hence can be available weekly, daily or intradaily.

[4]For example, Merl et al. (2010) implement model selection in an $8,509 \times 97$ di-

Nevertheless, probably the most important aspect of model selection is the improvement of forecast accuracy. In many problems, variables which are irrelevant in-sample (i.e. when estimating a model), are not expected to be relevant out-of-sample (i.e. when forecasting), and vice-versa. Of course, this simple rule is a naive (and not globally correct) simplification of the main idea: richly parametrized econometric models suffer from over-fitting the in-sample observations, which most often results in performing very poorly out-of-sample. The statistics (and especially the econometrics) literature is full of examples where simpler models usually perform much better than complex models (like exchange rate forecasting using random walks). However there are two issues that need to be taken care of when forecasting with a "best" model. The first one is that every model, no matter how well or not it fits the data, carries some information that might be useful for forecasting. Secondly, macroeconomic data are subject to structural changes, so one variable might be relevant for forecasting only in a part of the sample (so including or excluding the variable in the whole sample is a suboptimal strategy). This thesis deals with both these issues. The first issue is easily solved using what is called Bayesian model averaging (see Chapter 3). Instead of forecasting only with the model that has the maximum posterior model probability, we can use the predictive content of each model scaled by its sample model probability to obtain an average forecast. This average forecast over a range of models, takes into account model uncertainty and is optimal in a minimum mean square error sense.

mensional gene expression matrix, or equivalently, using "econometric language", 97 observations on 8,509 variables.

The second issue is also taken into account in this thesis by developing model selection in specifications with parameter instability. For that reason I use the popular state-space framework in order to specify drifitng parameters and drifting model probabilities (see Chapters 2 & 4).

1.3 The contribution of this thesis

In light of the motivation outlined above, this thesis develops model selection methods for specific forecasting problems in macroeconomics. In Chapter 2, I extend Bayesian model averaging and selection to the time-domain. Instead of allowing predictors to have a constant probability of inclusion in the "true" model over the whole sample, I use methods which define time-varying probabilities of inclusion of predictors. In the same respect, averaging across models can be implemented each time-period (quarter) using different weighting probabilities each quarter. I implement Dynamic Model Averaging and Dynamic Model Selection by allowing a regression model to have time-varying parameters. Subsequently I also use efficient approximations for estimating the time-varying parameters and the time-varying probabilities of each possible model specification which can make computations of recursive forecasts feasible. I show that these procedures are very relevant for forecasting inflation using the Phillips curve. Among 15 predictors, there is evidence that the forecasting content of different variables has changed over time. The striking result is that for short-term forecasts of US inflation (one quarter ahead), inflation expectations becomes an important predictor in the post-1984 era (i.e. after the Great Moderation).

This complies with the empirical observation that - after the monetarist experiment of the first chairmanship of Paul Volcker in 1979 to 1983 - the Fed put more focus on price stability so it was easier to anchor inflation expectations on the part of the private sector. I compare the results with benchmark models commonly used in the literature, and find that Dynamic Model Averaging and Dynamic Model Selection provide superior forecasts of inflation.

In Chapters 3 and 4, I develop model selection methods appropriate to vector autoregressive (VAR) models. VARs are multivariate econometric models which have been used extensively for many years by macroeconomists in order to analyze the effects of monetary policy and to create reliable forecasts. As I argue in these two chapters (and as similarly many other macroeconomists have done over the years), these models are heavily parametrized, and unconstrained estimation of these models can lead to erratic predictions. In these chapters I propose two powerful model selection methods based on the Gibbs sampler. A common feature of these algorithms is that model selection is implemented in one step, i.e. at the same time as estimation of the parameters. There is no need to estimate all possible 2^n VAR specifications; all that is needed is to estimate the full model, i.e. the model with all possible predictors and lags of dependent variables, and obtain probabilities of inclusion of each parameter. Additionally, as VAR models consist of multiple equations (hence the term "multivariate"), I define these algorithms in such a way that different predictors are allowed to affect different dependent variables in each equation.

In Chapter 3 I use linear VARs in order to forecast eight macroeconomic

variables, commonly monitored by Central Banks and other institutions. Despite the moderately large number of dependent variables, I consider a large number of exogenous predictor variables, which in the empirical application is of the order of 124. A practical problem that occurs in this case is that "variables with the clearest theoretical (i.e. from macroeconomic theory) justification for use as predictors often have scant empirical predictive content" (Stock and Watson, 2003). Subsequently this is a problem where data-based selection of predictions is very relevant. Nevertheless, the computational problem can be very demanding, since in the eight-variable VAR with 124 predictors, 13 lags and an intercept that I consider, there are almost 9,000 free parameters which implies that the number of possible models is approximately 2^{9000}. It is reasonable to expect also to use lagged values of these 124 exogenous predictors, increasing the number of parameters by even more. For that reason I consider a "factor-augmented VAR" specification, where the exogenous variables are replaced by a few factors estimated using principal components. However, in light of ignorance about the correct number of factors and lags of the factors, the true model might still be very demanding. Adding in my analysis efficient Bayesian model selection methods, I forecast using a maximum of 10 factors with 13 lags each. I show that the proposed stochastic search variable selection algorithm leads to parsimonious models, with favorable results in forecasting.

In Chapter 4 I extend the previous analysis by dropping the assumption of linearity in the VAR and I develop a generic method for model selection which can effectively be used in many popular nonlinear econometric models. I specifically consider the popular time-varying parameters VAR. This

is a VAR model where the coefficients change at each point in time (as in Chapter 2). Even in small traditional specifications with three variables, the number of total parameters is in the order of thousands, and over-parametrization is a major concern. Additionally, time-varying parameters models are very flexible (almost nonparametric) and can easily overfit the data, with additional negative impacts on forecasting performance. A third issue is that in these models marginal likelihood calculations are extremely hard, and computation is quite demanding, so that piecewise comparison is effectively impossible. Many recent papers using the time-varying parameter VAR fix the number of lags for convenience (references are provided in Chapter 4). These reasons call for Bayesian variable selection methods. I provide a conceptually simple algorithm for automatic model selection of parameters which may be constant or not. Again, there is no need to estimate a new model for each possible lag length; the researcher only has to select a maximum number of lags. Unlike the data- and observation-rich empirical application of Chapter 3, I test empirically the new variable selection algorithm using a four-variable VAR for quarterly data. The Bayesian model selection algorithm consistently provides improved forecasts.

Each chapter is self contained, so I provide specific motivation for each method used in each individual chapter. Additionally, each chapter contains the necessary information for the reader to understand the intuition behind the methods proposed in this thesis. All the technical details are contained in the Appendices. This is with the exception of Chapter 2 where the nature of the problem being analyzed calls for providing a few more estimation details within the main body of the chapter.

Each of the subsequent chapters has been considered for publication. As of May 2010, a revised working paper version of Chapter 2 is re-submitted to International Economic Review. This is joint work with Gary Koop. A working paper version of Chapter 3 has been published as: Korobilis, D. (2008) "Forecasting in vector autoregressions with many predictors". Advances in Econometrics, vol. 23, 403-431. A working paper version of Chapter 4 has beeen submitted to Journal of Applied Econometrics.

Chapter 2

Forecasting using Dynamic Model Averaging

2.1 Introduction

Forecasting inflation is one of the more important, but difficult, exercises in macroeconomics. Many different approaches have been suggested. Perhaps the most popular are those based on extensions of the Phillips curve. This literature is too voluminous to survey here, but a few representative and influential papers include Ang et al. (2007), Atkeson and Ohanian (2001), Groen et al. (2008), Stock and Watson (1999) and Stock and Watson (2008). The details of these papers differ, but the general framework involves a dependent variable such as inflation (or the change in inflation) and explanatory variables including lags of inflation, the unemployment rate and other predictors. Recursive, regression-based methods, have had some success. However, three issues arise when using such methods.

First, the coefficients on the predictors can change over time. It is commonly thought that the slope of the Phillips curve has changed over time. If so, the coefficients on the predictors that determine this slope will be changing. More broadly, there is a large literature in macroeconomics which documents structural breaks and other sorts of parameter change in many time series variables (see, among many others, Stock and Watson, 1996). Recursive methods are poorly designed to capture such parameter change. It is better to build models designed to capture it.

Second, the number of potential predictors can be large. For instance, Groen et al. (2008) consider ten predictors. Researchers working with factor models such as Stock and Watson (1999) typically have many more than this. The existence of so many predictors can result in a huge number of models. If the set of models is defined by whether each of m potential predictors is included or excluded, then the researcher has 2^m models. This raises substantive statistical problems for model selection strategies. In light of this, many authors have turned to Bayesian methods, either to do Bayesian model averaging (BMA) or to automate the model selection process. Examples in macroeconomics and finance include Avramov (2002), Cremers (2002) and Koop and Potter (2004). Furthermore, computational demands can become daunting when the researcher is facing 2^m models.

Third, the model relevant for forecasting can potentially change over time. For example, the set of predictors for inflation may have been different in the 1970s than now. Or some variables may predict well in recessions but not in expansions. Furthermore, papers such as Stock and Watson (2008) find that Phillips curve forecasts work well in some periods, but at

other periods simpler univariate forecasting strategies work better. Such arguments suggest that the forecasting model is changing over time. This kind of issue further complicates an already difficult econometric exercise. That is, if the researcher has 2^m models and, at each point in time, a different forecasting model may apply, then the number of combinations of models which must be estimated in order to forecast at time τ is $2^{m\tau}$. Even in relatively simple forecasting exercises, it can be computationally infeasible to forecast by simply going through all of these $2^{m\tau}$ combinations. For this reason, to our knowledge, there is no literature on forecasting inflation with many predictors where the coefficients on those predictors may change over time and where a different forecasting model might hold at each point in time. A purpose of this paper is to fill this gap.

In this paper, we consider a strategy developed by Raftery et al. (2007) which they refer to as dynamic model averaging or DMA. Their approach can also be used for dynamic model selection or DMS where a single (potentially different) model can be used as the forecasting model at each point in time. DMA or DMS seem ideally suited for the problem of forecasting inflation since they allow for the forecasting model to change over time while, at the same time, allowing for coefficients in each model to evolve over time. They involve only standard econometric methods for state space models such as the Kalman filter but (via some empirically-sensible approximations) achieve vast gains in computational efficiency so as to allow DMA and DMS to be done in real time despite the computational problem described in the preceding paragraph.

We use these methods in the context of a forecasting exercise with quar-

terly US data from 1959Q1 through 2008Q2. We use two measures of inflation and fifteen predictors and compare the forecasting performance of DMA and DMS to a wide variety of alternative forecasting procedures. DMA and DMS indicate that the set of good predictors for inflation changes substantially over time. Due to this, we find DMA and DMS to forecast very well (in terms of forecasting metrics such as log predictive likelihoods, MSFEs and MAFEs), in most cases leading to large improvements in forecast performance relative to alternative approaches.

2.2 Forecasting Inflation

Generalized Phillips curve models

Many forecasting models of inflation are based on the Phillips curve in which current inflation depends only on the unemployment rate and lags of inflation and unemployment. Authors such as Stock and Watson (1999) include additional predictors leading to the so-called generalized Phillips curve. We take as a starting point, on which all models used in this paper build, the following generalized Phillips curve:

$$y_t = \phi + x'_{t-1}\beta + \sum_{j=1}^{p} \gamma_j y_{t-j} + \varepsilon_t \tag{2.1}$$

where y_t is inflation which we define as $\ln\left(\frac{P_t}{P_{t-1}}\right)$, with P_t being a price index, and x_t a vector of predictors. This equation is relevant for forecasting at time t given information through time $t-1$. When forecasting $h > 1$ periods ahead, the direct method of forecasting can be used and y_t and ε_t

are replaced by y_{t+h-1} and ε_{t+h-1} in (2.1).

In this paper we use quarterly data. We provide results for inflation as measured by the GDP deflator and by the consumer price index (CPI). As predictors, authors such as Stock and Watson (1999) consider measures of real activity including the unemployment rate. Various other predictors (e.g. cost variables, the growth of the money supply, the slope of term structure, etc.) are suggested by economic theory. Finally, authors such as Ang et al. (2007) have found surveys of experts on their inflation expectations to be useful predictors. These considerations suggest the following list of potential predictors which we use in this paper. Precise definitions and sources are given in Appendix A.

- UNEMP: unemployment rate.
- CONS: the percentage change in real personal consumption expenditures.
- INV: the percentage change in private residential fixed investment.
- GDP: the percentage change in real GDP.
- HSTARTS: the log of housing starts (total new privately owned housing units).
- EMPLOY: the percentage change in employment (All Employees: Total Private Industries, seasonally adjusted).
- PMI: the change in the Institute of Supply Management (Manufacturing): Purchasing Manager's Composite Index.

- WAGE: the percentage change in average hourly earnings in manufacturing.
- TBILL: three month Treasury bill (secondary market) rate.
- SPREAD: the spread between the 10 year and 3 month Treasury bill rates.
- DJIA: the percentage change in the Dow Jones Industrial Average.
- MONEY: the percentage change in the money supply (M1).
- INFEXP: University of Michigan measure of inflation expectations.
- COMPRICE: the change in the commodities price index (NAPM commodities price index).
- VENDOR: the change in the NAPM vendor deliveries index.

This set of variables is a wide one reflecting the major theoretical explanations of inflation as well as variables which have found to be useful in forecasting inflation in other studies.

Time Varying Parameter Models

Research in empirical macroeconomics often uses time varying parameter (TVP) models which are estimated using state space methods such as the Kalman filter. A standard specification can be written, for $t = 1, .., T$, as

$$y_t = z_t\theta_t + \varepsilon_t \tag{2.2a}$$

$$\theta_t = \theta_t + \eta_t. \tag{2.2b}$$

In our case, y_t is inflation, $z_t = [1, x_{t-1}, y_{t-1}, \ldots, y_{t-p}]$ is an $1 \times m$ vector of predictors for inflation (including an intercept and lags of inflation), $\theta_t = \left[\phi_{t-1}, \beta_{t-1}, \gamma_{t-1}, \ldots, \gamma_{t-p}\right]$ is an $m \times 1$ vector of coefficients (states), $\varepsilon_t \overset{ind}{\sim} N(0, H_t)$ and $\eta_t \overset{ind}{\sim} N(0, Q_t)$. The errors, ε_t and η_t, are assumed to be mutually independent at all leads and lags. Examples of recent papers which use such models (or extensions thereof) in macroeconomics include Cogley and Sargent (2005), Cogley et al. (2005), Groen et al. (2008), Koop et al. (2009), Korobilis (2009a) and Primiceri (2005).

The model given by (2.2a) and (2.2b) is an attractive one that allows for empirical insights which are not available with traditional, constant coefficient models (even when the latter are estimated recursively). However, when forecasting, they have the potential drawback that the same set of explanatory variables is assumed to be relevant at all points in time. Furthermore, if the number of explanatory variables in z_t is large, such models can often over-fit in-sample and, thus, forecast poorly.

Popular extensions of (2.2a) and (2.2b) such as TVP-VARs also include the same set of explanatory variables at all times and suffer from the same problems. Even innovative extensions such as that of Groen et al. (2008) involve only a partial treatment of predictor uncertainty. In an inflation forecasting exercise, they use a model which modifies the measurement equation to be:

$$y_t = \sum_{j=1}^{m} s_j \theta_{jt} z_{jt} + \varepsilon_t,$$

where θ_{jt} and z_{jt} denote the j^{th} elements of θ_t and z_t. The key addition to their model is $s_j \in \{0, 1\}$. Details of the exact model used for s_j are provided in Groen et al. (2008). For present purposes, the important thing to note is that it allows for each predictor for inflation to either be included (if $s_j = 1$) or excluded (if $s_j = 0$), but that s_j does not vary over time. That is, this model either includes a predictor at all points in time or excludes it at all points in time. It does not allow for the set of predictors to vary over time. It is the treatment of this latter issue which is the key addition provided by DMA.

Dynamic Model Averaging

To define what we do this paper, suppose that we have a set of K models which are characterized by having different subsets of z_t as predictors. Denoting these by $z^{(k)}$ for $k = 1, .., K$, our set of models can be written as:

$$\begin{aligned} y_t &= z_t^{(k)} \theta_t^{(k)} + \varepsilon_t^{(k)} \\ \theta_t^{(k)} &= \theta_{t-1}^{(k)} + \eta_t^{(k)}, \end{aligned} \tag{2.3}$$

$\varepsilon_t^{(k)}$ is $N\left(0, H_t^{(k)}\right)$ and $\eta_t^{(k)}$ is $N\left(0, Q_t^{(k)}\right)$. Let $L_t \in \{1, 2, .., K\}$ denote which model applies at each time period, $\Theta_t = \left(\theta_t^{(1)\prime}, .., \theta_t^{(K)\prime}\right)'$ and $y^t = (y_1, .., y_t)'$. The fact that we are letting different models hold at each point in time and will do model averaging justifies the terminology "dynamic model averaging". To be precise, when forecasting time t variables using information through time $t-1$, DMA involves calculating $\Pr\left(L_t = k | y^{t-1}\right)$

for $k = 1, .., K$ and averaging forecasts across models using these probabilities. DMS involves selecting the single model with the highest value for $\Pr(L_t = k|y^{t-1})$ and using this to forecast. Details on the calculation of $\Pr(L_t = k|y^{t-1})$ will be provided below.

Specifications such as (2.3) are potentially of great interest in empirical macroeconomics since they allow for the set of predictors for inflation to change over time as well as allowing the marginal effects of the predictors to change over time. The problems with such a framework are that many of the models can have a large number of parameters (and, hence, risk being over-parameterized) and the computational burden which arises when K is large implies that estimation can take a long time (a potentially serious drawback when forecasting in real time).

To understand the source and nature of these problems, consider how the researcher might complete the model given in (2.3). Some specification for how predictors enter/leave the model in real time is required. A simple way of doing this would be through a transition matrix, P, with elements $p_{ij} = \Pr(L_t = i|L_{t-1} = j)$ for $i, j = 1, .., K$. Bayesian inference in such a model is theoretically straightforward, but will be computationally infeasible since P will typically be an enormous matrix. Consider the case where we have m potential predictors and our models are defined according to whether each is included or excluded. Then we have $K = 2^m$ and P is a $K \times K$ matrix. Unless m is very small, P will have so many parameters that inference will be very imprecise and computation very slow.[1] Thus, a full

[1]See, for instance, Chen and Liu (2000) who discuss related models and how computation time up to t typically involves mixing over K^t terms.

Bayesian approach to DMA can be quite difficult. In this paper, we use approximations suggested by Raftery et al. (2007) which have the huge advantage that standard state space methods (e.g. involving the Kalman filter) can be used, allowing for fast, real-time forecasting.

The approximations involve two parameters, λ and α, which they refer to as *forgetting factors* and fix to numbers slightly below one. To explain the role of these forgetting factors, first consider the standard state space model in (2.2a) and (2.2b). For given values of H_t and Q_t, standard filtering and smoothing results can be used to carry out recursive estimation or forecasting. That is, Kalman filtering begins with the result that

$$\theta_{t-1}|y^{t-1} \sim N\left(\widehat{\theta}_{t-1}, \Sigma_{t-1|t-1}\right) \tag{2.4}$$

where formulae for $\widehat{\theta}_{t-1}$ and $\Sigma_{t-1|t-1}$ are standard (and are provided below for the case considered in this paper). Note here only that these formulae depend on H_t and Q_t. Then Kalman filtering proceeds using:

$$\theta_t|y^{t-1} \sim N\left(\widehat{\theta}_{t-1}, \Sigma_{t|t-1}\right), \tag{2.5}$$

where

$$\Sigma_{t|t-1} = \Sigma_{t-1|t-1} + Q_t.$$

Things simplify substantially if this latter equation is replaced by:

$$\Sigma_{t|t-1} = \frac{1}{\lambda}\Sigma_{t-1|t-1} \tag{2.6}$$

or, equivalently, $Q_t = \left(1 - \lambda^{-1}\right) \Sigma_{t-1|t-1}$ where $0 < \lambda \leq 1$. Such approximations have long been used in the state space literature going back to Fagin (1964) and Jazwinsky (1970). It is a neat way to avoid the immense computational challenge of estimating the covariance matrix Q_t. Since θ_t is unobserved, estimating its covariance matrix is a challenging task for which prior information is usually not available. A different important aspect of the approximation in (2.6) is that it is related to statistical methods such as age-weighting and windowing. The name "forgetting factor" is suggested by the fact that this specification implies that observations j periods in the past have weight λ^j. An alternative way of interpreting λ is to note that it implies an effective window size of $\frac{1}{1-\lambda}$. It is common to choose a value of λ near one, suggesting a gradual evolution of coefficients. Raftery et al. (2007) set $\lambda = 0.99$. For quarterly macroeconomic data, this suggests observations five years ago receive approximately 80% as much weight as last period's observation. This is the sort of value consistent with fairly stable models where coefficient change is gradual. With $\lambda = 0.95$, observations five years ago receive only about 35% as much weight as last period's observations. This suggests substantial parameter instability where coefficient change is quite rapid. This seems to exhaust the range of reasonable values for λ and, accordingly, in our empirical work we consider $\lambda \in (0.95, 0.99)$. $\lambda = 0.99$ will be our benchmark choice and most of our empirical results will be reported for this (although we also include an analysis of the sensitivity to this choice).

An important point to note is that, with this simplification, we no longer have to estimate or simulate Q_t. Instead, all that is required (in addition to

the Kalman filter) is a method for estimating or simulating H_t (something which we will discuss below).

Forecasting in the one model case is then completed by the updating equation:

$$\theta_t|y^t \sim N\left(\widehat{\theta}_t, \Sigma_{t|t}\right), \tag{2.7}$$

where

$$\widehat{\theta}_t = \widehat{\theta}_{t-1} + \Sigma_{t|t-1}z_t\left(H_t + z_t\Sigma_{t|t-1}z_t'\right)^{-1}\left(y_t - z_t\widehat{\theta}_{t-1}\right) \tag{2.8}$$

and

$$\Sigma_{t|t} = \Sigma_{t|t-1} - \Sigma_{t|t-1}z_t\left(H_t + z_t\Sigma_{t|t-1}z_t'\right)^{-1}z_t\Sigma_{t|t-1}. \tag{2.9}$$

Recursive forecasting is done using the predictive distribution

$$y_t|y^{t-1} \sim N\left(z_t\widehat{\theta}_{t-1}, H_t + z_t\Sigma_{t|t-1}z_t'\right). \tag{2.10}$$

We stress that, conditional on H_t, these results are all analytical and, thus, no Markov chain Monte Carlo (MCMC) algorithm is required. This greatly reduces the computational burden.

The case with many models, (2.3), uses the previous approximation and an additional one. To explain this, we now switch to the notation for the multiple model case in (2.3) and let Θ_t denote the vector of all the coefficients. In the standard single model case, Kalman filtering is based on (2.4), (2.5) and (2.7). In the multi-model case, for model k, these three

equations become:

$$\Theta_{t-1}|L_{t-1}=k,y^{t-1}\sim N\left(\widehat{\theta}_{t-1}^{(k)},\Sigma_{t-1|t-1}^{(k)}\right), \tag{2.11}$$

$$\Theta_{t}|L_{t}=k,y^{t-1}\sim N\left(\widehat{\theta}_{t-1}^{(k)},\Sigma_{t|t-1}^{(k)}\right) \tag{2.12}$$

and

$$\Theta_{t}|L_{t}=k,y^{t}\sim N\left(\widehat{\theta}_{t}^{(k)},\Sigma_{t|t}^{(k)}\right), \tag{2.13}$$

where $\widehat{\theta}_{t}^{(k)},\Sigma_{t|t}^{(k)}$ and $\Sigma_{t|t-1}^{(k)}$ are obtained via Kalman filtering in the usual way using (2.8), (2.9) and (2.6),except with (k) superscripts added to denote model k. To make clear the notation in these equations, note that, conditional on $L_t=k$, the prediction and updating equations will only provide information on $\theta_t^{(k)}$ and not the full vector Θ_t. Hence, we have only written (2.11), (2.12) and (2.13) in terms of the distributions which hold for $\theta_t^{(k)}$.

The previous results were all conditional on $L_t=k$, and we need a method for unconditional prediction (i.e. not conditional on a particular model). In theory, a nice way of doing this would be through specifying a transition matrix, P, such as that given above and using MCMC methods to obtain such unconditional results. However, for the reasons discussed previously, this will typically be computationally infeasible and empirically undesirable due to the resulting proliferation of parameters. In this paper, we follow the suggestion of Raftery et al. (2007) involving a forgetting

factor for the state equation for the models, α, comparable to the forgetting factor λ used with the state equation for the parameters. The derivation of Kalman filtering ideas begins with (2.4). The analogous result, when doing DMA, is

$$p\left(\Theta_{t-1}|y^{t-1}\right) = \sum_{k=1}^{K} p\left(\theta_{t-1}^{(k)}|L_{t-1}=k, y^{t-1}\right) \Pr\left(L_{t-1}=k|y^{t-1}\right), \quad (2.14)$$

where $p\left(\theta_{t-1}^{(k)}|L_{t-1}=k, y^{t-1}\right)$ is given by (2.11). To simplify notation, let $\pi_{t|s,l} = \Pr\left(L_t = l|y^s\right)$ and thus, the final term on the right hand side of (2.14) is $\pi_{t-1|t-1,k}$.

If we were to use the unrestricted matrix of transition probabilities in P with elements p_{kl} then the model prediction equation would be:

$$\pi_{t|t-1,k} = \sum_{l=1}^{K} \pi_{t-1|t-1,l} p_{kl},$$

but Raftery et al. (2007) replace this by:

$$\pi_{t|t-1,k} = \frac{\pi_{t-1|t-1,k}^{\alpha}}{\sum_{l=1}^{K} \pi_{t-1|t-1,l}^{\alpha}}, \quad (2.15)$$

where $0 < \alpha \leq 1$ is set to a fixed value slightly less than one and is interpreted in a similar manner to λ. Raftery et al. (2007) argue that this is an empirically sensible simplification and, in particular, is a type of multiparameter power steady model used elsewhere in the literature. See also Smith and Miller (1986) who work with a similar model and argue approximations such as (2.15) are sensible and not too restrictive.

The huge advantage of using the forgetting factor α in the model prediction equation is that we do not require an MCMC algorithm to draw transitions between models nor a simulation algorithm over model space.[2] Instead, simple evaluations comparable to those of the updating equation in the Kalman filter can be done. In particular, we have a model updating equation of:

$$\pi_{t|t,k} = \frac{\pi_{t|t-1,k} p_k\left(y_t|y^{t-1}\right)}{\sum_{l=1}^{K} \pi_{t|t-1,l} p_l\left(y_t|y^{t-1}\right)}, \tag{2.16}$$

where $p_l\left(y_t|y^{t-1}\right)$ is the predictive density for model l (i.e. the Normal density in (2.10) with (l) superscripts added) evaluated at y_t.

Recursive forecasting can be done by averaging over predictive results for every model using $\pi_{t|t-1,k}$. In that case, DMA point predictions are given by:

$$E\left(y_t|y^{t-1}\right) = \sum_{k=1}^{K} \pi_{t|t-1,k} z_t^{(k)} \widehat{\theta}_{t-1}^{(k)}.$$

DMS proceeds by selecting the single model with the highest value for $\pi_{t|t-1,k}$ at each point in time and simply using it for forecasting.

To understand further how the forgetting factor α can be interpreted, note that this specification implies that the weight used in DMA which is attached to model k at time t is:

[2]Examples of simulation algorithms over model space include the Markov chain Monte Carlo model composition (MC^3) algorithm of Madigan and York (1995) or the reversible jump MCMC algorithm of Green (1995).

$$\begin{aligned}\pi_{t|t-1,k} &\propto \left[\pi_{t-1|t-2,k} p_k\left(y_{t-1}|y^{t-2}\right)\right]^{\alpha} \\ &= \prod_{i=1}^{t-1}\left[p_k\left(y_{t-i}|y^{t-i-1}\right)\right]^{\alpha^i}.\end{aligned}$$

Thus, model k will receive more weight at time t if it has forecast well in the recent past (where forecast performance is measured by the predictive density, $p_k\left(y_{t-i}|y^{t-i-1}\right)$). The interpretation of "recent past" is controlled by the forgetting factor, α and we have the same exponential decay at the rate α^i for observations i periods ago as we had associated with λ. Thus, if $\alpha = 0.99$ (our benchmark value and also the value used by Raftery et al., 2007), forecast performance five years ago receives 80% as much weight as forecast performance last period (when using quarterly data). If $\alpha = 0.95$, then forecast performance five years ago receives only about 35% as much weight. These considerations suggest that, as with λ, we focus on the interval $\alpha \in (0.95, 0.99)$.

Note also that, if $\alpha = 1$, then $\pi_{t|t-1,k}$ is simply proportional to the marginal likelihood using data through time $t-1$. This is what standard approaches to BMA would use. If we further set $\lambda = 1$, then we obtain BMA using conventional linear forecasting models with no time variation in coefficients. In our empirical work, we include BMA in our set of alternative forecasting procedures and implement this by setting $\alpha = \lambda = 1$.

We stress that, conditional on H_t, the estimation and forecasting strategy outlined above only involves evaluating formulae such as those in the Kalman filter. All the recursions above are started by choosing a prior for

$\pi_{0|0,k}$ and $\theta_0^{(k)}$ for $k = 1, .., K$.

The preceding discussion is all conditional on H_t. Raftery et al. (2007) recommend a simple plug in method where $H_t^{(k)} = H^{(k)}$ and is replaced with a consistent estimate. When forecasting inflation, however, it is likely that the error variance is changing over time. In theory, we could use a stochastic volatility or ARCH specification for $H_t^{(k)}$. However, to do this would greatly add to the computational burden. Thus, we prefer a simple plug-in approach which is a rolling version of the recursive method of Raftery et al. (2007). To be precise, let

$$\widetilde{H}_t^{(k)} = \frac{1}{t^*} \sum_{j=t-t^*+1}^{t} \left[\left(y_t - z_t^{(k)} \widehat{\theta}_{t-1}^{(k)} \right)^2 - z_t^{(k)} \Sigma_{t|t-1}^{(k)} z_t^{(k)\prime} \right].$$

Raftery et al. (2007) uses this with $t^* = t$, but to allow for more substantial change in the error variances (e.g. due to the Great Moderation of the business cycle), we set $t^* = 20$ and, thus, use a rolling estimator based on five years of data. Following Raftery et al. (2007), we can avoid the rare possibility that $\widetilde{H}_t^{(k)} < 0$, by replacing $H_t^{(k)}$ by $\widehat{H}_t^{(k)}$ where:

$$\widehat{H}_t^{(k)} = \begin{cases} \widetilde{H}_t^{(k)} \text{ if } \widetilde{H}_t^{(k)} > 0 \\ \widehat{H}_{t-1}^{(k)} \text{ otherwise} \end{cases}.$$

2.3 Empirical Work

Our empirical work is divided into three sub-sections. The first two of these sub-sections present results using DMA and DMS, implemented in our preferred way. This involves setting $\alpha = 0.99$, $\lambda = 0.99$, a noninformative prior

over the models (i.e. $\pi_{0|0,k} = \frac{1}{K}$ for $k = 1, .., K$ so that, initially, all models are equally likely) and a relatively diffuse prior on the initial conditions of the states: $\theta_0^{(k)} \sim N(0, 100)$ for $k = 1, .., K$. The first sub-section presents evidence on which variables are good for predicting inflation. The second sub-section investigates forecast performance by comparing DMA forecasts to those produced by several alternative forecasting strategies. The third sub-section presents evidence on the sensitivity of our results to the choice of the forgetting factors. We present results for short-term ($h = 1$), medium-term ($h = 4$) and long-term ($h = 8$) forecast horizons for two measures of inflation: one based on the CPI, the other based on the GDP deflator. The list of potential predictors (which specifies the transformation used on each variable) is given in sub-section 2.1 (see also Appendix A). All of our models include an intercept two lags of the dependent variable.[3]

Which Variables are Good Predictors for Inflation?

In theory, DMA has a large potential benefit over other forecasting approaches in that it allows the forecasting model to change over time. Of course, in a particular empirical application, this benefit may be small if the forecasting model does not change much over time. Accordingly, we begin by presenting evidence that, when forecasting inflation, the forecasting model is changing over time.

One striking feature of all of our empirical results is that, although we have 15 potential predictors (and, thus, tens of thousands of models),

[3]Preliminary experimentation with lag lengths up to four indicated two lags leads to the best forecast performance for both our measures of inflation.

most probability is attached to very parsimonious models with only a few predictors. If we let $Size_k$ be the number of predictors in model k (note that $Size_k$ does not include the intercept plus two lags of the dependent variable which are common to all models), then

$$E\left(Size_t\right) = \sum_{k=1}^{K} \pi_{t|t-1,k} Size_k$$

can be interpreted as the expected or average number of predictors used in DMA at time t. Figure 2.1 plots this for our six empirical exercises (i.e. two definitions of inflation and three forecast horizons).

For the short forecast horizon ($h = 1$), the shrinkage of DMA is particularly striking. It consistently includes (in an expected value sense) a single predictor for both our definitions of inflation. For GDP deflator inflation at horizons $h = 4$ and $h = 8$, slightly more predictors are included (i.e. roughly 2 predictors are included in the early 1970s, but the number of predictors increases to 3 or 4 by the end of the sample). It is only for CPI based inflation at longer horizons that DMA chooses larger numbers of predictors. For instance, for $h = 8$ the expected number of predictors gradually increases from about two in 1970 to about eight by 2000. But even this least parsimonious case (which is still very parsimonious before 1990) excludes (in an expected value sense) half of the potential predictors.

Figure 2.1 shows clear evidence that DMA will shrink forecasts and provides some evidence that the way this shrinkage is done changes over time. But it does not tell us which predictors are important and how the predictors are changing over time. It is to these issues we now turn.

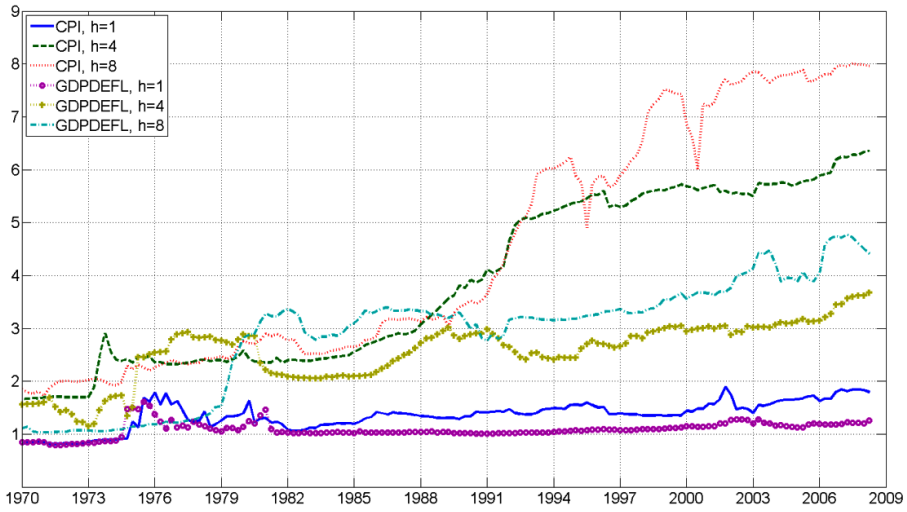

Figure 2.1: Expected Number of Predictors in Each Forecasting Exercise

Figures 2.2 through 2.7 shed light on which predictors are important at each point in time for each of our six empirical exercises. These graphs contain posterior inclusion probabilities. That is, they are the probability that a predictor is useful for forecasting at time t. Equivalently, they are the weight used by DMA attached to models which include a predictor. To keep the figures readable, we only present posterior inclusion probabilities for predictors which are important at least one point in time. To be precise, any predictor where the inclusion probability is never above 0.5 is excluded from the appropriate figure.

These figures confirm that DMS is almost always choosing parsimonious models and the weights in DMA heavily reflect parsimonious models. That is, with the partial exception of $h = 8$, it is rare for DMS to choose a model with more than two or three predictors.

Another important result is that for both measures of inflation and for

all forecast horizons, we are finding strong evidence of model change. That is, the set of predictors in the forecasting model is changing over time.

Results for CPI inflation for $h = 1$ are particularly striking. Before 1975, no predictors come through strongly. Between 1975 and 1985 money is the only predictor. After 1985 the measure of inflation expectations comes through strongly. With regards to the inflation expectations variable, similar patterns are observed for $h = 4$ and $h = 8$. Before the mid- to late- 1980s there is little or no evidence that it is a useful predictor for inflation. But after this, it often is a useful predictor. To a lesser extent, the same pattern holds with GDP deflator inflation. With $h = 1$ very few predictors are included, with money being an important predictor near the beginning of the sample and inflation expectations being important near the end. However, for GDP deflator inflation with $h = 1$, the predictor reflecting earnings (WAGE) comes through as being the strongest predictor after 1980 (this variable was not found to be an important predictor for CPI inflation).

Housing starts is another variable which often has strong predictive power for both measures of inflation. But, interestingly, only at medium or long horizons. For $h = 1$, there is no evidence at all that housing starts have predictive power for inflation.

The interested reader can examine Figures 2.2 through 2.7 for any particular variable of interest. Most of our potential explanatory variables come through as being important at some time, for some forecast horizon for some measure of inflation. Only CONS, DJIA, COMPRICE and PMI never appear in Figures 2.2 through 2.7. But it is clearly the case that

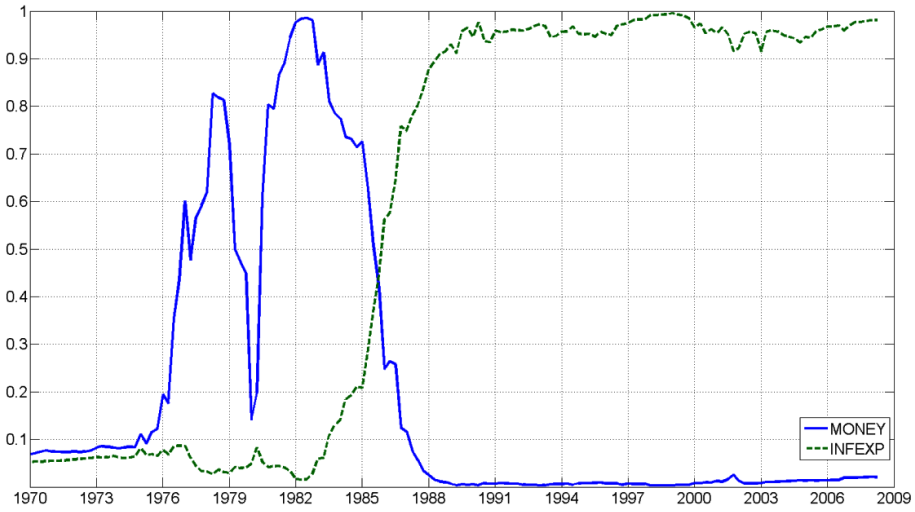

Figure 2.2: Posterior Probability of Inclusion of Main Predictors (CPI inflation, $h = 1$)

there is a large variation over time, over forecast horizons and over measures of inflation in what is a good predictor for inflation. We stress that the great benefit of DMA and DMS is that they will pick up good predictors automatically as the forecasting model evolves over time.

Figures 2.2 through 2.7 show how models evolve over time in our various empirical exercises. But they only indirectly address the issue of how the marginal effect of each predictor is changing over time. With fifteen predictors and six empirical exercises, the number of marginal effects to present is huge. Accordingly, to illustrate the kind of result that DMA is providing, we present $E(\theta_t|y^t)$ (averaged over all models) for one case. Figure 2.8 plots $E(\theta_t|y^t)$ using CPI inflation for $h = 1$ for the main predictors (i.e. the ones plotted in Figure 2.2). Consistent with Figure 2.2, it can be see that the

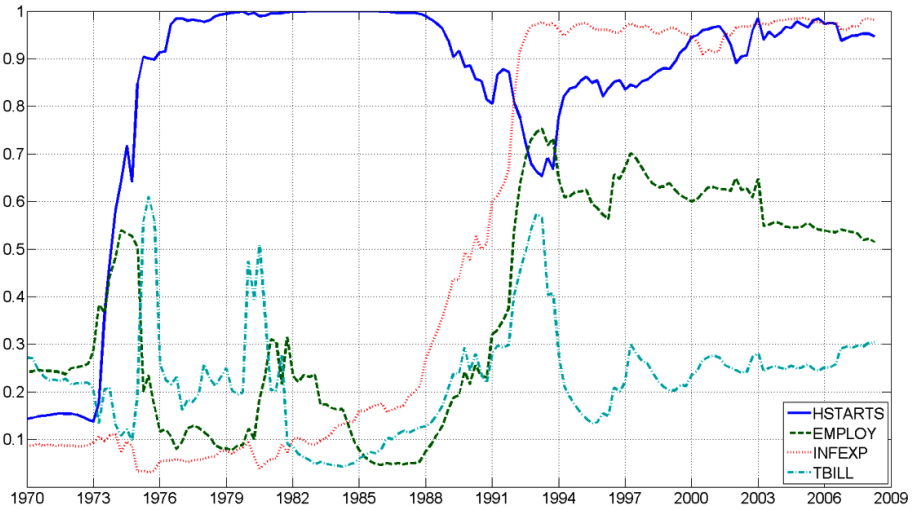

Figure 2.3: Posterior Probability of Inclusion of Main Predictors (CPI inflation, $h = 4$)

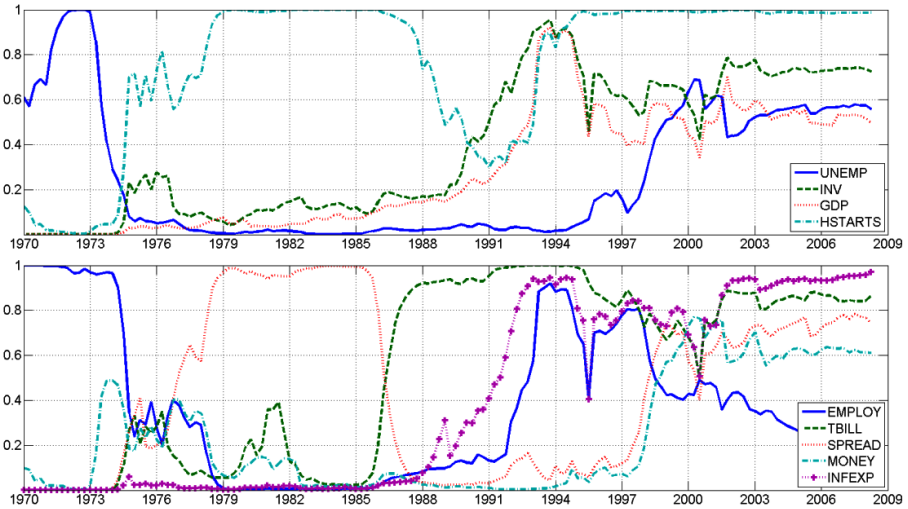

Figure 2.4: Posterior Probability of Inclusion of Main Predictors (CPI inflation, $h = 8$)

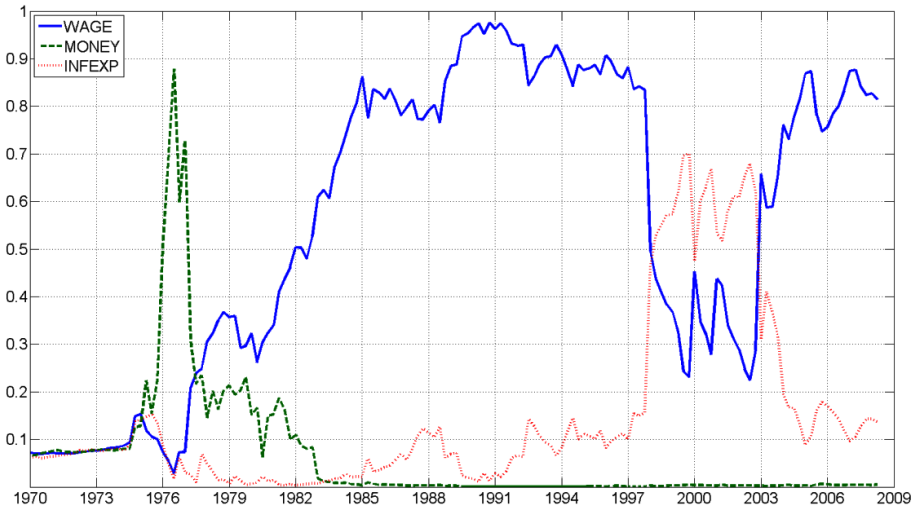

Figure 2.5: Posterior Probability of Inclusion of Main Predictors (GDP deflator inflation, $h = 1$)

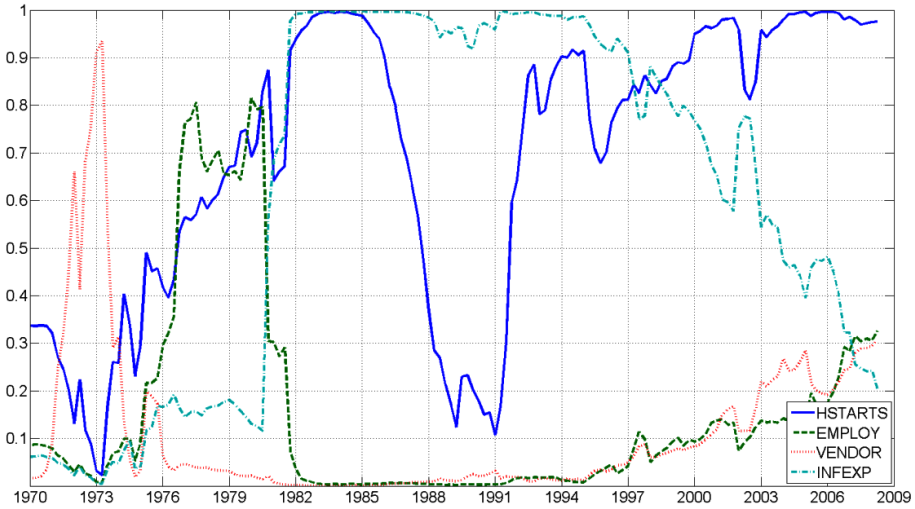

Figure 2.6: Posterior Probability of Inclusion of Main Predictors (GDP deflator inflation, $h = 4$)

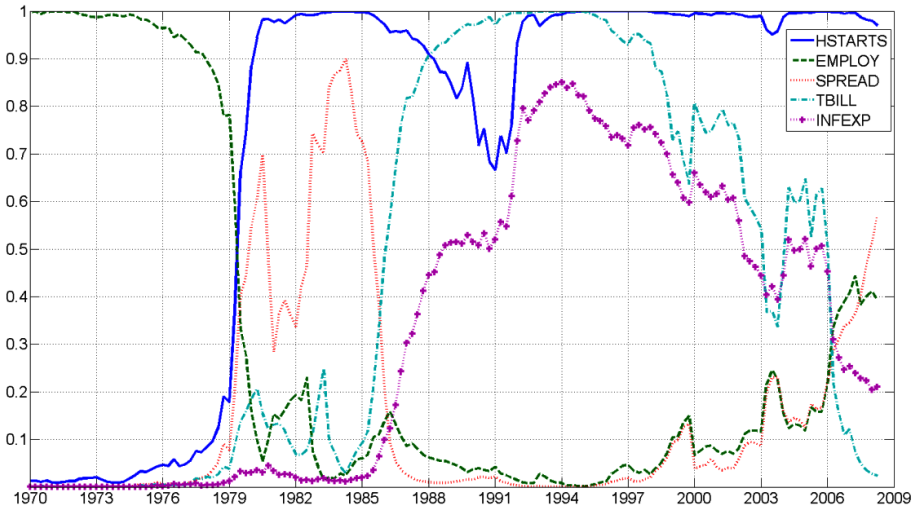

Figure 2.7: Posterior Probability of Inclusion of Main Predictors (GDP deflator inflation, $h = 8$)

marginal effect of MONEY on inflation is high until the mid to late 1980s. The marginal effect of inflation expectations becomes large only after this.

Forecast Performance: DMA versus Alternative Forecast Procedures

There are many metrics for evaluating forecast performance and many alternative forecasting methodologies that we could compare our DMA and DMS forecasts to. In this paper, we present two forecast comparison metrics involving point forecasts. These are mean squared forecast error (MSFE) and mean absolute forecast error (MAFE). We also present a forecast metric which involves the entire predictive distribution: the sum of log predictive likelihoods. Predictive likelihoods are motivated and described in many places such as Geweke and Amisano (2007). The predictive likelihood is

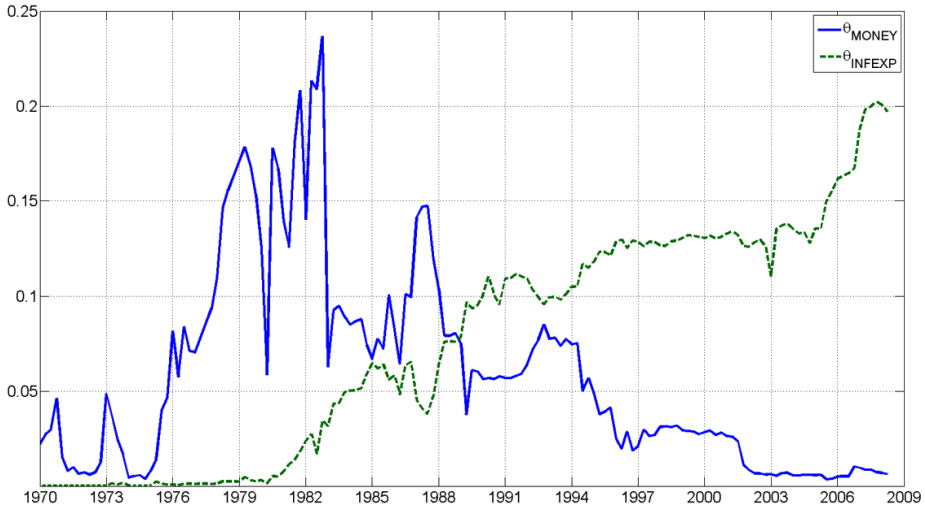

Figure 2.8: Posterior Means of Coefficients on Main Predictors (CPI inflation, $h = 1$)

the predictive density for y_t (given data through time $t-1$) evaluated at the actual outcome. The formula for the one-step ahead predictive density in model l was denoted by $p_l\left(y_t|y^{t-1}\right)$ above and can be calculated as described in Section 2.2. We use the direct method of forecasting and, hence, the log predictive density for the h-step ahead forecast is the obvious extension of this. We use the sum of log predictive likelihoods for forecast evaluation, where the sum begins in 1970Q1. MSFEs and MAFEs are also calculated beginning in 1970Q1.

In terms of alternative forecasting methods, we present results for:

- Forecasts using DMA with $\alpha = \lambda = 0.99$.
- Forecasts using DMS with $\alpha = \lambda = 0.99$.
- Forecasts using a single model containing all the predictors, but with time varying parameters (i.e. this is a special case of DMA or DMS

where 100% of the prior weight is attached to the model with all the predictors, but all other modelling choices are identical including $\lambda = 0.99$). This is labelled TVP in the tables.

- Forecasts using DMA, but where the coefficients do not vary over time in each model (i.e. this is a special case of DMA where $\lambda = 1$).
- Forecasts using BMA (i.e. this is a special case of DMA where $\lambda = \alpha = 1$).
- Recursive OLS forecasts using an AR(2) model.
- Recursive OLS forecasts using all of the predictors.
- Random walk forecasts.

The final three methods are not Bayesian, so no predictive likelihoods are presented for these cases. Note that there are many multivariate benchmark models that could be used as well. For example the time-varying parameters vector autoregression (TVP-VAR) - see also Chapter 4 - is a popular multivariate extension of the model we are using here, allowing for structural instabilities and stochastic variances and co-variances. However, this is a computationally demanding model since full MCMC methods must be used to forecast recursively. This is something which contradicts the spirit of this paper, which is to propose efficient computational methods for forecasting in the presence of structural instabilities and model uncertainty. Subsequently results of the TVP-VAR are not presented here.

Tables 1 and 2 present results for our forecasting exercise for our two different measures of inflation. The big picture story is a clear and strong

one: DMA and DMS forecast well. In most cases much better than other forecasting methods and in no case much worse than the best alternative method. We elaborate on these points below.

Consider first the log predictive likelihoods (the preferred method of Bayesian forecast comparison). These always indicate that DMA or DMS forecasts best. One message coming out of Tables 1 and 2 is that simply using a TVP model with all predictors leads to very poor forecasting performance. Of course, we are presenting results for only a single empirical exercise. But TVP models such as TVP-VARs are gaining increasing popularity in macroeconomics and the very poor forecast performance of TVP models found in Tables 1 and 2 should serve as a caution to users of such models (at least in forecasting exercises). Clearly, we are finding that the shrinkage provided by DMA or DMS is of great value in forecasting.

DMA and DMS extend conventional forecasting approaches by allowing for model evolution and parameter evolution. A message provided by the predictive likelihoods is that most of the improvements in forecasting performance found by DMA or DMS are due to their treatment of model evolution rather than parameter evolution. That is, DMA with constant coefficient models typically forecasts fairly well and occasionally even leads to the best forecast performance (see the results in Tables 1 and 2 for $h = 1$). Recently, macroeconomists have been interested in building models involving parameter change of various sorts. Our results suggest that allowing for the model to change is at least as important. At short horizons, conventional BMA forecasts fairly well, but at longer horizons it tends to forecast poorly.

Predictive likelihoods also consistently indicates that DMS forecasts a bit better than DMA (although this result does not carry over to MAFEs and MSFEs where DMA tends to do better). DMS and DMA can be interpreted as doing shrinkage in different ways. DMS puts zero weight on all models other than the one best model, thus "shrinking" the contribution of all models except one towards zero. It could be that this additional shrinkage provides some additional forecast benefits over DMA.

If we turn our attention to results using MSFE and MAFE, we can see that the previous picture still holds (although DMA does somewhat better relative to DMS than we found using predictive likelihoods). In addition, we can say that naive forecasting methods such as using an AR(2) or random walk model are clearly inferior to DMA and DMS for both measures of inflation at all forecast horizons. However, with CPI inflation, recursive OLS forecasting using all the predictors does well at the long horizon ($h = 8$). Forecasting at such a long horizon is difficult to do, so it is unclear how much weight to put on this result (and predictive likelihoods for this non-Bayesian method are not calculated). But it is worth noting that the good performance of recursive OLS in this case is not repeated for inflation measured using the GDP deflator nor at shorter horizons.

Table 2.1: Comparing Different Forecasting Methods: CPI inflation

Forecast Method	Log(PL)	MAFE	MSFE	Log(PL)	MAFE	MSFE	Log(PL)	MAFE	MSFE
Horizon	$h = 1$			$h = 4$			$h = 8$		
1970:Q1 - 1983:Q4									
DMA	−50.37	21.05	13.67	−54.65	25.15	19.50	−63.92	31.35	27.83
DMS	−47.48	21.46	14.49	−49.04	24.39	17.63	−62.22	31.23	28.49
TVP	−75.15	23.38	15.52	−79.56	34.83	31.99	−81.11	34.57	34.18
DMA ($\lambda = 1$)	−47.52	19.38	11.25	−54.21	24.68	19.44	−63.92	31.42	28.74
BMA (DMA with $\alpha = \lambda = 1$)	−48.74	19.94	12.09	−54.03	24.43	19.11	−68.03	34.06	34.18
Recursive OLS – AR(2)	–	28.93	25.23	–	41.25	45.66	–	45.31	57.72
Recursive OLS – All Preds.	–	25.62	19.61	–	27.89	23.63	–	24.49	18.85
Random Walk	–	23.38	16.73	–	40.64	40.18	–	58.70	89.94
1984:Q1 - 2008:Q2									
DMA	−34.93	26.42	12.69	−54.07	34.18	21.77	−57.08	36.20	29.48
DMS	−34.77	27.49	13.32	−54.87	34.62	23.37	−58.06	37.20	35.19
TVP	−107.19	31.31	16.67	−97.42	47.18	40.18	−72.92	38.25	28.22
DMA ($\lambda = 1$)	−34.11	25.61	11.77	−57.75	35.54	22.49	−65.84	38.02	31.63
BMA (DMA with $\alpha = \lambda = 1$)	−35.38	26.12	12.04	−71.28	41.00	28.17	−77.76	45.35	41.58
Recursive OLS – AR(2)	–	28.59	16.34	–	33.98	20.65	–	38.12	23.38
Recursive OLS – All Preds.	–	27.14	14.55	–	30.67	18.63	–	38.36	27.46
Random Walk	–	31.21	18.41	–	36.47	26.49	–	40.53	27.41
1970:Q1 - 2008:Q2									
DMA	−85.31	47.48	26.37	−108.73	59.34	41.28	−121.02	67.56	57.31
DMS	−82.26	48.96	27.82	−103.91	59.02	41.02	−120.29	68.44	63.69
TVP	−182.36	54.70	32.20	−178.30	82.28	72.24	−154.04	72.82	62.40
DMA ($\lambda = 1$)	−81.63	45.00	23.02	−111.97	60.22	41.94	−129.76	69.45	60.38
BMA (DMA with $\alpha = \lambda = 1$)	−84.12	46.07	24.14	−123.32	65.43	47.28	−145.80	79.42	75.77
Recursive OLS – AR(2)	–	57.52	41.58	–	75.00	66.17	–	83.43	81.11
Recursive OLS – All Preds.	–	52.76	34.16	–	58.36	42.08	–	62.85	46.32
Random Walk	–	54.59	35.14	–	77.30	66.75	–	99.24	117.35

Table 2.2: Comparing Different Forecasting Methods: GDP Deflator inflation

Forecast Method	Log(PL)	MAFE	MSFE	Log(PL)	MAFE	MSFE	Log(PL)	MAFE	MSFE
Horizon	$h = 1$			$h = 4$			$h = 8$		
1970:Q1 - 1983:Q4									
DMA	−42.99	19.85	8.82	−33.46	17.36	9.58	−49.52	25.01	19.52
DMS	−40.69	20.58	9.30	−28.87	17.41	9.41	−50.45	26.82	21.27
TVP	−75.19	22.05	12.62	−75.45	23.02	15.92	−78.62	28.72	20.34
DMA ($\lambda = 1$)	−39.72	17.99	8.01	−41.18	21.85	14.02	−49.97	24.42	18.34
BMA (DMA with $\alpha = \lambda = 1$)	−41.35	18.44	8.83	−46.67	23.52	15.56	−56.05	27.41	22.19
Recursive OLS – AR(2)	–	21.42	12.15	–	27.32	21.39	–	31.63	26.62
Recursive OLS – All Preds	–	19.87	9.70	–	20.12	11.63	–	23.13	17.49
Random Walk	–	18.37	9.13	–	27.14	19.16	–	36.57	37.58
1984:Q1-2008:Q2									
DMA	15.90	18.19	4.15	9.64	17.76	5.65	−10.26	22.61	9.47
DMS	15.72	18.19	4.39	9.87	18.30	6.01	−8.97	21.95	9.09
TVP	−101.70	16.79	4.37	−104.38	21.152	7.54	−105.97	29.11	13.46
DMA ($\lambda = 1$)	18.25	16.47	4.01	7.79	17.82	5.55	−12.98	23.07	9.42
BMA (DMA with $\alpha = \lambda = 1$)	16.36	17.04	4.27	0.26	19.33	6.08	−25.16	25.85	10.98
Recursive OLS – AR(2)	–	18.67	5.18	–	21.76	7.98	–	31.27	12.85
Recursive OLS – All Preds	–	17.47	4.59	–	23.38	8.53	–	30.29	14.83
Random Walk	–	19.01	6.05	–	17.21	5.42	–	22.58	8.49
1970:Q1-2008:Q2									
DMA	−27.10	34.47	12.98	−23.81	35.13	15.23	−59.79	47.73	29.01
DMS	−24.97	35.61	13.70	−18.20	35.72	15.42	−59.43	48.88	30.38
TVP	−176.90	38.85	16.99	−180.20	43.82	23.33	−184.60	57.84	33.82
DMA ($\lambda = 1$)	−21.47	33.17	12.02	−33.38	39.68	19.57	−62.95	47.60	27.78
BMA (DMA with $\alpha = \lambda = 1$)	−25.00	34.58	13.10	−46.41	42.85	21.64	−81.22	53.37	33.20
Recursive OLS – AR(2)	–	40.10	17.34	–	48.54	28.97	–	62.90	39.48
Recursive OLS – All Preds	–	37.34	14.30	–	43.09	19.92	–	53.42	32.33
Random Walk	–	37.39	15.19	–	44.28	24.57	–	59.16	46.07

Sensitivity Analysis

Our previous DMA and DMS results were for our benchmark case where $\alpha = \lambda = 0.99$. As discussed previously, researchers in this field choose pre-selected values for α and λ and the interval $(0.95, 0.99)$ is the empirically sensible one for most empirical applications. It would be possible to choose α and λ in a data-based fashion, but this is typically not done for computational reasons. For instance, the researcher could select a grid of values for these two forgetting factors and then do DMA at every possible combination of values for α and λ. Some metric (e.g. an information criteria or the sum of log predictive likelihoods through time $t-1$) could be used to select the preferred combination of α and λ at each point in time. However, this would turn an already computationally demanding exercise to one which was g^2 times as demanding (where g is the number of values in the grid). Accordingly, researchers such as Raftery et al. (2007) simply go with $\alpha = \lambda = 0.99$ and argue that results will be robust to reasonable changes in these factors. In order to investigate such robustness claims, Tables 3 and 4 present results for our forecasting exercise using different combinations of the forgetting factors.

Overall, Tables 3 and 4 reveal a high degree of robustness to choice of α and λ. If anything, these tables emphasize the benefits of DMA in that measures of forecast performance are sometimes better than those in Tables 1 and 2 and rarely much worse.

One finding of particular interest is that the combination $\alpha = 0.95$ and $\lambda = 0.99$ tends to forecast very well, for both of our measures of inflation.

Table 2.3: Sensitivity Analysis: CPI inflation

Forecast Method	log PL	MSFE	MAFE
	$h = 1$		
DMA, $\alpha = \lambda = 0.95$	-107.19	47.29	24.87
DMS, $\alpha = \lambda = 0.95$	-74.57	43.94	23.16
DMA, $\alpha = 0.99, \lambda = 0.95$	-95.39	58.30	26.61
DMS, $\alpha = 0.99, \lambda = 0.95$	-87.24	48.92	28.81
DMA, $\alpha = 0.95, \lambda = 0.99$	-91.48	45.63	22.57
DMS, $\alpha = 0.95, \lambda = 0.99$	-52.04	40.53	21.65
	$h = 4$		
DMA, $\alpha = \lambda = 0.95$	-106.25	54.26	34.31
DMS, $\alpha = \lambda = 0.95$	-57.34	46.44	26.19
DMA, $\alpha = 0.99, \lambda = 0.95$	-101.91	56.16	35.94
DMS, $\alpha = 0.99, \lambda = 0.95$	-98.64	59.34	42.43
DMA, $\alpha = 0.95, \lambda = 0.99$	-100.38	54.87	34.46
DMS, $\alpha = 0.95, \lambda = 0.99$	-61.37	47.63	26.36
	$h = 8$		
DMA, $\alpha = \lambda = 0.95$	-98.15	56.19	33.84
DMS, $\alpha = \lambda = 0.95$	-51.29	47.28	28.68
DMA, $\alpha = 0.99, \lambda = 0.95$	-111.58	64.04	51.40
DMS, $\alpha = 0.99, \lambda = 0.95$	-114.02	67.02	56.72
DMA, $\alpha = 0.95, \lambda = 0.99$	-92.93	56.48	36.06
DMS, $\alpha = 0.95, \lambda = 0.99$	-66.48	51.30	31.35

Note that the value $\alpha = 0.95$ allows for quite rapid change in forecasting model over time. This is consistent with a story we have told before: that it appears that allowing for models to change over time is more important in improving forecast performance than allowing for parameters to change (at least in our data sets).

2.4 Conclusions

This paper has investigated the use of DMA and DMS methods for forecasting US inflation. These extend conventional approaches by allowing

Table 2.4: Sensitivity Analysis: GDP Deflator inflation

Forecast Method	log PL	MSFE	MAFE
	$h = 1$		
DMA, $\alpha = \lambda = 0.95$	-66.23	36.20	14.03
DMS, $\alpha = \lambda = 0.95$	-46.81	38.12	15.96
DMA, $\alpha = 0.99, \lambda = 0.95$	-48.89	38.04	15.75
DMS, $\alpha = 0.99, \lambda = 0.95$	-48.56	38.77	16.52
DMA, $\alpha = 0.95, \lambda = 0.99$	-34.45	32.17	11.33
DMS, $\alpha = 0.95, \lambda = 0.99$	-0.11	30.76	10.84
	$h = 4$		
DMA, $\alpha = \lambda = 0.95$	-26.49	35.52	15.26
DMS, $\alpha = \lambda = 0.95$	-10.48	37.48	18.60
DMA, $\alpha = 0.99, \lambda = 0.95$	-37.50	42.46	22.79
DMS, $\alpha = 0.99, \lambda = 0.95$	-36.97	43.24	24.02
DMA, $\alpha = 0.95, \lambda = 0.99$	-13.04	32.25	13.48
DMS, $\alpha = 0.95, \lambda = 0.99$	25.36	28.87	11.81
	$h = 8$		
DMA, $\alpha = \lambda = 0.95$	-42.43	37.90	18.04
DMS, $\alpha = \lambda = 0.95$	-19.91	39.48	22.48
DMA, $\alpha = 0.99, \lambda = 0.95$	-57.95	46.96	29.40
DMS, $\alpha = 0.99, \lambda = 0.95$	-58.65	48.58	30.31
DMA, $\alpha = 0.95, \lambda = 0.99$	-36.93	38.19	18.36
DMS, $\alpha = 0.95, \lambda = 0.99$	-12.26	37.47	20.51

for the set of predictors for inflation to change over time. When you have K models and a different one can potentially hold at each of T points in time, then the resulting K^T combinations can lead to serious computational and statistical problems (regardless of whether model averaging or model selection is done). As shown in this paper, DMA and DMS handle these problems in a simple, elegant and sensible manner.

In our empirical work, we present evidence indicating the benefits of DMA and DMS. In particular, it does seem that the best predictors for forecasting inflation are changing considerably over time. By allowing for this change, DMA and DMS lead to substantial improvements in forecast

performance.

Chapter 3

Forecasting with many predictors

3.1 Introduction

It is common practice today to collect observations on many variables that potentially help explain economic variables of interest such as inflation and unemployment. Technological progress has allowed the collection, storage, and exchange of huge amounts of information without much effort and cost. In turn, this has significantly affected recent macroeconomic modeling techniques. Current academic research is focused on finding solutions on how to efficiently handle large amounts of information with, for example, Stock and Watson (2002) using 215 predictors to forecast 8 major macroeconomic variables for the U.S. economy. Bernanke and Boivin (2003), among others, argue that this is also the case nowadays in central banks, where it is customary for researchers and decision makers to monitor hundreds of

subsidiary variables during the decision-making process.

These reasons justify the current trend in applied modeling with large datasets. The modern econometrician has tools adequate enough to successfully extract information from hundreds of predictor variables and compute more accurate forecasts than ever before. It is noteworthy that these tools mainly do not rely on economic theory in an explicit way; rather they are statistical and consequently atheoretical methods that are used to cover the unfortunate gap between theoretical models and their empirical validation. Within the sum of all possible options, two methods in particular have recently gained ground: dimension reduction and model averaging. Among many others, Bernanke et al. (2005), Favero et al. (2005), Giannone et al. (2004), Stock and Watson (2002, 2005a, 2005b) and Koop and Potter (2004) show how forecasts can be improved over univariate or multivariate autoregressions, using either dynamic factors or Bayesian model averaging (BMA), or both techniques, when a rich dataset is in hand.

In this paper I examine empirically the merit of using factors extracted from a large set of explanatory variables and at the same time implementing Bayesian model averaging/selection in the context of macroeconomic vector autoregressions (VARs). While factor methods have already been examined thoroughly in multivariate models, the challenging task of model averaging/selection is implemented with a stochastic search variable selection algorithm (henceforth SSVS) proposed by George and McCulloch (1993, 1997) and George et al (2008).

The proposed approach is flexible as its output can easily be used for selection of a single best model or model averaging. The SSVS adds to a

recent and expanding literature on different approaches to BMA in VARs (Strachan and van Dijk, 2007; Andersson and Karlsson, 2008). The innovation of the specific prior formulation is that it is more appropriate for VAR models compared to previous model selection priors used in multivariate regressions (Brown et al., 1998, 2002). That is because each right-hand side variable is allowed to enter in all, some, or none of the VAR equations, and not only in all or none of them. The additional advantages come from the fact that this class of restriction search algorithms is extremely simple to use and automated. Furthermore, certain versions of these algorithms can incorporate variable selection when the number of predictors is larger than the number of time series observations.

The following section defines the Bayesian VAR model when many variables are available. Within this "large model approach" the large number of variables is replaced with a small number of factors and several aspects of this approach are discussed. In Section 3, the stochastic restriction search is introduced as a means of efficiently selecting a subset of macroeconomic variables or factors that should be restricted from the VAR specification, based only on the information in the data. The prior specification necessary for model selection is analyzed, as well as the interpretation of model selection probabilities as a special case of BMA. Section 4 outlines the setting of the empirical section (data, forecasting models, prior hyperparameters, and comparison statistics), and the results of the forecasting performance of various VAR specifications. Section 5 concludes the paper with a summary and thoughts for further extension of the basic framework presented in this paper.

3.2 Methodology

Let y_t be an $m \times 1$ vector of variables of interest (that we want to forecast) observed for $t = 1, ..., T$. Unlike previous univariate studies (Stock and Watson, 2002, Koop and Potter, 2004), $m > 1$ and I define a forecasting model for y using a general VAR representation

$$y_t = \sum_{i=1}^{p_1} a_i y_{t-i} + c_0 w_t + \varepsilon_t \tag{3.1}$$

where the parameter matrices a_i and c_0 are of dimensions $m \times m$ and $m \times N$ respectively, y_{t-i}, $i = 1, ..., p_1$, are lagged values of the dependent variable, w_t is a $N \times 1$ vector containing current and lagged values of some exogenous predictor variables, and the errors are iid Gaussian, $\varepsilon_t \sim N(0, \Sigma)$.This model can be estimated both by OLS and Bayesian methods, provided that the total number of explanatory variables will not exceed the total number of time series observations T. I propose to adopt a Bayesian setting which allows for a unified treatment of this model in high dimensions. For a review of the VAR under standard prior specifications and different sampling methods, the reader is referred to Kadiyala and Karlsson (1997).

Assume we have available observations $x_t = (x_{1t}, ..., x_{nt})'$ on some macroeconomic quantities, where n is large (in the order of hundreds). A popular and simple method to incorporate into an econometric model all the information inherent in a large set of variables, is to reduce their dimension into a lower-dimensional vector of $k \ll n$ latent factors and insert these in the

VAR model as explanatory variables

$$x_t = \lambda f_t + u_t \tag{3.2}$$

$$y_t' = \sum_{i=1}^{p_1} y_{t-i}' a_i + \sum_{j=1}^{p_2} f_{t-j}' b_j + \varepsilon_t \tag{3.3}$$

where f_t is an $k \times 1$ vector of unobserved factors, λ is the matrix of factor loadings and u_t are iid normal errors, $u_t \sim N(0, W)$. In equation (3.3) the same assumptions hold as in the base model in (3.1), with the only difference that now $w_t = (f_{t-1}, ..., f_{t-p_2})'$ and $N = k \times p_2$, and the b_j are of appropriate dimensions. For simplicity x_t is demeaned which is equivalent to imposing a constant term in the factor equation, equal to the sample mean $\overline{x} = \frac{1}{T}\sum x_t$ (which in this model coincides both with the MLE of the constant or the mode of its posterior under a diffuse prior). The factors are unobserved quantities and usually it is assumed that they follow a normal distribution with diagonal covariance matrix. One more convention in the factor model literature is to impose the covariance matrix of the innovations, W, to be also diagonal so that (3.2) reduces to independent equations. Estimation methods vary from principal component analysis (PCA) to full likelihood-based approaches. The ultimate goal of using the factor model is to obtain the factor scores f_t as a valid reduced representation of the manifest vector x_t, so that factor identifiability issues play no actual role here and will not be further discussed.

In terms of the general forecasting VAR model in equation (3.1), I replace the predictors w_t with the principal components (PC) estimates of the factors $\widehat{F}_t = \left[\widehat{f}_t, \widehat{f}_{t-1}, ..., \widehat{f}_{t-p_2}\right]$, i.e., as if they were observed data. Note

that this specification is slightly different from the dynamic factor model (or factor-augmented VAR) used in Bernanke et al. (2005). From their point of view, the dynamic factor model (DFM) is treated as a state-space model, which has the advantage of a probably more efficient one-step estimation of the factors (i.e., along with the parameters of the model) through the Kalman filter algorithm. But this comes at a huge computational cost which makes the application of this model prohibitive in the recursive forecasting setting adopted in this study. After all, Stock and Watson (2005a) have already implemented a large-scale forecasting exercise involving DFMs where they compare several frequentist, full Bayes, and empirical Bayes approaches.

The factors replace the original variables in order to allow richer dynamics and subsequently are allowed to have up to p_2 lags. If the original observed series $x_t = (x_{1t}, ..., x_{nt})'$ were included as predictors then – for a typical macroeconomic dataset with monthly observations on many variables – a degrees of freedom problem would occur if more than one or two lags were assumed. However, even in the case of reducing the dimension of our data with factors the fact that we would ideally allow for many lags does not resolve the problem of overparameterization. For $N = k \times p_2$ larger than 20 the number of all possible models will tend virtually to infinity so that pairwise model comparison is practically infeasible using an AIC/BIC-type criterion or prior predictive (marginal) densities and Bayes factors. A reasonable proposed solution from a Bayesian point of view is to use shrinkage subjective priors. For example, the Minnesota prior imposes restrictions on parameters which correspond to higher order lags of y, whereas the prior

weight (i.e., the prior mean) for the parameter on the first own lag in each of the m equations is equal to one, and zero on the first lag of the rest $m-1$ dependent variables. While this approach will work well in VARs which include only lags of the dependent variables, it is difficult to adopt this approach in the models examined here. This happens because there is no theoretical or empirical justification for constructing a subjective prior on exogenous predictor variables, especially if these exogenous variables are latent (constructed) factors.

Introducing any kind of subjective prior information in this model is not an easy task, anyway. These priors may not be specified concretely because of the lack of prior information regarding joint distributions or the large amount of models involved in the analysis. In that respect, subjective prior beliefs require a huge amount of input from the researcher. It is unrealistic to assume that uncertainty about the true model specification can be described meaningfully using ones' own beliefs; hence prior elicitation should be based mainly on economic theory. The problem with this approach is that in many cases economic theory has empirically proven to be bad guidance in proposing relevant predictors. Stock and Watson (2003) argue that this is the case when forecasting inflation: "the literature does suggest [. . .] variables with the clearest theoretical justification for use as predictors often have scant empirical predictive content."

The discussion so far has focused on the "large-n" case, avoiding to mention anything about how small or large the dimension m of the dependent variable y should be. Although macroeconomic VARs typically contain as dependent variables three or four fundamental quantities that describe the

economy, when forecasting, the actual number of variables of interest can grow large. A decision maker would be interested to forecast future values of many series, like production, employment/unemployment, short- and long-term interest rates, consumer and producer price inflation, exchange rates, and many other nominal or real quantities. This is easily handled with the model selection algorithm which is the focus of the next section. The methods described below apply to large VARs in a general sense, that is (i) when the number of predictors n grows large and the number of dependent variables m is small, (ii) when m grows large and n is small, or (iii) when both $m, n \to \infty$, although the empirical application is centered upon the first case.

3.3 Bayesian model selection and averaging

As was mentioned in the introductory section, when the number of candidate models is too large to enumerate, posterior sampling methods are necessary for the computation of marginal likelihoods for model comparison. Stochastic search algorithms that base on a Markov chain on model space identify regions of high posterior probability and can be used for model selection or to obtain posterior weighted estimates for model averaging. When applied to small models, these algorithms have the ability to search the entire model space, while in large settings only more plausible models are visited. An indicator (zero/one) variable γ, epitomizes the core of Bayesian model selection using stochastic search techniques. Let us define the vector $\gamma = (\gamma_1, ..., \gamma_s)$ as the complete set of indicators, where s

is the maximum number of variables in the model. When $\gamma_i = 0$ ($\gamma_i = 1$) then variable j exits (enters) the "true" model, $i = 1, ..., s$. This allows to index all possible 2^s models as combinations of the available variables. Then we can proceed by defining a prior $p(\gamma)$ which combined with the likelihood $p(data|\gamma)$, will give zero or one value for each γ_i, $i = 1, ..., s$, from the (updated based on data) posterior distribution $p(\gamma|data)$. This posterior distribution entails all the necessary information for model selection and averaging. The main idea is to impose the vector of parameters, say $\theta = (\theta_1, ..., \theta_s)$, to have a structure conditional on the values of γ, so that when $\gamma_i = 1$ the associated parameter θ_i will be estimated according to its unrestricted posterior density, and when $\gamma_i = 0$ this would imply $\theta_i = 0$.

There are many ways to implement this general strategy and many alternative methods exist which involve several prior specifications. Analytical reviews of model averaging and selection is offered in Hoeting et al. (1998), O'Hara and Sillanpää (2009), George (2000), Clyde and George (2004), and Chipman et al. (2001). Recent references of BMA in economics include Fernandez, Ley and Steel (2001), Ciccone and Jarocinski (2007), and Sala-I-Martin, Doppelhofer and Miller (2004). Ciccone and Jarocinski (2007) in particular implement an investigation of BMA in the presence of data revisions. A computationally fast restriction search is described in this section which is based on the SSVS algorithm of George and McCulloch (1993, 1997). The general idea is to use a mixture prior on the parameters we want to restrict, conditional on the model selection indicators γ. That is, we can

write a prior for θ of the form

$$\theta_i|\gamma_i \sim (1-\gamma_i)\,\delta(0) + \gamma_i N(0,\eta) \tag{3.4}$$

where $\delta(0)$ is the Dirac delta function with mass at 0. Thus when $\gamma_i = 0$, θ_i will be shrunk towards zero, due to very tight prior. When $\gamma_i = 0$, θ_i will be left unrestricted (Normal prior with "large" variance η). The exact SSVS implementation of the above prior from George and McCullogh (1993), sets the first mixture component to be Normal (instead of the Dirac delta at zero) with a very tight variance (see below for more details).

A different approach is to set in (3.4) a double exponential prior in the second mixture component. This can be implemented if we assume at a second level a hyperprior for η of the form

$$\eta \sim \exp\left(\tau^2/2\right).$$

This double exponential prior can better accommodate large regression coe cients due to its heavier tail probability. Additionally it can achieve the adaptive minimax convergence rates that are not obtainable using normal priors. Most importantly, this empirical Bayes estimator is closely related to the LASSO algorithm (see Liang et al., 2008).

Define $z_t = \left[y'_{t-1}, ..., y'_{t-p_1}, w'_t\right]'$, then the VAR model in familiar matrix form is obtained by stacking the row vectors y_{t+1}, z_t and ε_t for $t = 1, ..., T$

$$y = z\phi + \varepsilon, \quad \varepsilon \sim N(0, \Sigma) \tag{3.5}$$

where $y = \left[y_2', ..., y_{T+1}'\right]'$, $z = [z_1', ..., z_T']'$, $\phi = [a_0, ..., a_{p_1}, c_0]$, and $\varepsilon = \left[\varepsilon_2', ..., \varepsilon_{T+1}'\right]'$. Note that when forecasts are projected h-steps ahead, y is the matrix $y = \left[y_{1+h}', ..., y_{T+h}'\right]'$ (see next section for a definition). Let $n_u = m \times (m \times (p_1 + 1) + k \times (p_2 + 1))$ be the total number of elements in $\varphi = vec(\phi)$. From these elements the m in total constants are always included in the models and admit a typical normal prior of the form

$$\varphi^c \sim N\left(\underline{\varphi}^c, vI_m\right) \tag{3.6}$$

where φ^c is the block of φ which contains the constant terms. Let φ^k be the vector of the remaining $n_\varphi = n_u - m$ parameters in φ which are subject to restriction search and let $\gamma = \left(\gamma_1, ..., \gamma_{n_\phi}\right)$ be the vector of indicator variables associated with the elements of φ^k. Then each element φ_i^k conditional on γ_i, $i = 1, ..., n_\varphi$, follows a scale mixture of normals prior of the form

$$\varphi_i^k|\gamma_i \sim (1 - \gamma_i) N\left(0, \tau_{0i}^2\right) + \gamma_i N\left(0, \tau_{1i}^2\right) \tag{3.7}$$

The hyperparameters τ_{0i}, τ_{1i} are selected in such a way so that τ_{0i}^2 is small (or even zero) and τ_{1i}^2 is large. Subsequently each parameter φ_i^k is restricted with zero prior mean and very small (or zero) prior variance when $\gamma_i = 0$, while for $\gamma_i = 1$ has a large (locally uninformative) prior variance and in that respect is left unrestricted.

It would not make sense to define the γ_i's if these were defined subjectively and not updated by the information in the data. Hence a Bernoulli prior on these variables is placed, which updated by the likelihood will result in a conditional posterior which is also Bernoulli. The elements of the

vector γ follow an independent Bernoulli $p_i \in (0,1)$ prior of the form

$$\gamma \sim \prod p_i^{\gamma_i} (1-p_i)^{(1-\gamma_i)}, \quad i = 1, ..., n_\varphi \tag{3.8}$$

This prior choice reduces computational costs and leads to a posterior density which is easy to derive. In this case $p(\gamma_i = 1) = p_i = 1 - p(\gamma_i = 0)$ so that p_i reflects the prior belief that φ_i^k is large enough and should be left unrestricted. By selecting $p_i < 1/2$, models with an unreasonably large number of parameters are downweighted in order to highlight the significance of parsimonious models. The special case where $p_i = 1/2 \ \forall \ i$, is equivalent to a constant uniform prior $p(\gamma) \equiv 1/2^{n_\varphi}$. This prior is uninformative in the sense that it favors each parameter equally; see Section 4.2 in this paper for more details, and the discussion in Chipman et al. (2001).

The hierarchical mixture prior described above is straightforward to interpret and can be applied virtually to any model for which a normal prior can be specified3 as the conjugate prior that leads to easy derivation of the underlying posterior. A different version of the SSVS is used in Brown et al. (1998) for a multivariate regression model used to predict three variables using 160 predictors. Following the suggestions of George and McCulloch (1997) and Smith and Kohn (1996) they set in equation (3.7) $\tau_{0i}^2 = 0$ and $\tau_{1i}^2 = g \times \left(z_\gamma' z_\gamma\right)^{-1}$. This prior implies that the first component of the mixture is a Dirac delta function at zero, i.e., a function that puts point mass at zero and hence whenever $\gamma_i = 0$, φ_i^k will be exactly zero. The second component is Zellner's g-prior specification and suggestions for setting uninformative values of g (although in a univariate context) are given

in Fernandez et al. (2001). An updated and computationally more efficient version of this prior specification appears in Brown et al. (2002), where more variables than observations can be handled. The shortcoming of their approach is that it is able to treat each equation in the VAR individually, but instead is choosing the variables in z which are more probable to be included in *all* VAR equations together. Put simply, if, say, z contains only the first lag of the dependent variables, then the latter approach will allow the y_{it-1} to be a predictor of the whole vector y_t, while the approach proposed here $y_{it-1}$to explain the dependent variable in equation j of the VAR (denoted y_{jt}), but not the dependent variable in the l-th VAR equation (denoted y_{lt}). Nevertheless, the Brown et al. (2002) implementation of the SSVS algorithm is a valuable complement to the one used here, and undoubtedly a useful tool in empirical analysis with focus on prediction.

Smith and Kohn (2002) extend the stochastic search for parameter restrictions to the covariance matrix of longitudinal data. George et al. (2008) apply their idea to the covariance matrix of structural VARs: motivated by the fact that identifying restrictions on the covariance are usually imposed on the elements of a reparametrization of Σ, they focus on restricting the elements of the $m \times m$ upper triangular matrix Ψ satisfying

$$\Sigma^{-1} = \Psi'\Psi \tag{3.9}$$

They then derive a mixture of normals prior, as in equation (3.7), for the nondiagonal elements of Ψ, while the diagonal is integrated out with a

gamma prior. Matrix Ψ has the form

$$\Psi = \begin{bmatrix} \psi_{11} & \psi_{12} & \cdots & \psi_{1m} \\ 0 & \psi_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \psi_{mm} \end{bmatrix} \tag{3.10}$$

so let $\boldsymbol{\psi} = (\psi_{11}, ..., \psi_{mm})'$ and $\eta = (\eta_2', ..., \eta_m')' = \left(\psi_{12}, \psi_{13}, \psi_{23}, ..., \psi_{(m-1)m}\right)'$be the vectors of the diagonal and upper diagonal elements respectively, where $\eta_j = \left(\psi_{1j}, ..., \psi_{(j-1)j}\right)'$ for $j = 2, ..., m$. Let $\omega_j = \left(\omega_{1j}, ..., \omega_{(j-1)j}\right)'$ be a vector of 0-1 variables so that each element of η_j has prior conditional on ω_j of the form

$$\eta_{ij}|\omega_{ij} \sim (1 - \omega_{ij}) N\left(0, \kappa_{0ij}^2\right) + \omega_{ij} N\left(0, \kappa_{1ij}^2\right) \tag{3.11}$$

for $i = 1, ..., j-1$ and $j = 2, ..., m$. As in the case of the vector γ, assume that the elements of the vector $\omega = (\omega_2', ..., \omega_m')'$ are independent Bernoulli $q_{ij} \in (0, 1)$ random variables so that

$$\omega \sim \prod_i \prod_j q_{ij}^{\omega_{ij}} (1 - q_{ij})^{(1-\omega_{ij})} \tag{3.12}$$

For $i = 2, ..., m$, each ψ_{ii} has a gamma prior density

$$\psi_{ii}^2 \sim Gamma\left(\alpha_i, \beta_i\right)$$

For more information on these priors the reader is referred to the ana-

lytical calculations of George et al. (2008) where it is shown that finding restrictions on the covariance matrix based solely on the data provides an attractive alternative to identifying restrictions imposed in structural VARs. It should be clear from the prior specification that the SSVS is an intuitive extension of the Bayesian conjugate (normal – inverse Wishart) prior. In the empirical application I adopt a fast sampling scheme (see Section 4.2) to draw from the posteriors of γ and ω, which makes computation feasible in multivariate models. The parameter posteriors are given in detail in Appendix B. Although selection of prior hyperparameters seems to be fairly automatic in this setting, prior elicitation is an important factor in model selection.

3.4 Empirical Application

Data

I use the Stock and Watson (2005b) dataset which is an updated version of the Stock and Watson (2002) dataset that is widely used in empirical applications. This version consists of 132 monthly variables pertaining to the US economy measured from 1960:01 to 2003:12. The 132 predictors can be grouped in 14 categories: real output and income; employment and hours; real retail, manufacturing, and trade sales; consumption; housing starts and sales; real inventories; orders; stock prices; exchange rates; interest rates and spreads; money and credit quantity aggregates; price indexes; average hourly earnings; and miscellaneous. The data were transformed to

eliminate trends and nonstationarities. All the data and transformations are summarized in Appendix C.

Selection of prior hyperparameters

Implementation of Bayesian model selection requires all the priors to be proper, as the ones described in Section 3. Noninformative improper priors are not suitable to calculate Bayes factors and posterior model probabilities. Even though there are certain methods which overcome this difficulty (BIC approximations, intrinsic, or fractional Bayes factors), the standard practice in the Bayesian model selection literature is to use only proper priors. This does not necessarily mean that noninformative proper priors cannot be specified. It is easy to choose the hyperparameters in such a way that all the priors are locally noninformative.

Selection of τ_{0i}, τ_{1i} and κ_{0ij}, κ_{1ij} can be made along the guidelines of Chipman et al. (2001, p. 86). For instance, given a non-negative scalar threshold ξ_i, higher posterior weighting can be allocated to those values of γ for which $\left|\varphi_i^k\right| > \xi_i$ when $\gamma_i = 1$, iff τ_{0i}, τ_{1i} satisfy

$$\log\left(\frac{\tau_{1i}/\tau_{0i}}{\tau_{0i}^{-1} - \tau_{1i}^{-1}}\right) = \xi_i^2$$

A similar argument can be made for the choice of κ_{0ij} and κ_{1ij}. Alternatives for a more objective selection of these hyperparameters exist, but at the cost of a substantial increase in computational calculations. The first one is to use empirical Bayes criteria in the spirit of George and Foster (2000), while a fully Bayes approach would require to place an inverted-Gamma hyperprior

on each τ_{0i}, τ_{1i} and κ_{0ij}, κ_{1ij}. Selection based on the formula above is a simple task which can easily be implemented in large models. George et al. (2008) argue that even if the restriction search algorithm is not effective in selecting the correct restrictions on ϕ, the results can still be used to obtain improved forecasts.

The only source of difficulty may arise in eliciting the hyperparameters of the Bernoulli random variables γ (similarly ω). The prior structure that appears in equation (3.8) (similarly in equation (3.12)) is an "independence prior," in the sense that each element of γ (ω) is assumed to be independent of the rest. This simplification makes it difficult to account for similarities or differences between models when the correlation between the explanatory variables is high. Just using a prior probability of inclusion for all variables, equal to 1/2 has implications in this case. A clear example of this is when three different but highly correlated measures of the same quantity are used (say three different measures of unemployment). Then with the uniform prior, the prior probability of unemployment having an effect would be 7/8, not 1/2.While priors that "dilute" probability across neighborhoods of similar models (Chipman et al., 2001; Yuan & Lin, 2005) are able to correct this shortcoming, it is preferable to use an orthogonal transformation of the variables[1] in z, by applying a singular value decomposition. This allows exploring the model space in considerably less iterations, which subsequently decreases the computational cost in multivariate mod-

[1]Note that when orthogonalizing variables, their economic interpretation is lost. Assume that x_1 and x_2 are predictors and we implement model selection to their orthogonalized equivalents, x_1^* and x_2^*. If x_1^* is included in the "true" model, this does not imply that x_1 should be selected.

els. Hence, in the forecasting exercise, I apply the restriction search to the model

$$y_{T+h} = G_T \mu_h + \varepsilon_{T+h}$$

where $G = zH$ are orthogonal variables and $\mu = H^{-1}\phi$; see Koop and Potter (2004). This approach will speed up computations, even though orthogonality does not lead to posterior independence of elements of γ. The default choice $p_i = 1/2$ in equation (3.8) and $q_{ij} = 1/2$ in equation (3.12) may result in a uniform prior, but this would not be a noninformative prior about model size. A rule of thumb is that if the researcher anticipates many (few) restrictions on the model then the choice should be $p_i, q_{ij} < 1/2$ ($p_i, q_{ij} > 1/2$). Prior sensitivity analysis using real and simulated data showed that $p_i = q_{ij} = 1/2$ is able to identify restrictions quite well and hence is left as the default reasonable choice.

Following the suggestions of George et al. (2008) and George and McCulloch (1997), I adopt a fast sampling scheme for γ and ω, which requires to set τ_{0i} and κ_{0ij} small, but different from 0. According to the preceding discussion in this subsection and the absence of prior beliefs about specific parameters I set $\tau_{0i} = \tau_0 = 0.01$, $\tau_{1i} = \tau_1 = 70$ for all $i = 1, ..., n_\varphi$, and $\kappa_{0ij} = \kappa_0 = 0.01$, $\kappa_{1ij} = \kappa_1 = 30$ for all $j = 2, ..., m$ and $i = 1, ..., j-1$. For the intercept term, the typical normal prior has mean $\underline{\varphi}^c = 1$ and variance $v = 100$. A default noninformative choice for the parameters of the Gamma density is $\alpha_i, \beta_i = 0.01$.

Implementation of Bayesian Model Averaging/Selection

At this point, as it is practically impossible to summarize model selection results from the recursive forecasting exercise, I summarize the average posterior probability of some of the variables in the dataset without extracting factors, i.e., replacing w_t with $x_t = (x_{1t}, ..., x_{nt})'$ in specification (3.1), and using the full sample of observations from 1960:1 to 2003:12. I am not presenting the results with factors (which is the main model used for forecasting in the empirical application), since those factors are latent and their selection does not have economic contect. Subsequently this demonstration of model averaging/selection is to show how the SSVS algorithm can choose which variables can affect (or not) the dependent variables in a VAR, and then verify if these results comply with expectations from economic theory and the empirical literature. I consider a New Keynesian VAR with three variables (unemployment, consumer price index, and federal funds rate) regressed on an intercept, 14 autoregressive lags, and the remaining 129 variables in the dataset which are used as exogenous predictors. This gives a total of $129 + 13 \times 3 = 168$ right-hand side variables (excluding the intercept which is always included) to choose from in each equation. The horizon chosen in this illustration is $h = 12$. The unemployment and interest rate are transformed to stationarity by taking first differences. The consumer price index is transformed by taking the second difference of the logarithm.

A parameter should either be included or excluded, hence the number

of all possible models is 2^{168} in each VAR equation and $2^{168\times 3} = 5.2e+151$ in total. The BMA posterior probabilities are computed for each parameter $i = 1, ..., n_{\varphi}$ as

$$E\left(\gamma_i|y\right) = \frac{1}{S}\sum_{s=1}^{S}\gamma_i^{(s)}$$

where S is the total number of iterations from the posterior sampler, and $\gamma_i^{(s)}$ are draws from the conditional posterior of γ_i. This suggests that the average probability is actually the proportion of models visited by the Gibbs sampler, which contain the corresponding variable. Exactly similar inference and interpretation holds for the parameters ω, although these index elements of the covariance matrix and not columns of predictors in mean VAR equation.

Tables 3.1 and 3.2 summarize the results for those predictor variables and own lags, respectively, that have the highest probabilities. Variables which had average posterior probability less than 0.5 in all of the three equations are not included at all in the tables. Each element in these tables is the BMA posterior probability and can be interpreted simply as the probability that the corresponding right-hand side variable should be included. For this specific application the variables are not orthogonalized in order to retain the interpretation of the probabilities as the amount of belief that the respective variable is included in the model. The results are based on 150,000 iterations with a burn-in period of 50,000, which leaves 100,000 draws to evaluate the posterior of γ. Elicitation of prior hyperparameters is based on the values described earlier.

Note that the probabilities ω for Ψ are 0.52, 1, and 1 for each of the upper

Table 3.1: Average Posterior Probabilities of Explanatory Variables in the 3-variable VAR

Explanatory variables	u_{t+12}	cpi_{t+12}	r_{t+12}
Personal income	0.141	0.001	0.949
IP index - Final products	0.251	0.003	0.564
IP index - Manufacturing	0.593	0.016	0.17
Capacity Utilization	1	0.124	0.032
Employment ratio	0.011	0.002	0.992
Civilian labor force: Total employed	0.428	0.003	0.652
Employees on nonfarm payrolls - Total private	0.811	0.018	0.317
Employees on nonfarm payrolls - Manufacturing	0.5	0.014	0.33
Employees on nonfarm payrolls - Service-providing	1	0.023	0.826
Employees on nfm prl - Trade, transportation and utilities	0.878	0.003	0.682
Employees on nonfarm payrolls - Wholesale trade	0.296	0.003	1
Employees on nonfarm payrolls - Financial activities	0.687	0.008	0.697
Average weekly hours of production	0.001	0.082	0.941
Housing starts: Total	0.879	0.001	0.04
Housing authorized: Total	1	0.001	1
Houses authorized by building permits: Northeast	1	0.105	0.003
Houses authorized by building permits: Midwest	1	0.025	0.018
Houses authorized by building permits: South	1	0.001	0.006
Houses authorized by building permits: West	1	0	1
Consumer installment credit to Personal income (ratio)	0.013	0.001	1
S&P'S common stock price index: Composite	0.962	0.132	0.004
S&P's composite common stock: Dividend yield	0.092	0.001	0.937
Commercial paper rate (spread from Fed Funds Rate)	0.028	0.7452	0.851
3-month interest rate (spread from FFR)	0.002	0.087	1
6-month interest rate (spread from FFR)	0.005	0.002	1
1-year interest rate (spread from FFR)	0.941	0.752	0.992
5-year interest rate (spread from FFR)	1	0.982	1
10-year interest rate (spread from FFR)	1	0.861	1
Bond yield: Moody's BAA corporate (spread from FFR)	0.001	0	0.978
NAPM commodity prices index	0.0012	0.867	0.857
CPI-U: Durables	0.172	0.002	0.543
CPI-U: All items less shelter	0.246	0.006	0.692

Table 3.2: Average Posterior Probabilities of autoregressive lags in the 3-variable VAR

Dependent Variable	Most important lags (probability>0.5)	Average posterior probability
u_{t+12}	r_{t-8}	0.56
	r_{t-8}	0.74
cpi_{t+12}	Own lags 1 to 7 (cpi_{t-1}-cpi_{t-7})	1
	cpi_{t-8}	0.83
r_{t+12}	r_{t-7}	1

diagonal elements ψ_{12}, ψ_{13}, and ψ_{23} respectively. Once all these probabilities are available, it is straightforward to interpret them. This output can be used to implement BMA if all variables contribute to the final forecast according to their probability, no matter how high or low this probability is. Looking for example at Table 3.1, the spread of the 10-year interest rate from the federal funds rate variable will contribute to the final forecast of the unemployment rate, the consumer price index, and the interest rate in 100, 86.1, and 100%of the occasions (models visited by the sampler), respectively. In contrast the same output can be used to select the best single model. Barbieri and Berger (2004) show that in the context of Bayesian model selection the optimal model is the median probability model. According to this result, only the variables which have average probability larger than 0.5 in each equation will be unrestricted. These probabilities are presented in Tables 3.1 and 3.2. Hence, in this "best" model, the 1, 5, and 10-year interest rate spreads should be included in all three equations, while capacity utilization should enter only the unemployment equation.

The results presented in Table 3.1 are also subject to economic interpretation. Space restrictions, however, do not allow further analysis in

this study. Structural interpretation is not the main focus, but forecast improvement is. This is an issue examined in the following section.

Forecasting in Large VAR Models

The first estimation period is set to 1960:1 and a simulated real-time forecasting of y_{t+h} is done from 1983:1 through 2003:12-h, for horizons $h = 1, 6$, and 12. Each VAR model has eight dependent variables of interest (with their short mnemonic from the dataset in parentheses): Personal Income ($A0M052$), Industrial Production ($IPS10$), Employment Rate ($CES002$), Unemployment Rate ($LHUR$), 3-month Treasury Bill Rate ($FYGM3$), Producer Price Index ($PWFSA$), Consumer Price Index ($PUNEW$), and PCE Deflator ($GMDC$). This leaves a total of 124 variables to explore their predictive content. There are methods available to forecast with non-stationary variables. Nevertheless in this paper I am forecasting with stationary VARs only. All the variables are transformed to stationarity as in the Appendix, however the dependent variables in y_{t+h} are transformed as follows for the purpose of forecasting. Let v_{it} denote the untransformed value of y_{it} for each of the eight monthly dependent variables i, then $y_{it+h} = (1200/h)\log(v_{it+h}/v_{it})$ for $i = (A0M052, IPS10, CES002)$, $y_{it+h} = v_{it+h} - v_{it}$ for $i = (LHUR, FYGM3)$, and $y_{it+h} = (1200/h)\{\log(v_{it+h}/v_{it}) - h\Delta\log(v_{it})\}$ for $i = (PWFSA, PUNEW, GMDC)$.

The principal components are estimated from the 124 variables in the dataset using the same sample period as the VAR. Several multivariate forecasting exercises in the literature (cf. Stock and Watson, 2002) focus

on finding the best performing model. In contrast, here the main challenge is to improve forecasts when the number of predictors grows large and the researcher has no prior information about which is the correct model size. Thus, the maximum potential number of factors and lags is deliberately set to large, "uninformative" values. In particular, 10 principal components ($k = 10$) are extracted from the factor model in equation (3.2), while the VAR specification in equation (3.3) contains an intercept, 13 autoregressive lags ($p_1 = 13$), and 13 lagged factors ($p_2 = 13$). This gives a maximum of 221 (plus the intercepts, which are unrestricted) potential predictors on each of the 8 dependent variables. For the purpose of the empirical application forecasts are computed from: (i) VAR with SSVS and model averaging, (ii) VAR with SSVS and model selection, and (iii) VAR estimated using OLS with selection of predictors with the Bayesian information criterion (which has a larger penalty for less parsimonious models relative to the Akaike information criterion, and is a rough approximation to the Bayes factors). The predictors in the latter method are orthogonalized and the total number of possible models considered is equal to the maximum number of right-hand side variables and subsequently selection of the best model is implemented in a finite number of calculations.

An appropriate common way to quantify out-of-sample forecasting performance is to compute the root mean square forecast error (RMSFE) statistic for each forecast horizon h:

$$RMSFE_{ij}^{h} = \sqrt{\sum_{t=1982:12}^{2003:12-h} \left(y_{i,t+h}^{*} - \widetilde{y}_{i,t+h,j}\right)^{2}} \tag{3.13}$$

where $y^*_{i,t+h}$ is the realized (observed) value of y at time $t+h$ for the i-th series, and $\widetilde{y}_{i,t+h,j}$ is the mean of the posterior predictive density at time $t+h$, for the i-th series, from the j-th forecasting model. The RMSFE of each model is reported relative to the RMSFE of a benchmark VAR with an intercept and seven lags of the dependent variables, estimated with OLS

$$rRMSFE^h_{ij} = \frac{RMSFE^h_{ij}}{RMSFE^h_{iVAR(7)}} \tag{3.14}$$

This VAR(7) model is not chosen because of its higher forecasting ability compared to other alternatives. Following the standard convention in the literature an AR(2) model would be a better candidate to serve as the benchmark model. But note that the VAR(7) is nested to the VAR with factors, which will give a better picture of whether the restrictions found by the SSVS are actually the ones that will lead to reduced RMSFE statistics, compared to a more parsimonious alternative. The forecasting performance of the models based on the relative RMSFE for horizons $h = 1, 6, 12$, is summarized in Table 3.3. These are the averaged values of the RMSFEs over the forecast period, 1983:1 through 2003:12-h.

The results are encouraging about the application of the restriction search algorithm in large models. In most occasions the BMA and Bayesian model selection give improved results compared to the BIC selection. Note that the improvement is not only due to the fact that the models of interest contain more predictors than the benchmark model. It is noteworthy that in some occasions only lags of the dependent variable are selected from the restriction search, while for most samples the number of important lagged

Table 3.3: Forecast Comparison - relative RMSFE

	PI	IP	EMP	UR	TBILL	PPI	CPI	PCED
BVAR with factors (Bayesian Model Averaging)								
$h = 1$	0.94	1	0.9	0.96	1.08	0.88	0.95	1.09
$h = 4$	1.06	0.96	0.93	0.94	0.95	0.92	1.05	0.94
$h = 12$	0.97	0.92	0.99	1.02	0.98	0.92	0.95	0.96
BVAR with factors (Model Selection)								
$h = 1$	0.86	0.98	0.87	0.96	1.06	0.91	0.93	0.91
$h = 4$	0.9	0.97	0.85	0.92	0.94	0.94	0.98	0.93
$h = 12$	0.87	0.99	0.91	0.98	0.89	0.87	0.99	0.96
VAR with factors (BIC Selection)								
$h = 1$	0.92	0.99	0.94	0.99	1.22	0.99	1.01	0.97
$h = 4$	0.93	0.97	0.94	0.94	1.12	0.97	1.06	0.94
$h = 12$	0.97	1.04	0.98	1.05	0.99	0.9	1.1	0.95

Note: The variables of interest are: PI: Personal Income (A0M052), IP: Industrial Production (IPS10), EMP:Employment Rate (CES002), UR: Unemployment Rate (LHUR), TBILL: 3-month Treasury Bill Rate (FYGM3), PPI: Producer Price Index (PWFSA), CPI: Consumer Price Index (PUNEW), andPCED: PCE Deflator (GMDC)

factors, for each dependent variable, is not more than five. This is supported by the fact that the average RMSFE (results not reported here) of the large VAR with factors but without selection of predictors (i.e., a heavily overparametrized model) is, as expected, extremely high relative to the VAR(7). An important feature of the restriction search algorithm applied to the specific VAR is that the forecasts from Bayesian model selection are better than the forecasts from BMA. The practical difference of the two approaches is that BMA shrinks the posterior means of the parameter with low probability toward zero, while Bayesian model selection imposes that these parameters (with probability less than 0.5) will be exactly zero.

3.5 Conclusions

This paper addresses the forecasting performance of Bayesian VAR models with many predictors using a flexible prior structure which leads to output that can be used for model selection and model averaging. For eight U.S. monthly macroeconomic variables of interest forecasting accuracy is improved over least squares estimation and selection of predictors using the Bayesian information criterion. Without arguing that the choice of prior hyperparameters was the best possible and done with a strict "objective" criterion (like in other BMA applications, see Fernandez et al., 2001), the gains from the standard automated choices are appreciable. As already mentioned, there are many proposals in the Bayesian literature for efficient elicitation of prior hyperparameters for model selection and some of them were discussed in the paper. Nevertheless, the merit of the SSVS for VAR models lies in its simplicity and intuitive interpretation.

With regard to other macroeconometric specifications, the flexibility of the restriction search algorithm suggests many interesting extensions. Firstly, note that it is straightforward to adopt it in general piecewise-linear multivariate regressions that allow for thresholds, Markov switching or structural breaks; an interesting area for future research. Secondly, I only considered the case where the number of dependent variables, m, is small and the number of predictors grows large. But as already mentioned the restriction search algorithm may also be used when the number of dependent variables grows large. Banbura et al. (2010) examine this case using shrinkage priors and find huge gains from this large VAR specification. Lastly, an

interesting direction for future research would be the empirical application of the restriction search algorithm in the Bayesian dynamic factor model. This approach will probably improve forecasting performance and impulse response analysis in DFMs that lack parsimony (see Bernanke et al., 2005, and Stock and Watson, 2005b).

Chapter 4

Forecasting using Bayesian variable selection

4.1 Introduction

Since the pioneering work of Sims (1980), a large part of empirical macroeconomic modeling is based on vector autoregressions (VARs). Despite their popularity, the flexibility of VAR models entails the danger of overparameterization which can lead to problematic predictions. This pitfall of VAR modelling was recognized early and shrinkage methods have been proposed; see for example the so-called Minnesota prior (Doan et al., 1984). Nowadays the toolbox of applied econometricians includes numerous efficient modelling tools to prevent the proliferation of parameters and eliminate parameter/model uncertainty, like variable selection priors (George et al. 2008), steady-state priors (Villani, 2009), Bayesian model averaging (Andersson and Karlsson, 2008) and factor models (Stock and Watson,

2005a, 2005b), to name but a few.

This paper develops a stochastic search algorithm for variable selection in linear and nonlinear vector autoregressions (VARs) using Markov Chain Monte Carlo (MCMC) methods; see Gilks et al., (1996). The term "stochastic search" simply means that if the model space is too large to assess in a deterministic manner (say estimate all possible model combinations and decide on the best model), the algorithm will visit only the most probable models. In this paper the general model form that I am studying is the reduced-form VAR model, which can be written using the following linear regression specification

$$y_t = Bx_t + \varepsilon_t \tag{4.1}$$

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

$$\begin{aligned} B_{ij} &= 0 \text{ if } \gamma_{ij} = 0 \\ B_{ij} &\neq 0 \text{ if } \gamma_{ij} = 1 \end{aligned} \tag{4.2}$$

where B_{ij} is an element of the matrix B, for $i = 1, .., m$ and $j = 1, ..., k$.

There are various benefits of using this approach over the shrinkage

methods mentioned previously. First, variable selection is automatic, meaning that along with estimates of the parameters we get associated probabilities of inclusion of each parameter in the "best" model. In that respect, the variables γ_{ij} indicate which elements of B should be included or excluded from the final optimal model, thus implementing a selection among all possible $2^{m\times k}$ VAR model combinations, without the need to estimate each and everyone of these models. Second, this form of Bayesian variable selection is independent of the prior assumptions about the parameters B. That is, if the researcher has defined any desirable prior for her parameters of the unrestricted model (4.1), adopting the variable selection restriction (4.2) needs no other modification than one extra block in the posterior sampler that draws from the conditional posterior of the γ_{ij}'s. Finally, unlike other proposed stochastic search variable selection algorithms for VAR models (George et al. 2008, Korobilis, 2008), this form of variable selection may be adopted in many nonlinear extensions of the VAR models.

In fact, in this paper I show that variable selection is very easy to adopt in the non-linear and richly parameterized, time-varying parameters vector autoregression (TVP-VAR). These models are currently very popular for measuring monetary policy, see for example Canova and Gambetti (2009), Cogley and Sargent (2005), Cogley et al. (2005), Koop et al. (2009) and Primiceri (2005). Common feature of these papers is that they all fix the number of autoregressive lags to 2 for parsimony. That is because marginal likelihoods are difficult to obtain in the case of time-varying parameters. Therefore, automatic variable selection is a convenient and fast way to overcome the computational and practical problems associated with mod-

els where parameters drift at each point in time. Although the methods described in this paper can be used for structural analysis (by providing data-based restrictions on parameters, useful for identifying monetary policy for instance), the aim is to show how more parsimonious models can be selected with positive impact in macroeconomic forecasting.

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

4.2 Variable selection in vector autoregressions

The standard VAR model

To allow for different equations in the VAR to have different explanatory variables, rewrite equation (4.1) as a system of seemingly unrelated regressions (SUR)

$$y_t = z_t\beta + \varepsilon_t \tag{4.3}$$

where $z_t = I_m \otimes x_t'$ is a matrix of dimensions $m \times n$, $\beta = vec(B)$ is $n \times 1$, and $\varepsilon_t \sim N(0, \Sigma)$. When no parameter restrictions are present in equation

(4.3), this model will be referred to as the unrestricted model. Bayesian variable selection is incorporated by defining and embedding in model (4.3) indicator variables $\gamma = (\gamma_1, ..., \gamma_n)'$, such that $\beta_j = 0$ if $\gamma_j = 0$, and $\beta_j \neq 0$ if $\gamma_j = 1$. These indicators γ are treated as random variables by assigning a prior on them, and allowing the data likelihood to determine their posterior values. We can explicitly insert these indicator variables multiplicatively in the model[1] using the following form

$$y_t = z_t\theta + \varepsilon_t \tag{4.4}$$

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

The Gibbs sampler provides a natural framework to estimate these parameters, by drawing sequentially from the conditional posterior of each parameter. In fact, sampling the restriction indices γ just adds one more block to the Gibbs sampler of the unrestricted VAR model[2]. For example, the full conditional (i.e. conditional on the data and Γ) densities of β and Σ are of standard form, assuming the so-called independent Normal-Wishart prior. For the restriction indicators we need to sample the n elements in the column vector $\gamma = (\gamma_1, ..., \gamma_n)'$, and then recover the diagonal matrix

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

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

$\Gamma = diag\{\gamma_1, ..., \gamma_n\}$ only when computations require it. Derivations are simplified if the indicators γ_j are independent of each other for $j = 1, ..., n$, i.e. $p(\gamma) = \prod_{j=1}^{n} p(\gamma_j) = \prod_{j=1}^{n} p(\gamma_j|\gamma_{\setminus -j})$, where $_{\setminus -j}$ indexes all the elements of a vector but the $j-th$, so that a conjugate prior for each γ_j is the independent Bernoulli density. In particular, define the priors

$$\beta \sim N_n(b_0, V_0) \tag{4.5}$$

$$\gamma_j|\gamma_{\setminus -j} \sim Bernoulli(1, \pi_{0j}) \tag{4.6}$$

$$\Sigma^{-1} \sim Wishart\left(\alpha, S^{-1}\right) \tag{4.7}$$

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

Exact expressions for the conditional densities of the parameters are provided in Appendix D. Here I provide a pseudo-algorithm which demonstrates that the algorithm for the restricted model (4.4) actually adds only one block which samples γ, in the standard algorithm of the unrestricted VAR model (4.3).

Bayesian Variable Selection Algorithm

1. Sample β as we would do in the unrestricted VAR in (4.3), but conditional on data being Z_t^*, with $Z_t^* = Z_t\Gamma$.

2. Sample each γ_j conditional on $\gamma_{\backslash -j}$, β, Σ and the data from

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

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

$$l_{0j} = p\left(y|\theta_j, \Sigma, \gamma_{\backslash -j}, \gamma_j = 1\right)\pi_{0j} \quad (4.8)$$

$$l_{1j} = p\left(y|\theta_j, \Sigma, \gamma_{\backslash -j}, \gamma_j = 0\right)(1 - \pi_{0j}) \quad (4.9)$$

3. Sample Σ as in the unrestricted VAR in (4.3), where now the mean equation parameters are $\theta = \Gamma\beta$.

In this type of model selection, what we care about is which of the parameters θ_j are equal to zero, so that identifiability of β_j and γ_j plays no role. In a Bayesian setting identifiability is still possible, since if the likelihood does not provide information about a parameter, its prior does. When $\beta_j = 0$ then γ_j is identified by drawing from its prior: notice that in this case in equations (4.8) - (4.9) it holds that $p\left(y|\theta_j, \gamma_{\backslash -j}, \gamma_j = 1\right) = p\left(y|\theta_j, \gamma_{\backslash -j}, \gamma_j = 0\right)$, so that the posterior probability of the Bernoulli density, $\widetilde{\pi}_j$, will be equal to the prior probability π_{0j}. Similarly, when $\gamma_j = 0$ then β_j is identified from the prior: the j-th column of $z_t^* = z_t\Gamma$ will be zero, i.e. the likelihood provides no information about β_j, and drawing from the posterior of β_j collapses to getting a draw from its prior, i.e. $\widetilde{b}_j = b_{0j}$ and

$\widetilde{V}_{jj} = V_{0,jj}$[3]. Nevertheless, in both of the above cases the result of interest is that $\theta_j = 0$, whether because $\beta_j = 0$ or because $\gamma_j = 0$, and the respective parameter is restricted.

There are several other approaches to automatic Bayesian model selection for regression models which can be generalized to VAR models. Most of them are based on introducing and sampling indicator variables γ as we saw above. For the specific case of the simple VAR the restriction search proposed in Algorithm 1 is computationally more intensive than other approaches, like the variable selection algorithm of George et al. (2008); see also Chapter 3. Nevertheless, Algorithm 1 can be easily adopted in the case of nonlinearity in the parameters, or specifications which admit non-conjugate priors. In that respect, the remainder of this paper develops a useful extension, namely model selection in time-varying parameters VARs. Since a different value of the parameters at each time period t occurs in these models, they tend to be non-parsimonious representations of macroeconomic data. Additionally, marginal likelihood calculations may be difficult to obtain (at least in terms of computer time). In this case, variable selection offers a very easy and fast method for selection of lag length and/or exogenous predictors.

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

Time-varying parameters VAR model

Modern macroeconomic applications increasingly involve the use of VARs with mean regression coefficients and covariance matrices which are time-varying. In fact, Granger (2008) shows that any nonlinear econometric model can be cast as a special case of a time-varying parameters model. Nonetheless, forecasting with time-varying parameters VARs is not a new topic in economics. During the "Minnesota revolution" efficient approximation methods of forecasting with TVP-VARs were developed, with most notable contributions the ones by Doan et al. (1984) and Sims (1989); for a large-scale application in an 11-variable VAR see Canova (1993). However, the development of accurate Bayesian sampling methods through the '90s (Gibbs sampler), combined with modern computing power has resulted in a recent rising interest in forecasting structural instability using time-varying parameters models. Canova and Ciccarelli (2004), Clark and McCracken (2010) and D'Agostino et al. (2009) are examples of forecasting multiple time-series using TVP-VAR's, while Stock and Watson (2007), Groen et al. (2009) and Koop and Korobilis (2009b) are focusing on univariate predictions but with the use of a large set of exogenous variables.

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

$$y_t = z_t \beta_t + \varepsilon_t \tag{4.10}$$

$$\beta_t = \beta_t + \eta_t \tag{4.11}$$

where $z_t = I_m \otimes x_t'$ is an $m \times n$ matrix, β_t is an $n \times 1$ vector of parameters

for $t = 1, ..., T$, $\varepsilon_t \sim N(0, \Sigma)$ with Σ an $m \times m$ covariance matrix, and $\eta_t \sim N(0, Q)$ with Q an $n \times n$ covariance matrix. Models of the form described above, and variations of it, have extensively been used for structural analysis (cite papers) and forecasting (cite papers). Obviously, the model in (4.10) is not a parsimonious representation of our data, and in practice most of the studies mentioned rely on quarterly data while using 2 lags of the dependent variable in order to prevent the proliferation of parameters and the estimation error.

Variable selection in this model is a simple extension of the VAR model with constant parameters[4]. For that reason replace (4.10) with

$$y_t = z_t \theta_t + \varepsilon_t \tag{4.12}$$

where, as before, $\theta_t = \Gamma \beta_t$ and Γ is the $n \times n$ matrix defined in (4.4). For this model, the priors on Σ and γ_j are the same as in the VAR case, i.e. $\Sigma^{-1} \sim Wishart(\alpha, S^{-1})$ and $\gamma_j | \gamma_{\setminus -j} \sim Bernoulli(1, \pi_{0j})$ respectively. For the time varying parameters, a prior on the initial condition is necessary which is of the form $\beta_0 \sim N_n(b_0, V_0)$. The random walk evolution of β_t, it would be desirable to restrict their prior variance in order to avoid explosive behavior. The (implied) priors for β_1 to β_T are provided by the state equation (4.11), and they are of the form $\beta_t | \beta_{t-1}, Q \sim N\left(\beta_{t-1}, Q\right)$. The covariance matrix Q is considered to be unknown, so it will have its own prior of the form

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

$Q^{-1} \sim Wishart\,(\xi, R^{-1})$. A Gibbs sampler for the unrestricted TVP-VAR model exists, so that sampling from the TVP-VAR with model selection, requires only one extra block which samples γ, in the spirit of Algorithm 1 of the previous section. Full details are provided in Appendix D.

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

In that respect, in this forecasting exercise I use a Minnesota-based prior to elicit the prior hyperparameters of the TVP-VAR model. This results to an Empirical Bayes prior that can be tuned using the full sample, without the need to waste useful observations in a training sample. In order to avoid explosive draws, I subjectively choose the hyperparameters for the initial condition β_0 and the covariance matrix Q, in order to get a very

tight prior. This prior allows to overcome stationarity restrictions which make the MCMC sampler inefficient. The reader should note that I use only a single choice of hyperparameters for the Minnesota prior, without searching and comparing other choices for the TVP-VAR model. The main purpose of the paper is to compare the unrestricted to the restricted (with variable selection) model, so only one benchmark prior is used for the shake of this comparison. Nevertheless, examining the forecasting performance of different priors in models with time-varying parameters is a challenging, but very important idea for future research.

Prior elicitation for variable selection

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

informative. For the purpose of comparison, I use a benchmark Minnesota-type prior.

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

4.3 Simulated numerical examples

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

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

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

with a constant and one lag and constant parameters[6]

$$B = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \Psi = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{4.13}$$

where it holds that $\Sigma = (\Psi\Psi')^{-1}$. 100 samples of size $T = 50$ are generated, using a random initial condition for the dependent variables $y_{-1} \sim U(0, 1)$. Hence 100 VAR models are estimated using the generated samples, saving 30.000 draws from the posterior of the estimated parameters after discarding an initial 20.000 draws to ensure convergence. The prior hyperparameters (see equations (4.5) - (4.7)) are set to be uninformative: $b_0 = 0_{n\times 1}$, $V_0 = 9 \times I_n$, $\pi_{0j} = 0.5$ for all $j = 1, ..., n$, $\alpha = 0$ and $S = 0 \times I_m$. The intercepts are always included in each model, so that the variable selection applies only to lags of the dependent variables.

The unrestricted VAR is just a special case of variable selection, where we impose all $\gamma_j = 1$. Thus, a fully unrestricted estimate of B can be obtained using the variable selection priors and imposing always Γ to be the identity matrix I of dimensions $n \times n$. The average over the 100 samples

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

of the posterior mean (posterior standard deviations of non-zero parameters are in parentheses) of B using the unrestricted prior is

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

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

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

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

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

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

George et al. (2008) have used the exact same setup to evaluate the efficiency of the SSVS algorithm described in Section 3.1. Comparing the matrix of restrictions, $\widehat{\gamma}$, reported above, with their equivalent matrix it is obvious that their reported probabilities of inclusion of the parameters are more decisive. In their case, the highest probability which a parameter which should be restricted gets is equal to 0.5 in only a few cases. This happens because the SSVS restricts a parameter if it is too low[7], while variable selection in this paper requires a parameter to be restricted exactly to zero. However, using both algorithms model selection implies that the optimal predictive model is the one which has parameters with probability of inclusion higher than 0.5 (see Barbieri and Berger (2004) for a proof).

Example 2: Homoskedastic TVP-VAR. In the second example, 100 samples of size $T = 100$ are generated from a 4-variable Homoskedastic TVP-VAR with one autoregressive lag (no constant). The covariance matrix

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

Σ is set equal to the upper left 4x4 block of the covariance matrix specified in Example 1 (in equation (4.13)), and $\beta_t = vec(B_t)$ is let to evolve according to the random walk specification (4.11), by setting initial condition, $B_0 = \{B_t\}_{t=0}$, and a simple diagonal covariance matrix Q, of the form

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

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

The usual practice in the TVP-VAR models is to use tight data-based priors. However, for the purpose of this exercise relatively uninformative

priors are defined

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

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

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

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

4.4 Macroeconomic Forecasting with VARs

Data and set-up

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

The data are obtained from the Office for National Statistics (ONS) website, http://www.statistics.gov.uk/. The available sample runs from 1971Q1 to 2008Q4. Inflation, unemployment and interest rates are measured on a monthly basis. Quarterly series are calculated by the ONS by taking averages over the quarter (for inflation), the value at the mid-month

of the quarter (for unemployment), and the value at the last-month of the quarter (for interest rate), respectively. The data are plotted in Figure 4.1.

The VAR and Homoskedastic TVP-VAR models include a constant and a maximum of 2 lags of the dependent variables. One could argue that the use of variable selection would allow to define a higher maximum lag length, say 4 lags, and then let the data decide on the optimal number of lags in each VAR equation. However, given the fact that the total observations are only 152, we would be asking too much from the variable selection algorithm in a recursive forecasting exercise. The forecast horizons used for comparison are $h = 1$,4 and 8. The sample 1971:Q1 - 1988:Q4 is used for initial estimation, and the forecasts are computed recursively with expansion of the estimation sample each quarter. Subsequently for the period 1989:Q1 - 2008:Q4 we obtain a total of 80-h forecasts.

Iterated forecasts are obtained by estimating the models (4.4) and (4.10), writing the models in companion (VAR(1)) form, and iterating the forward up to $h = 8$ periods ahead in order to obtain $[\widehat{y}_{T+1}, ..., \widehat{y}_{T+h}]$, where T is the last observation of the sample. Direct forecasts are obtained from the VAR and the TVP-VAR specifications in (4.4) and (4.10) respectively by estimating separately for $h = 1$,4 and 8 the models with the dependent variable y_t replaced by y_{t+h}, while the R.H.S. variables are still measured up to, and including, time t. When $h = 1$ direct and iterated forecasts are exactly the same, since we are estimating and forecasting with exactly the same specifications.

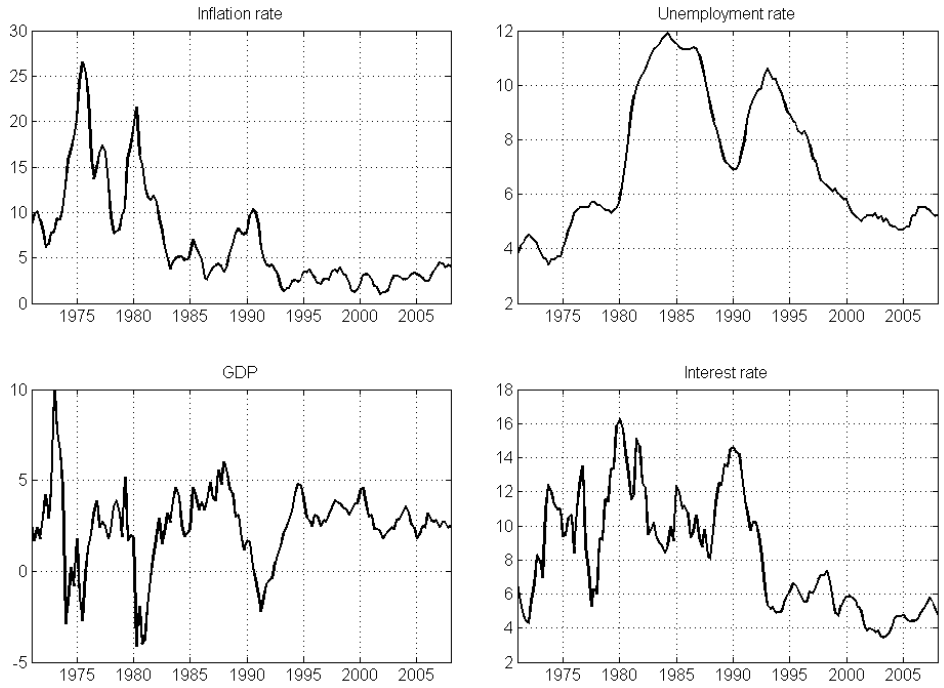

Figure 4.1: Graph of the data for UK inflation, unemployment, GDP, and interest rate.

When forecasting, the parameters of the VAR model remain constant in the out-of-sample period. However this is not the case for the autoregressive coefficients of the TVP-VAR model. The iterative nature of the Gibbs sampler allows to easily simulate the out-of-sample path of β_t. At each iteration, conditional on obtaining a draw of β_T and the covariances Q, we can use the random walk evolution equation (4.11) to simulate $\left[\widehat{\beta}_{T+1}, ..., \widehat{\beta}_{T+h}\right]$ in a recursive manner. Then conditional on knowing these parameters, calculation of direct or iterated forecasts can be computed as in the VAR case, separately for each out-of-sample period.

Forecasting models

Dependent on the model specification and the choice of prior hyperparameters, there are 5 competing forecasting models. Common place in all models, in order to evaluate the performance of variable selection of mean equation coefficients, is that the covariance matrix is integrated out using an uninformative prior of the form $p(\Sigma) \propto |\Sigma|^{-(m+1)/2}$ which is equivalent to the Wishart prior defined in (4.7) with the additional restriction that $\alpha = 0$ and $S^{-1} = 0_{m \times m}$.

The 5 models are

1. VAR with variable selection (VAR VS)

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

2. VAR with Minnesota prior (VAR MIN)

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

$$V_{i,l}^{MIN} = \begin{cases} g_1/p & \text{for parameters on own lags} \\ g_3/s_i^2 & \text{for intercepts} \\ \frac{g_2 s_i^2}{p s_l^2} & \text{for parameters } j \text{ on variable } l \neq i, l, i = 1, .., m \end{cases} \tag{4.14}$$

   Here s_i^2 is the residual variance from the p-lag univariate autoregression for variable i. The prior mean vector b_{MIN} is set equal to 0.9 for parameters on the first own lag of each variable and zero otherwise.

The hyperparameters are set to the values $g_1 = 0.5$, $g_2 = 0.005$ and $g_3 = 100$.

3. Benchmark VAR

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

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

   The initial condition is set to $\beta_0 \sim N(0, 4^2 V_{MIN})$, and $\gamma_j|\gamma_{\setminus -j} \sim Bernoulli(1, 0.5)$. The covariance Q of the varying coefficients has the prior $Q^{-1} \sim Wishart(\xi, R)$ where $\xi = n+1$ and $R^{-1} = 0.001(n+1)V^{MIN}$, where V^{MIN} is the matrix defined in (4.14) above.

5. Benchmark TVP-VAR

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

The priors for the VAR are fairly uninformative, however the TVP-VAR prior is quite tight. Alternatively we can assign to β_0 a large variance, say $100I$, on the basis that this is a desirable uninformative choice. However doing so, means that we increase the probability that the whole sequence of draws for β_t will be in the nonstationary region. This approach can be computationally cumbersome as many draws may be required as in the case of Cogley et al. (2005) who use 100.000 draws, discard the first 50.000 and save every 10-th draw. Given the dimension of the parameter space and

the compuational demands of a recursive forecasting exercise, the informative variance $4^2 V_{MIN}$ on the initial state is used here in order enhance the efficiency of the Gibbs sampler. All models are based on a run of 20.000 draws from the posterior, discarding the first 10.000 draws.

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

Forecast evaluation

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

$$\begin{aligned} MSFE_{i,t}^{h} &= \sqrt{\left(\widehat{y}_{i,t+h|t} - y_{i,t+h}^{o}\right)^2} \\ MAFE_{i,t}^{h} &= \left|\widehat{y}_{i,t+h|t} - y_{i,t+h}^{o}\right| \end{aligned}$$

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

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

where τ_0 is 1989:Q1 and τ_1 is 2008:Q4.

In-sample variable selection results

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

The first thing to observe from Tables 1 and 2 is that for both models and for all horizons variable selection imposes many restrictions. This result is not surprising, both from an empirical and theoretical point of view. The most recent own lag of each variable is important in most cases for all forecast horizons. Other than that, variable selection indicates only a few extra variables as important in each VAR equation, leading to quite parsimonious models. This pattern complies with the empirical results of Korobilis (2008) and Jochmann et al. (2009) using the SSVS algorithm for VAR models (see Section 3.3 above).

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

A comparison of the parameters of the VAR models with the respective parameters of the TVP-VAR models, reveals quite a few differences, but also many similarities at the same time, as to which variables are selected to enter the "best" model. For example, in the VAR strong (probability equal to 1) predictors for 1-quarter ahead inflation ($\Delta\pi_{t+1}$) are current inflation ($\Delta\pi_t$) and interest rate(r_t), as well as inflation in the previous quarter ($\Delta\pi_{t-1}$). In the TVP-VAR it is only $\Delta\pi_t$ and $\Delta\pi_{t-1}$ which are selected, and the current level of interest rate has only probability of 0.28.

In the same equation, there is weaker evidence that gdp_t, $u_{t-1}$and r_{t-1} are good predictors, which vanishes in the TVP-VAR case (for example r_{t-1} has a probability of 0.61 of entering the VAR model, but only a probability of 0.41 of entering the TVP-VAR model). Similar inference can be made for the rest of variables and equations.

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

Table 4.1: Average of posterior of restrictions γ for $h = 1, 4, 8$ (VAR model)

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

Table 4.2: Average of posterior of restrictions γ for $h = 1, 4, 8$ (TVP-VAR model)

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

Out-of-sample forecasting results

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

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

For short-term forecasts ($h = 1$) the multivariate VAR models, whether restricted or unrestricted, offer more accurate forecasts compared to the

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

While MSFE and MAFE measures are very informative in our case, since the purpose is just to evaluate point forecasts, full predictive densities can be compared using predictive likelihoods. In fact predictive likelihoods

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

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

4.5 Concluding remarks

Vector autoregressive models have been used extensively over the past for the purpose of macroeconomic forecasting, since they can fit the observed data better than competing theoretical and large-scale structural macro-

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

Table 4.3: Forecast evaluation, $h = 1$

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

Table 4.4: Forecast evaluation, $h = 4$

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

Table 4.5: Forecast evaluation, $h = 8$

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

econometric models. Nowadays, Bayesian dynamic stochastic general equilibrium (DSGE) models like the one of Smets and Wouters (2003) have been shown to challenge the forecasting performance of unrestricted VAR models, while at the same time having all the advantages of being structural, i.e. connected to economic theory. While DSGE models provide restrictions based on theory, this paper shows that Bayesian variable selection methods can be used to find restrictions based on the evidence in the data, and at the same time improve over the forecasts of unrestricted VAR models as well. Additionally, Bayesian variable selection methods for vector autoregressions can be used for structural analysis, like measuring monetary policy shocks in identified VARs. A different route for VAR variable selection algorithms

would be to uncover empirically the relationship between variables, which could potentially help in the development of new theoretical relationships.

Chapter 5

Conclusion

5.1 Summary & policy implications

Nowadays, Central Banks and policy makers monitor hundreds of variables during the decision process (Bernanke and Boivin, 2003). Additionally, it is currently recognized that forward-looking expectations are very important during the price setting behavior of agents (Rudd and Whelan, 2007), which results in an increasing importance of accurate forecasts of economic fundamentals on behalf of inflation-targeting Central Banks. Another strand of literature has identified a large decrease in volatility (persistence) in most macroeconomic variables of many developed countries (most notably the US; see Giannone, Lenza and Reichlin, 2008, for a review), which suggests that economic relationships are far from being constant over the course of the last years.

This thesis deals with all these three important modeling issues using empirically and computationally attractive methods. In three distinct - but

methodologically interconnected - settings I show how model selection can i) be adopted to models which capture nonlinearities in macroeconomic relationships, ii) improve the forecasting performance of econometric models by preserving parsimony, iii) select the most relevant indicators for policy-making (among a set of hundreds of variables), and iv) provide multivariate forecasts in cases where the limited number of macroeconomic time-series observations would otherwise not allow econometric estimation to be implemented in the first place.

Chapters 2 & 4 develop two methods to implement model selection in models with time-varying parameters and volatility. Time-varying parameters models are very popular in modern macroeconomics as they are able to capture many important features of the observed data (Cogley and Sargent, 2005). For that reason in Chapter 2 I extend the New Keynesian Phillips curve regression with drifting parameters, but I additionally allow the probabilities of inclusion of predictor variables to be drifting as well. This proves to be a feature supported by the data, since the relationship between inflation and many traditional predictor variables has changed dramatically in the post WWII era. Inflation itself has changed as well and it is regarded as more persistent since the mid-80s (Stock and Watson, 2007). Allowing the model relevant for prediction to change over time is thus a very prominent contribution, which has not been examined before in the literature.

In Chapter 4 I use a multivariate model, namely the time-varying parameters VAR (TVP-VAR). The methods of Chapter 2 can be extended in a straightforward way in the case of the TVP-VAR. Nevertheless, due to the high dimensionality of this model I develop a simple model selection method

which assumes that a predictor is either relevant for all the periods in the sample, or it is not relevant at all (and its coefficient shrinks to zero). The algorithm is computationally simple to adopt and doesn't add unreasonably too much computer time in the estimation of the (already computationally demanding) TVP-VAR model.

In a context without structural instabilities and nonlinearities, but with a large information set available, Chapter 3 addresses the issue of forecasting variables of interest to Central Banks (and the general public) using factor methods. A stochastic search variable selection algorithm is used successfully to select among hundreds of predictors, and preserve degrees of freedom. The results suggest that the benefits in forecasting are large.

5.2 Further research

Given the promising results of this thesis, there are many aspects of practical Bayesian variable selection which are relevant for macroeconometric applications (as well as other fields in economics). For instance, Korobilis and Moretti (under preparation) introduce dynamic model averaging in a nowcasting problem using dynamic factor models (see Giannone, Reichlin and Small, 2008). The DMA methodology can provide the additional flexibility of allowing different number of factors to be selected at each time series observation, and thus is not restricted by keeping the optimal number of factors constant for the full sample.

Other than selecting variables/models relevant for forecasting, Bayesian model selection has been used in numerous other applications involving

flexible modeling. Using appropriate modifications one can use variable selection of the form of Chapter 4 to select whether a parameter is constant or time-varying, or obtain a degree of the "amount" of variation in time-varying parameters models; see the Appendix of Korobilis (2009b) for a brief review and Korobilis (2009a).

Nevertheless, an important field which might benefit from Bayesian model selection is the modern Dynamic Stochastic General Equilibrium (DSGE) models. These models have benefited extremely from Bayesian sampling methods, and the fact that parameters can be identified from the priors (see An and Schorfheide, 2007). However, prior elicitation in these models is sometimes driven from the need of identification. It is also the case that some parameters have to be calibrated, a procedure that is completely subjective and has been the matter of criticism for decades in economics. Using priors which data-based restrictions is an attractive alternative that should definetely be explored in the future.

Bibliography

[1] An, S. and Schorfheide, F. (2007). Bayesian Analysis of DSGE Models. *Econometric Reviews*, 26, 113-172.

[2] Andersson, M. K. and Karlsson, S. (2008). Bayesian forecast combination for VAR models. In: S. Chib, W. Griffiths, G. Koop & D. Terrell (Eds), *Bayesian Econometrics* (pp. 501–524). Advances in Econometrics, 23,

[3] Ang, A. Bekaert, G. and Wei, M. (2007). Do macro variables, asset markets, or surveys forecast inflation better?, *Journal of Monetary Economics* 54, 1163-1212.

[4] Atkeson, A. and Ohanian, L. (2001). Are Phillips curves useful for forecasting inflation?, *Federal Reserve Bank of Minneapolis Quarterly Review* 25, 2-11.

[5] Avramov, D. (2002). "Stock return predictability and model uncertainty," *Journal of Financial Economics* 64, 423-458.

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

[7] Berger, J. O. and Molina, G. (2004). Some recent developments in Bayesian variable selection. In R. Fischer, R. Preuss and U. von Toussaint (Eds.). *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, AIP Conference Proceedings 735, 417–428.

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

[9] Bernanke, B. S. and Boivin, J. (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics*, 50, 525–546.

[10] Bernanke, B. S., Boivin, J. and Eliasz, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *Quarterly Journal of Economics*, 120, 387–422.

[11] Brown, P. J., Vannucci, M. and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60, 627–641.

[12] Brown, P. J., Vannucci, M. and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64, 519–536.

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

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

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

[16] Carter, C., and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81, 541–553.

[17] Chen, R. and Liu, J., "Mixture Kalman filters," *Journal of the Royal Statistical Society, Series B* 62 (2000), 493-508.

[18] Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96, 270-281.

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

[20] Ciccone, A. and Jarocinski, S. (2007). Determinants of Economic Growth: Will Data Tell?. CERP working paper 6544 .

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

[22] Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science*, 19, 81-94.

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

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

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

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

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

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

[29] Fagin, S. (1964). Recursive linear regression theory, optimal filter theory, and error analyses of optimal systems. *IEEE International Convention Record Part i*, 216-240.

[30] Favero, C. A., Marcellino, M. and Neglia, F. (2005). Principal components at work: The empirical analysis of monetary policy with large datasets. *Journal of Applied Econometrics*, 20, 603–620.

[31] Fernandez, C., Ley, E. and Steel, M. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100, 381–427.

[32] George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association*, 95, 1304-1308.

[33] George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, 87, 731–747.

[34] George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88, 881–889.

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

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

[37] Geweke, J. and Amisano, G. (2007). Hierarchical Markov Normal mixture models with applications to financial asset returns," Manuscript available at

http://www.biz.uiowa.edu/faculty/jgeweke/papers/paperA/paper.pdf, 2007.

[38] Giannone, D., Lenza, M. and Reichlin, L. (2008). Explaining the great moderation: It is not the shocks. *Journal of the European Economic Association*, 6, 621-633.

[39] Giannone, D., Reichlin, L. and Sala, L. (2004). Monetary policy in real time (and comments). In: M. Gertler & K. Rogoff (Eds), *NBER Macroeconomics Annual 2004* (pp. 161–224). Cambridge: The MIT Press.

[40] Giannone, D., Reichlin, L. and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55, 665-676.

[41] Gilks, W. R., Richardson, S. and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in practice*. Chapman and Hall.

[42] Granger, C. W. J. (2008). Non-linear models: Where do we go next - time varying parameter models?. *Studies in Nonlinear Dynamics & Econometrics*, 12, 1-34.

[43] Green, P. (1995). Reversible jump Markov Chain Monte Carlo Computation and Bayesian model determination. *Biometrika* 82, 711-732.

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

[45] Hoeting, J., Madigan, D., Raftery, A. and Volinsky, C. (1998). Bayesian model averaging: A tutorial. *Statistical Science*, 14, 382–417.

[46] Jazwinsky, A. (1970). *Stochastic processes and filtering theory.* New York: Academic Press.

[47] Jochmann, M., Koop, G. and Strachan, R.W. (2008). Bayesian forecasting using stochastic search variable selection in a VAR subject to breaks. Unpublished manuscript.

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

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

[50] Koop, G. and D. Korobilis. (2009a). Bayesian multivariate time series methods for empirical macroeconomics. RCEA Working Paper 47-09.

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

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

[53] Koop, G. and Potter, S. M. (2004). Forecasting in dynamic factor models using Bayesian model averaging. *Econometrics Journal*, 7, 550-565.

[54] Koop, G. and Potter, S. M. (2008). Time-varying VARs with inequality restrictions. Unpublished manuscript.

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

[56] Korobilis, D. (2009a). Assessing the transmission of monetary policy shocks using dynamic factor models. Discussion Paper 9-14, University of Strathclyde.

[57] Korobilis, D. (2009b). VAR forecasting using Bayesian variable selection. MPRA Paper 21124, University Library of Munich, Germany.

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

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

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

[61] Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review* 63, 215-232.

[62] Merl, D., Lucas, J.E., Nevins, J.R., Shen, H. and West, M. (2010). Trans-study projection of genomic biomarkers using sparse factor re-

gression models. forthcoming in the *The Handbook of Applied Bayesian Analysis.*

[63] Mittelhammer, R. C., Judge, G. G. and Miller, D. J. (2000). *Econometric foundations*. Cambridge: Cambridge University Press.

[64] O'Hara, R. B. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, 4, 85-118.

[65] Pesaran, M.H., Pick, A., and A. Timmermann. (2009). Variable selection and inference for multi-period forecasting problems. Cambridge Working Papers in Economics 0901, Faculty of Economics, University of Cambridge.

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

[67] Raftery, A.E. (1995). Bayesian model selection in social research (with Discussion). *Sociological Methodology*, 25, 111-196.

[68] Raftery, A., Karny, M., Andrysek, J. and Ettler, P. (2007). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. Technical report 525, Department of Statistics, University of Washington.

[69] Rudd, J. and Whelan, K. (2007). Modeling inflation dynamics: A critical review of recent research. *Journal of Money, Credit and Banking*, 39, 155-170.

[70] 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.

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

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

[73] Smets, F., and R. Wouters. (2003). An estimated Dynamic Stochastic General Equilibrium model of the Euro-area. *Journal of the European Economic Association*, 1, 1123-1175.

[74] Smith, M., and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75, 317–343.

[75] Smith, M., and Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. J*ournal of the American Statistical Association*, 97, 1141–1153.

[76] Smith, J. and Miller, J. (1986). A non-Gaussian state-space model and application to prediction records. *Journal of the Royal Statistical Society, Series B*, 48, 79-88.

[77] Stock, J. H. and Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. J*ournal of Business and Economic Statistics*, 14, 11-30.

[78] Stock, J. H. and Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, 44, 293-335.

[79] Stock, J. H. and Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics*, 20, 147–162.

[80] Stock, J. H. and Watson, M. W. (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, 41, 788–829.

[81] Stock, J. H. and Watson, M. W. (2005a). Forecasting with many predictors. Unpublished Manuscript. Princeton University, Princeton, NJ (prepared for The Handbook of Economic Forecasting).

[82] Stock, J. H. and Watson, M. W. (2005b). Implications of factor models for VAR analysis. Unpublished Manuscript. Princeton University, Princeton, NJ.

[83] Stock, J. H. and Watson. M. W. (2007). Why has U.S. inflation become harder to forecast?. *Journal of Money, Credit and Banking*, 39,3-33.

[84] Stock, J. H. and Watson, M. W. (2008). Phillips curve inflation forecasts. NBER Working Paper No. 14322.

[85] Strachan, R. W. and van Dijk, H. K. (2007). Bayesian model averaging in vector autoregressive processes with an investigation of stability of the US great ratios and risk of a liquidity trap in the USA, UK and

Japan. *Econometric Institute Report* EI 2007-11. Erasmus University Rotterdam, the Netherlands.

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

[87] Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44, 92-107.

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

[89] Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, 100, 1215–1225.

Appendices

A Data Appendix (Chapter 2)

The variables used in this study are described in the table below. All series were seasonally adjusted, where applicable, and run from 1959:Q1 to 2008:Q2. Some series in the database were observed 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. 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, $x_{i,t} = z_{i,t} - z_{i,t-1}$; 4 - logarithm, $x_{i,t} = \log z_{i,t}$; 5 - first difference of logarithm, $x_{i,t} = \log z_{i,t} - \log z_{i,t-1}$.

#	Mnemonic	Tcode	Description
1	GDPDEFL	5	Gross Domestic Product: Implicit Price Deflator
2	CPI	5	Consumer Price Index For All Urban Consumers
3	UNEMP	1	Civilian Unemployment Rate
4	CONS	5	Real Personal Consumption Expenditures
5	INV	5	Private Residential Fixed Investment
6	GDP	5	Real Gross Domestic Product, 3 Decimal
7	HSTARTS	4	Housing Starts: Total Units Started
8	EMPLOY	5	All Employees: Total Private Industries
9	PMI	2	ISM Manufacturing: PMI Composite Index
10	COMPRICE	2	NAPM Commodity Prices Index (Percent)
11	VENDOR	2	NAPM Vendor Deliveries Index (Percent)
12	WAGE	5	Average Hourly Earnings: Manufacturing
13	TBILL	1	3-Month Treasury Bill: Secondary Market Rate
14	SPREAD	1	Spread 10-year / 3-month rate (GS10 -TB3MS)
15	DJIA	5	Dow Jones Industrial Average
16	MONEY	5	M1 Money Stock
17	INFEXP	1	University of Michigan Inflation Expectations

B Technical Appendix (Chapter 3)

A Gibbs sampler for SSVS in VAR models

The priors described in Section 3 combined with the likelihood function of a VAR model, will allow us to derive and draw from the full conditional distributions. The likelihood of the VAR model $y = z\phi + \varepsilon$, $\varepsilon \sim N(0, \Sigma)$ with $\Sigma^{-1} = \Psi'\Psi$, is

$$\begin{aligned} L(y|\phi, \Psi) &\propto |\Psi|^{-T} \exp\left\{-\frac{1}{2} tr\left[\Psi'(y - z\phi)'(y - z\phi)\Psi\right]\right\} \\ &= |\Psi|^{-T} \exp\left\{-\frac{1}{2}\left(\phi - \widehat{\phi}\right)'\left[\Psi\Psi' \otimes (z'z)\right]\left(\phi - \widehat{\phi}\right)\right. \\ &\quad \left. -\frac{1}{2} tr\left[\left(y - z\widehat{\phi}\right)'\Psi'\Psi\left(y - z\widehat{\phi}\right)\right]\right\} \end{aligned}$$

where $\widehat{\phi}$ is the MLE of ϕ. This form of the likelihood function allows to derive the posterior of the ϕ parameters. In order to derive the posterior of the elements of Ψ we need to first rewrite the likelihood function in convenient form. Define $S(\phi) = (y - z\phi)'(y - z\phi)$ and write $S(\phi) = s_{ij}$. For $j = 2, ..., m$ define the $(m-1)$ vectors $s_j = \left(s_{1j}, ..., s_{(j-1)j}\right)'$ containing the upper diagonal elements of $S(\phi)$, and the $(m-1)$ matrices S_j containing the upper left $j \times j$ submatrix of $S(\phi)$. Define also $\rho_1 = s_{11}$ and $\rho_i = |S_i| / |S_{i-1}| = s_{ii} - s_i' S_{i-1}^{-1} s_i$ for $i = 2, ..., m$. The likelihood function now cam take the following form

$$\begin{aligned} L(y|\phi, \Psi) &\propto \prod_{i=1}^{m} (\psi_{ii})^T \\ &\quad \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{m} \psi_{ii}^2 \rho_i + \sum_{j=2}^{m} \left(\eta_j + \psi_{jj} S_{j-1}^{-1} s_j\right)' S_{j-1}\left(\eta_j + \psi_{jj} S_{j-1}^{-1} s_j\right)\right]\right\} \end{aligned}$$

Now define $D = diag\left\{h_1, ..., h_{n_\varphi}\right\}$ with

$$h_i = \begin{cases} \tau_{0i}, \text{ if } \gamma_i = 0 \\ \tau_{1i}, \text{ if } \gamma_i = 1 \end{cases}, \text{ for } i = 1, ..., n_\varphi$$

and, similarly, define $D_j = diag\left\{h_{1j}, ..., h_{(j-1)j}\right\}$ with

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

for $i = 1, ..., j$ and $j = 2, ..., m$. Then we can rewrite equations (3.7) and (3.11) in the main text, as

$$\begin{aligned} \varphi_i^k|\gamma &\sim N\left(0, DD\right) \\ \eta_j|\omega_j &\overset{iid}{\sim} N_{j-1}\left(0, D_j D_j\right) \end{aligned}$$

respectively. Denote the combined prior of the unrestricted coefficients φ^c and the restricted coefficients φ^k as $\varphi \sim N\left(\underline{\varphi}, \underline{V}\right)$. Given starting values, model parameters are drawn from their conditionals for $r = 1, ..., R$ iterations:

1. Draw $\left(\psi^{(r)}|\eta^{(r-1)}, \omega^{(r-1)}, \gamma^{(r-1)}, \varphi^{(r-1)}, data\right)$ by sampling each element from the Gamma distribution

$$\psi_{ii}^2 \sim Gamma\left(\alpha_i + \frac{1}{2}T, B_i\right)$$

where

$$B_i = \begin{cases} \beta_1 + \frac{1}{2}s_{11} & \text{for } i = 1 \\ \beta_i + \frac{1}{2}\left[s_{ii} - s_i'\left(S_{i-1} + \left(D_i D_i\right)^{-1}\right)^{-1} s_i\right] & \text{for } i = 2, ..., m \end{cases}$$

2. Draw $\left(\eta^{(r)}|\psi^{(r)}, \gamma^{(r-1)}, \varphi^{(r-1)}, \omega^{(r-1)}, data\right)$ by sampling each element from the Normal distribution

$$\eta_j \sim N_{j-1}\left(\mu_j, \Delta_j\right)$$

where for $j = 2, ..., m$.

$$\begin{aligned} \mu_j &= -\psi_{jj} \left\{ S_{j-1} + (D_j D_j)^{-1} \right\}^{-1} s_j \\ \Delta_j &= \left\{ S_{j-1} + (D_j D_j)^{-1} \right\}^{-1} \end{aligned}$$

3. Draw $\left(\omega^{(r)} | \eta^{(r)}, \psi^{(r)}, \gamma^{(r-1)}, \varphi^{(r-1)}, data\right)$ by sampling each element from the Bernoulli distribution

$$\omega_{ij} \sim Bernoulli\left(1, \frac{u_{1ij}}{u_{1ij} + u_{2ij}}\right)$$

where for $j = 2, ..., m$ and $i = 1, ..., j-1$

$$\begin{aligned} u_{1ij} &= \frac{1}{\kappa_{0ij}} \exp\left(-\frac{\psi_{ij}^2}{2\kappa_{0ij}^2}\right) q_{ij} \\ u_{2ij} &= \frac{1}{\kappa_{1ij}} \exp\left(-\frac{\psi_{ij}^2}{2\kappa_{1ij}^2}\right) (1 - q_{ij}) \end{aligned}$$

4. Draw $\left(\varphi^{(r)} | \eta^{(r)}, \psi^{(r)}, \omega^{(r)}, \gamma^{(r-1)}, data\right)$ by sampling $\varphi = vec(\phi)$ from the Normal distribution

$$\varphi \sim N_{n_u}(\mu, \Delta)$$

where

$$\begin{aligned} \mu &= \left\{ (\Psi\Psi') \otimes (z'z) + \underline{V}^{-1} \right\}^{-1} \left\{ ((\Psi\Psi') \otimes (z'z)) \widehat{\varphi} + \underline{V}^{-1} \underline{\varphi} \right\} \\ \Delta &= \left\{ (\Psi\Psi') \otimes (z'z) + \underline{V}^{-1} \right\}^{-1} \end{aligned}$$

where $\widehat{\varphi}$ is the vector occurring from stacking the elements of the matrix of MLE coefficients, i.e. $\widehat{\varphi} = vec\left(\widehat{\phi}\right) = vec\left((z'z)^{-1} z'y\right)$.

5. Draw $\left(\gamma^{(r)} | \eta^{(r)}, \psi^{(r)}, \omega^{(r)}, \varphi^{(r)}, data\right)$ by sampling each element from the Bernoulli density

$$\gamma_i \sim Bernoulli\left(1, \frac{u_{1i}}{u_{1i}+u_{2i}}\right)$$

1. where for $i = 1, ..., n_u$

$$\begin{aligned} u_{1i} &= \frac{1}{\tau_{0i}} \exp\left(-\frac{\varphi_i^2}{2\tau_{0i}^2}\right) p_i \\ u_{2i} &= \frac{1}{\tau_{1i}} \exp\left(-\frac{\varphi_i^2}{2\tau_{1i}^2}\right) (1-p_i)\,. \end{aligned}$$

C Data Appendix (Chapter 3)

This table lists the 132 variables in the dataset used. The third column indexes the respective transformation applied to each of the variables to ensure stationarity (at least approximately). Let v_t and x_t be the untransformed value and transformed values respectively, then there are five cases:(1) lv: $x_t = v_t$ (level), (2) ln: $x_t = \log(v_t)$ (logarithm), (3) Δlv: $x_t = v_t - v_{t-1}$ (first difference), (4) $\Delta \ln$: $x_t = \log(v_t/v_{t-1})$ (growth rate), and (5) $\Delta^2 \ln$: $x_t = \Delta \log(v_t/v_{t-1})$ .This table is from Stock and Watson (2005b) and the reader should seek in this reference the original source of the data.

#	Mnemonic	Trans	Description
1	A0M052	$\Delta \ln$	Personal income (ar, bil. chain 2000 $)
2	A0M051	$\Delta \ln$	Personal income less transfer payments (ar, bil. chain 2000 $)
3	A0M224	$\Delta \ln$	Real consumption ($A0M224/GMDC$)
4	A0M057	$\Delta \ln$	Manufacturing and trade sales (mil. chain 1996 $)
5	A0M059	$\Delta \ln$	Sales of retail stores (mil. chain 2000 $)
6	IPS10	$\Delta \ln$	Industrial production index - total index
7	IPS11	$\Delta \ln$	Industrial production index - products, total
8	IPS299	$\Delta \ln$	Industrial production index - final products
9	IPS12	$\Delta \ln$	Industrial production index - consumer goods
10	IPS13	$\Delta \ln$	Industrial production index - durable consumer goods
11	IPS18	$\Delta \ln$	Industrial production index - nondurable consumer goods
12	IPS25	$\Delta \ln$	Industrial production index - business equipment
13	IPS32	$\Delta \ln$	Industrial production index - materials
14	IPS34	$\Delta \ln$	Industrial production index - durable goods materials
15	IPS38	$\Delta \ln$	Industrial production index - nondurable goods materials
16	IPS43	$\Delta \ln$	Industrial production index - manufacturing17
17	IPS307	$\Delta \ln$	Industrial production index - residential utilities
18	IPS306	$\Delta \ln$	Industrial production index - fuels
19	PMP	lv	NAPM production index (percent)
20	A0M082	Δlv	Capacity utilization (mfg)
21	LHEL	Δlv	Index of help-wanted advertising in newspapers (1967=100;sa)
22	LHELX	Δlv	Employment: ratio; help-wanted ads/ no. unemployed clf
23	LHEM	Δlv	Civilian labor force: employed, total (thous.)

#	Mnemonic	Trans	Description
24	LHNAG	Δlv	Civilian labor force: employed, nonagricultural industries (thous.)
25	LHUR	Δlv	Unemployment rate: all workers, 16 years & over (%)
26	LHU680	Δlv	Unemployment by duration: average (mean) duration in weeks
27	LHU5	$\Delta \ln$	Unemployment by duration: persons unemployed less than 5 wks (thous.)
28	LHU14	$\Delta \ln$	Unemployment by duration: persons unemployed 5 to 14 wks (thous.)
29	LHU15	$\Delta \ln$	Unemployment by duration: persons unemployed 15 wks + (thous.)
30	LHU26	$\Delta \ln$	Unemployment by duration: persons unemployed 15 to 26 wks (thous.)
31	LHU27	$\Delta \ln$	Unemployment by duration: persons unemployed 27 wks + (thous.)
32	A0M005	$\Delta \ln$	Average weekly initial claims, unemployment insurance (thous.)
33	CES002	$\Delta \ln$	Employees on nonfarm payrolls - total private
34	CES003	$\Delta \ln$	Employees on nonfarm payrolls - goods-producing
35	CES006	$\Delta \ln$	Employees on nonfarm payrolls - mining
36	CES011	$\Delta \ln$	Employees on nonfarm payrolls - construction37
38	CES017	$\Delta \ln$	Employees on nonfarm payrolls - durable goods
39	CES033	$\Delta \ln$	Employees on nonfarm payrolls - nondurable goods
40	CES046	$\Delta \ln$	Employees on nonfarm payrolls - service-providing
41	CES048	$\Delta \ln$	Employees on nonfarm payrolls - trade, transportation, and utilities
42	CES049	$\Delta \ln$	Employees on nonfarm payrolls - wholesale trade
43	CES053	$\Delta \ln$	Employees on nonfarm payrolls - retail trade
44	CES088	$\Delta \ln$	Employees on nonfarm payrolls - financial activities
45	CES140	$\Delta \ln$	Employees on nonfarm payrolls - government
46	A0M048	$\Delta \ln$	Employee hours in nonagricultural establishments (ar, bil. hours)
47	CES151	lv	Average weekly hours of production or nonsupervisory workers on private nonfarm payrolls
48	CES155	Δlv	Average weekly hours of production or nonsupervisory workers on private nonfarm payrolls
49	AOM001	lv	Average weekly hours: manufacturing (hours)
50	PMEMP	lv	NAPM employment index (percent)
51	HSFR	ln	Housing starts: nonfarm (1947-58); total farm
52	HSNE	ln	Housing starts: Northeast (thousands of units)
53	HSMW	ln	Housing starts: Midwest (thousands of units)
54	HSSOU	ln	Housing starts: South (thousands of units)

#	Mnemonic	Trans	Description
55	HSWST	ln	Housing starts: West (thousands of units)
56	HSBR	ln	Housing authorized: total new priv housing units (thousands)
57	HSBNE	ln	Houses authorized by build. permits: Northeast (thousands of units)
58	HSBMW	ln	Houses authorized by build. permits: Midwest (thousands of units)
59	HSBSOU	ln	Houses authorized by build. permits: South (thousands of units)
60	HSBWST	ln	Houses authorized by build. permits: West (thousands of units)
61	PMI	lv	Purchasing managers' index (sa)
62	PMNO	lv	NAPM new orders index (percent)
63	PMDEL	lv	NAPM vendor deliveries index (percent)
64	PMNV	lv	NAPM inventories index (percent)
65	A0M008	$\Delta \ln$	Mfrs' new orders, consumer goods and materials (bil. chain 1982 $)
66	A0M007	$\Delta \ln$	Mfrs' new orders, durable goods industries (bil. chain 2000 $)
67	A0M027	$\Delta \ln$	Mfrs' new orders, nondefense capital goods (mil. chain 1982 $)
68	A1M092	$\Delta \ln$	Mfrs' unfilled orders, durable goods indus. (bil. chain 2000 $)
69	A0M070	$\Delta \ln$	Manufacturing and trade inventories (bil. chain 2000 $)
70	A0M077	Δlv	Ratio, mfg. and trade inventories to sales (based on chain 2000 $)
71	FM1	$\Delta^2 \ln$	Money stock: M1 (bil$,sa)
72	FM2	$\Delta^2 \ln$	Money stock: M2 (bil$,sa)
73	FM3	$\Delta^2 \ln$	Money stock: M3 (bil$,sa)
74	FM2DQ	$\Delta \ln$	Money supply - M2 in 1996 dollars (bci)
75	FMFBA	$\Delta^2 \ln$	Monetary base, adjusted for reserve requirement changes(mil$,sa)
76	FMRRA	$\Delta^2 \ln$	Depository inst. reserves: total, adjusted for reserve req changes (mil$,sa)
77	FMRNBA	$\Delta^2 \ln$	Depository inst. reserves: non-borrowed, adj reserve req changes (mil$,sa)
78	FCLNQ	$\Delta^2 \ln$	Commercial & industrial loans outstanding in 1996 dollars (bci)
79	FCLBMC	lv	Wkly rp lg com'l banks:net change com'l & indus loans (bil$,saar)
80	CCINRV	$\Delta^2 \ln$	Consumer credit outstanding – non-revolving

#	Mnemonic	Trans	Description
81	A0M095	Δlv	Ratio, consumer installment credit to personal income (pct.)
82	FSPCOM	$\Delta \ln$	S&P's common stock price index: composite (1941-43=10)
83	FSPIN	$\Delta \ln$	S&P's common stock price index: industrials (1941-43=10)
84	FSDXP	Δlv	S&P's composite common stock: dividend yield (% per annum)
85	FSPXE	$\Delta \ln$	S&P's composite common stock: price-earnings ratio (%)
86	FYFF	Δlv	Interest rate: Federal funds (effective) (% per annum)87
88	FYGM3	Δlv	Interest rate: u.s. Treasury bills, sec market, 3-mo.(% per annum)
89	FYGM6	Δlv	Interest rate: u.s. Treasury bills, sec market, 6-mo.(% per annum)
90	FYGT1	Δlv	Interest rate: u.s. Treasury const maturities, 1-yr.(% per annum)
91	FYGT5	Δlv	Interest rate: u.s. Treasury const maturities, 5-yr.(% per annum)
92	FYGT10	Δlv	Interest rate: u.s. Treasury const maturities, 10-yr.(% per annum)
93	FYAAAC	Δlv	Bond yield: Moody's AAA corporate (% per annum)
94	FYBAAC	Δlv	Bond yield: Moody's BAA corporate (% per annum)
95	SCP90	lv	CP90 – FYFF (spread)
96	SFYGM3	lv	FYGM3 – FYFF (spread)
97	SFYGM6	lv	FYGM6 – FYFF (spread)
98	SFYGT1	lv	FYGT1 – FYFF (spread)
99	SFYGT5	lv	FYGT5 – FYFF (spread)
100	SFYGT10	lv	FYGT10 – FYFF (spread)
101	SFYAAAC	lv	FYAAAC – FYFF (spread)
102	SFYBAAC	lv	FYBAAC – FYFF (spread)
103	EXRUS	$\Delta \ln$	United States; effective exchange rate (merm) (index no.)
104	EXRSW	$\Delta \ln$	Foreign exchange rate: Switzerland (Swiss franc per U.S.$)
105	EXRJAN	$\Delta \ln$	Foreign exchange rate: Japan (yen per U.S.$)
106	EXRUK	$\Delta \ln$	Foreign exchange rate: United Kingdom (cents per pound)
107	EXRCAN	$\Delta \ln$	Foreign exchange rate: Canada (Canadian$ per U.S.$)

#	Mnemonic	Trans	Description
108	PWFSA	$\Delta^2 \ln$	Producer price index: finished goods (82=100,sa)
109	PWFCSA	$\Delta^2 \ln$	Producer price index: finished consumer goods (82=100,sa)
110	PWIMSA	$\Delta^2 \ln$	Producer price index: intermed mat. supplies & components (82=100,sa)
111	PWCMSA	$\Delta^2 \ln$	Producer price index: crude materials (82=100,sa)
112	PSCCOM	$\Delta^2 \ln$	Spot market price index: bls & crb: all commodities(1967=100)
113	PSM99Q	$\Delta^2 \ln$	Index of sensitive materials prices (1990=100)(bci-99a)
114	PMCP	lv	NAPM commodity prices index (percent)
115	PUNEW	$\Delta^2 \ln$	CPI-u: all items (82-84=100,sa)116
117	PU84	$\Delta^2 \ln$	CPI-u: transportation (82-84=100,sa)
118	PU85	$\Delta^2 \ln$	CPI-u: medical care (82-84=100,sa)
119	PUC	$\Delta^2 \ln$	CPI-u: commodities (82-84=100,sa)
120	PUCD	$\Delta^2 \ln$	CPI-u: durables (82-84=100,sa)
121	PUS	$\Delta^2 \ln$	CPI-u: services (82-84=100,sa)
122	PUXF	$\Delta^2 \ln$	CPI-u: all items less food (82-84=100,sa)
123	PUXHS	$\Delta^2 \ln$	CPI-u: all items less shelter (82-84=100,sa)
124	PUXM	$\Delta^2 \ln$	CPI-u: all items less medical care (82-84=100,sa)
125	GMDC	$\Delta^2 \ln$	PCE, impl price deflator (1987=100)
126	GMDCD	$\Delta^2 \ln$	PCE, impl price deflator: Durables (1987=100)
127	GMDCN	$\Delta^2 \ln$	PCE, impl price deflator: Nondurables (1996=100)
128	GMDCS	$\Delta^2 \ln$	PCE, impl price deflator: Services (1987=100)
129	CES275	$\Delta^2 \ln$	Average hourly earnings of production or nonsupervisory workers on private nonfarm payrolls: goods
130	CES277	$\Delta^2 \ln$	Average hourly earnings of production or nonsupervisory workers on private nonfarm payrolls: construction
131	CES278	$\Delta^2 \ln$	Average hourly earnings of production or nonsupervisory workers on private nonfarm payrolls: manufacturing
132	HHSNTN	Δlv	U. of Michigan index of consumer expectations

D Technical Appendix (Chapter 4)

Posterior inference in the VAR with variable selection

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

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

Algorithm 1

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

1. Sample β from the density

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

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

2. Sample γ_j from the density

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

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

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

The expressions $p\left(y|\theta_j, \gamma_{\setminus -j}, \gamma_j = 1\right)$ and $p\left(y|\theta_j, \gamma_{\setminus -j}, \gamma_j = 0\right)$ are conditional likelihood expressions. Define θ^* to be equal to θ but with

its $j-th$ element $\theta_j = \beta_j$ (i.e. when $\gamma_j = 1$). Similarly, define θ^{**} to be equal to θ but with the $j-th$ element $\theta_j = 0$ (i.e. when $\gamma_j = 0$). Then in the case of the VAR likelihood of model (4.4), we can write l_{0j}, l_{1j} analytically as

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

3. Sample Σ^{-1} from the density

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

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

Algorithm 2

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

Begin from formulation (4.1), and let $y=(y_1',....,y_T')$, $x=(x_1',...,x_T')$ and $\varepsilon=(\varepsilon_1',...,\varepsilon_T')$. A different SUR formulation of the VAR takes the form

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

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

where $Y = vec(y)$ is a $(Tn) \times 1$ column vector, $W = I_m \otimes x$ is a block diagonal matrix of dimensions $(Tn) \times m$ with the matrix x replicated m times on its diagonal, $\theta = \Gamma\beta^*$ is a $m \times 1$ vector, $\beta^* = vec(B')$ and $e = vec(\varepsilon)$ $\sim N(0, \Sigma \otimes I_T)$. To clarify notation, $vec(\circ)$ is the operator that stacks the columns of a matrix and $\otimes$ is the Kronecker product. In this formulation, $W = I_m \otimes x$ is not equal to $z = (z_1', ..., z_T') = \left((I_m \otimes x_1)', ..., (I_m \otimes x_T)'\right)$ which was defined in (4.4). Additionally, note that while β and b are both $n \times 1$ vectors, they are not equal. It holds that $\beta = vec(B)$ and $\beta^* = vec(B')$.

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

1. Sample b from the density

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

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

2. Sample γ_j from the density

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

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

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

3. Sample Σ^{-1} from the density

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

where $\widetilde{\alpha} = T + \alpha$ and $\widetilde{S}^{-1} = \left(S + (Y - x\Theta)'(Y - x\Theta)\right)^{-1}$, where Θ

is the $k \times n$ matrix obtained from the vector $\theta = \Gamma\beta^*$, which has elements $(\Theta_{ij}) = \theta_{(j-1)k+i}$, for $i = 1, ..., k$ and $j = 1, ..., n$.

This sampler has slight modifications compared to the one above because of the different specification of the likelihood function, but the two SUR specifications are equivalent and produce the same results. Posterior inference in the TVP-VAR model is just a simple generalization of the VAR case and it is described in the next section. Unfortunately it is not possible to formulate a TVP-VAR in the form (D.9), in order to take advantage of matrix computations.

Posterior inference in the TVP-VAR with variable selection

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

$$y_t = z_t\theta_t + \varepsilon_t \tag{D.13}$$
$$\beta_t = \beta_{t-1} + \eta_t \tag{D.14}$$

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

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

Estimating these parameters means sampling sequentially from the following conditional densities

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

2. Sample γ_j from the density

$$\gamma_j|\gamma_{\setminus -j}, \beta, \Sigma, y, z \sim Bernoulli\left(1, \widetilde{\pi}_j\right) \tag{D.15}$$

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

$$l_{0j} = p\left(y|\theta_j, \gamma_{\setminus -j}, \gamma_j = 1\right)\pi_{0j} \tag{D.16}$$
$$l_{1j} = p\left(y|\theta_j, \gamma_{\setminus -j}, \gamma_j = 0\right)(1-\pi_{0j}) \tag{D.17}$$

The expressions $p\left(y|\theta_j^{1:T}, \gamma_{\setminus -j}, \gamma_j = 1\right)$ and $p\left(y|\theta_j^{1:T}, \gamma_{\setminus -j}, \gamma_j = 0\right)$ are conditional likelihood expressions, where $\theta_j^{1:T} = [\theta_{1,j}, ..., \theta_{t,j}, ..., \theta_{T,j}]'$. Define θ_t^* to be equal to θ_t but with its $j-th$ element $\theta_{t,j} = \beta_{t,j}$ (i.e. when $\gamma_j = 1$). Similarly, define θ_t^{**} to be equal to θ_t but with the $j-th$ element $\theta_{t,j} = 0$ (i.e. when $\gamma_j = 0$), for all $t = 1, ..., T$. Then in the case of the TVP-VAR likelihood of model (D.13), we can write l_{0j}, l_{1j} analytically as

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

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

$$Q^{-1}|\beta, \gamma, \Sigma, y, z \sim Wishart\left(\widetilde{\xi}, \widetilde{R}^{-1}\right) \tag{D.18}$$

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

4. Sample Σ^{-1} from the density

$$\Sigma^{-1}|\beta, Q, \gamma, y, z \sim Wishart\left(\widetilde{\alpha}, \widetilde{S}^{-1}\right) \tag{D.19}$$

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

Carter and Kohn (1994) algorithm:

Consider a state-space model of the following form

$$y_t = z_t a_t + u_t \tag{D.20a}$$
$$a_t = a_{t-1} + v_t \tag{D.20b}$$
$$u_t \sim N(0, R),\ v_t \sim N(0, W)$$

where (D.20a) is the measurement equation and (D.20b) is the state equation, with observed data y_t and unobserved state a_t. If the errors u_t, v_t are iid and uncorrelated with each other, we can use the Carter and Kohn (1994) algorithm to obtain a draw from the posterior of the unobserved states.

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

$$\begin{aligned} a_{t|t-1} &= a_{t-1|t-1} \\ P_{t|t-1} &= P_{t-1|t-1} + W \\ K_t &= P_{t|t-1}z_t'\left(z_t P_{t|t-1} z_t + R\right)^{-1} \\ a_{t|t} &= a_{t|t-1} + K_t\left(y_t - z_t a_{t|t-1}\right) \\ P_{t|t} &= P_{t|t-1} - K_t z_t P_{t|t-1} \end{aligned} \tag{D.21}$$

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

obtain the smooth estimates $a_{t|t+1}$ and $P_{t|t+1}$ given by the backward recursion:

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

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

Efficient sampling of the variable selection indicators

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

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

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

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

- If $\gamma_j^s = 1$ and $g < \pi_{0j}$ or $\gamma_j^s = 0$ and $g < 1 - \pi_{0j}$, then generate γ_j^{s+1} from the Bernoulli density $\gamma_j|\gamma_{\setminus -j}, b, y, z \sim Bernoulli\left(1, \widetilde{\pi}_j\right)$, where $\widetilde{\pi}_j = \frac{l_{0j}}{l_{0j}+l_{1j}}$ and l_{0j}, l_{1j} are given in equations (D.6)-(D.7) and (D.16)-(D.17), for the VAR and TVP-VAR models respectively.