Essays on Forecasting and Bayesian Model Averaging

STOCKHOLM SCHOOL
OF ECONOMICS
HANDELSHÖGSKOLAN I STOCKHOLM

EFI, The Economic Research Institute

EFI Mission

EFI, the Economic Research Institute at the Stockholm School of Economics, is a scientific institution that works independently of economic, political and sectional interests. It conducts theoretical and empirical research in the management and economic sciences, including selected related disciplines. The Institute encourages and assists in the publication and distribution of its research findings and is also involved in the doctoral education at the Stockholm School of Economics. At EFI, the researchers select their projects based on the need for theoretical or practical development of a research domain, on their methodological interests, and on the generality of a problem.

Research Organization

The research activities at the Institute are organized into 22 Research Centres.
Centre Directors are professors at the Stockholm School of Economics.

EFI Research Centre:	*Centre Director:*
Management and Organisation (A)	Sven-Erik Sjöstrand
Centre for Entrepreneurship and Business Creation (E)	Carin Holmquist
Public Management (F)	Nils Brunsson
Information Management (I)	Mats Lundeberg
Centre for People and Organization (PMO)	Andreas Werr (acting)
Centre for Innovation and Operations Management (T)	Christer Karlsson
Centre for Risk Research (CFR)	Lennart Sjöberg
Economic Psychology (P)	Guje Sevón
Centre for Consumer Marketing (CCM)	Magnus Söderlund
Centre for Information and Communication Research (CIC)	Per Andersson (acting)
Marketing, Distribution and Industrial Dynamics (D)	Björn Axelsson
Centre for Strategy and Competitiveness (CSC)	Örjan Sölvell
Centre for Business and Economic History (BEH)	Håkan Lindgren
Accounting and Managerial Finance (B)	Johnny Lind
Centre for Financial Analysis and Managerial Economics in Accounting (BFAC)	Kenth Skogsvik
Finance (FI)	Clas Bergström
Centre for Health Economics (CHE)	Bengt Jönsson
International Economics and Geography (IEG)	Mats Lundahl
Economics (S)	Paul Segerstrom
Economic Statistics (ES)	Anders Westlund
Law (RV)	Erik Nerep
Centre for Tax Law (SR)	Bertil Wiman

Chair of the Board: Professor Carin Holmquist

Director: Associate Professor Filip Wijkström

Address

EFI, Box 6501, SE-113 83 Stockholm, Sweden • Website: `www.hhs.se/efi`

Telephone: +46(0)8-736 90 00 • Fax: +46(0)8-31 62 70 • E-mail efi@hhs.se

ESSAYS ON FORECASTING AND BAYESIAN MODEL AVERAGING

Jana Eklund

STOCKHOLM SCHOOL
OF ECONOMICS
HANDELSHÖGSKOLAN I STOCKHOLM

EFI, The Economic Research Institute

Dissertation for the Degree of Doctor of Philosophy, Ph.D
Stockholm School of Economics

ISBN 91-7258-710-5

Keywords:
Bayesian model averaging, Forecast combination, GDP forecasts, Inflation forecasts, Forecasting with a large number of predictors, Markov chain Monte Carlo

Printed by:
Elanders Gotab, Stockholm 2006

Distributed by:
EFI, The Economic Research Institute
Stockholm School of Economics
Box 6501, S-113 83 Stockholm, Sweden
`www.hhs.se/efi`

To stubbornness

Contents

Acknowledgments

The journey of writing this thesis has been an adventure, and along the way I met many wonderful persons. To thank all of them would be the right thing to do, but rather overwhelming. I will take the risk of offending those forgotten and mention only a few.

First of all I would like to thank my supervisor, Sune Karlsson. He is an enormous source of knowledge, on virtually any topic. You just have to find the right question to ask. Throughout the years he has been there not only as a mentor, but also as a friend supporting me whenever necessary.

Secondly, my deepest gratitude goes to my husband Bruno. Going through the process of writing and defending a thesis himself a couple of years ago, he is my strongest supporter and most careful critic.

A special place in my heart will always belong to Andrés González and Johan Parmler, the two other musketeers who entered the PhD program at the department at the same time as I did. Nor can I forget Ganesh Munnorcode and Tomoaki Nakatani for their unconditional help in all practical matters surrounding network simulations.

I also wish to thank my Slovak friends, Andrea Klampárová and Martin Stanček, who I met after coming to Sweden, for all those moments that brought a piece of home into the Swedish dark winter evenings.

Special gratitude also goes to Þórarinn Pétursson, Magnús Fjalar Guðmunsson and Ragnhildur Jónsdóttir from Seðlabanki Íslands, the Central Bank of Iceland, for making my last year and a half much more enjoyable and interesting.

Many of my thanks go to my dear volleyball friends at SSIF in Stockholm for politely asking from time to time how the simulations are going and allowing me to take my occasional frustration out over the ball. Also the training with the volleyball teams of ÍS and Þróttur Reykjavík provided an excellent environment for clearing my mind of research and recharging it with energy.

Special thanks are dedicated to Herman van Dijk for his subtle comments that significantly influenced the direction and outcome of this thesis.

Finally, the financial support of Sveriges Riksbank, the Central Bank of Sweden, is gratefully acknowledged.

Stockholm, September 2006
Jana Eklund

Part I

Summary of Thesis

Introduction

The process of systematically searching for models with apparent superior performance is an easy task nowadays for empirical modellers using modern computers and tailor-made software. Statistical modelling in a general-to-specific approach, see for example Hendry (1995), Krolzig and Hendry (2001), Hoover and Perez (2001), or direct model selection methods may be used in this process. However, when the set of possible models increases, these modelling approaches quickly become time consuming. In macroeconomic modelling exercises, the large number of possible explanatory variables available today gives rise to vast sets of potential models. Alternative modelling methods have been developed, specially designed not to reduce the amount of initial data, but to make use of all possible information available.

Methods that exploit large sets of predictors can be divided into two groups. The first group consists of methods based on factor models, in which comovements in the large number of variables are treated as arising from a small number of unobserved factors. Estimates of these common factors, derive via the method of principal components, are used to augment an otherwise standard regression. The second group considers Bayesian model averaging (BMA), where many different models are averaged together to form an aggregate 'model' that takes into account information from each of the single models.

Both of these groups have developed substantially in the past decade. The use of factors to achieve dimensional reduction has been found to be empirically useful in analyzing macroeconomic time series. Studies, for example, by Stock and Watson (2002b), Stock and Watson (2003), Shintani (2003), Artis, Banerjee, and Marcellino (2005) are just a few in the vast literature. BMA has been applied successfully to several statistical model classes. See for example Raftery, Madigan, and Hoeting (1997), Fernández, Ley, and Steel (2001) for applications on linear regression, Madigan and Raftery (1994) for studies on discrete graphical models. Furthermore, Raftery, Madigan, and Volinsky (1995) apply BMA in survival analysis, Koop and Potter (2004) consider

factor-based models in BMA and Jacobson and Karlsson (2004) analyse large macroeconomic panels.

The research in both areas is mainly oriented towards extracting the relevant information from the available data and its use in forecasting. Basing the forecast model on data summaries, in the form of principal components, allows information from all the predictors to enter into the forecasts. BMA, on the other hand, can be viewed as a Bayesian approach to combination forecasting. In forecast combination, the forecast is a weighted average of the individual model forecasts, where the weights typically depend on some measure of historical accuracy of the individual forecasts. In BMA, the posterior probability associated with the model being correct serves as the weight assigned to each model in the forecast combination.

There are two major challenges in implementing BMA in practice: the specification of the prior distributions and the calculation of the posterior distribution. In the case when there are many possible predictors and models available, it is difficult to find well thought out priors for each of the models. Default, uninformative priors are used instead, such as suggested by Fernández, Ley, and Steel (2001). In this way the prior specification problem is reduced to selecting appropriate hyperparameter values. The second major challenge in BMA is how to explore or evaluate the posterior distribution over all possible models. When the number of models is large this is usually achieved using Markov chain Monte Carlo methods, (MCMC). This is a simulation technique that approximates the unknown distribution by drawing random numbers from a closely related distribution, and is designed to improve the approximation as the number of draws increases.

The main theme of this thesis, consisting of four papers, is forecasting using many predictors. The first chapter evaluates the predictive performance of the factor models and the BMA approach to forecast combination, using macroeconomic datasets consisting of up to 161 predictors. The second chapter suggests using as weights in the forecast combination not the usual posterior model probabilities based on marginal likelihood, but weights that are based on a measure of out-of-sample performance - the predictive likelihood. Chapter 3 applies the results obtained in Chapter 2 to forecast output growth in six countries, adding confidence interval to the standard point forecasts. The last chapter touches on the problem of simulating from the posterior models space, by analyzing existing methods and determining through simulations the most accurate and efficient.

Summary and main results

Chapter 1. An embarrassment of riches: Forecasting using large panels[1]

In the past, forecasting models for macroeconomic variables were functions of only a few variables. It is well known that such forecasts tend to be unreliable and unstable, Stock and Watson (2003a). On the other hand, including all predictors, if feasible, is thought to lead to overfitting and poor out-of-sample forecast accuracy. In recent years, researchers have made substantial progress in forecasting using many predictors, where several studies consider datasets consisting of hundreds of variables. Standard methods of comparing all possible combinations of predictors by means of an information criterion function, however, become computationally infeasible when the number of potential predictors is larger then 20 or so. One strategy in this situation is to combine forecasts from many models with alternative subsets of predictors, and sophisticated methods of model averaging weigh individual forecasts by the posterior probabilities of each forecast model. Another strategy is to reduce the dimensionality of the regressor set by extracting its principal components. If the data are generated by an approximate dynamic factor model, then factors estimated by principal components can be used for efficient forecasting under quite general conditions.

This chapter explores the idea of combining forecasts from various indicator models by using Bayesian model averaging and compares the predictive performance of BMA with the predictive performance of factor models. The combination of these two methods, as suggested in Koop and Potter (2004), is also implemented in the comparison, together with a benchmark AR(p) model. In addition, forecasts based on the median model are considered. The median model is defined as the model consisting of variables that have overall posterior inclusion probability of at least 1/2. Barbieri and Berger (2004)

[1] This is a joint work with Sune Karlsson.

show that for selection among linear models the median probability model is often the optimal predictive one. The forecast comparison is conducted in a pseudo out-of-sample framework for three distinct datasets measured at different frequencies. These datasets include a monthly US dataset consisting of 146 predictor variables, a quarterly US dataset containing 161 predictors, and a quarterly Swedish dataset having 77 possible predictors. We compute 4- and 8-step ahead forecasts and the forecasting performance of the various methods described is examined by comparing their simulated out-of-sample RMSFE to that of the benchmark model. The result show that none of the methods is uniformly superior and that no method consistently outperforms or underperforms a simple AR(p) process.

Chapter 2. Forecast combination using predictive measures[2]

In the context of forecasting the idea of model averaging has a long tradition starting with Bates and Granger (1969). Examples of formal evaluations of forecast methods, where forecast combination has performed well, include with a focus on macroeconomic forecasting, Stock and Watson (1999) and the M-competitions Makridakis, Andersen, Carbone, Fildes, Hibon, Lewandowski, Newton, Parzen, and Winkler (1982), Makridakis, Chatfield, Hibon, Lawrence, Mills, Ord, and Simmons (1993) and Makridakis and Hibon (2000). Combining forecasts explicitly accounts for model uncertainty enabling more accurate and realistic inference. While the literature on forecast combination is extensive, see Elliott and Timmermann (2004) for recent theoretical contributions, relatively little attention has been given to the use of predictive measures of fit as the base for forecast combination. Recently, Kapetanios, Labhard, and Price (2006) substitute an out-of-sample measure of fit into standard information criteria when constructing weights for forecast combination in frequentist information theoretic approach.

This chapter proposes the use of the out-of-sample predictive likelihood as the basis for Bayesian model averaging and forecast combination. In addition to the intuitive appeal, the use of the predictive likelihood relaxes the requirement to specify proper priors for the parameters of each model. We show that the forecast weights based on the predictive likelihood have desirable asymptotic properties and that the weights based on the predictive

[2] This is a joint work with Sune Karlsson and has been accepted for publication in the Econometric Reviews, Special Issue on Bayesian Dynamic Econometrics.

likelihood will have better small sample properties than the traditional in-sample marginal likelihood when uninformative priors are used.

In order to calculate the weights for the combined forecast, a number of observations, a hold-out sample, is needed. There is a trade off involved in the size of the hold-out sample. The number of observations available for estimation is reduced, which might have a detrimental effect. On the other hand, as the hold-out sample size increases, the predictive measure becomes more stable and this should improve performance.

A simulation study shows that the use of predictive measures of fit offers greater protection against in-sample overfitting, and improves forecast performance. When there is a true model in the model set, the predictive likelihood will select the true model asymptotically, but the convergence to the true model is slower than for the marginal likelihood. It is this slower convergence, coupled with the protection against overfitting, that is the reason the predictive likelihood performs better when the true model is not in the model set.

Chapter 3. Forecasting GDP with factor models and Bayesian forecast combination

In this chapter the predictive likelihood approach developed in previous chapter is applied to forecasting GDP growth. The analysis is performed on quarterly economic dataset from six countries: Canada, Germany, Great Britain, Italy, Japan and United States. GDP growth is a volatile series and found to be difficult to forecast. Given the ability of predictive likelihood weights to adapt to structural changes, this could improve forecast accuracy. The forecast combination technique based on both in-sample and out-of-sample weights is compared to forecasts based on factor models, an alternative method to forecast with many predictors. The traditional point forecast analysis is extended by considering confidence intervals.

The results indicate that forecast combinations based on the predictive likelihood weights have better forecasting performance compared to the factor models and forecast combinations based on the traditional in-sample weights. In contrast to common findings, the predictive likelihood does improve upon an AR process for longer horizons. The largest improvement over the in-sample weights is for small values of hold-out sample sizes, which provides protection against structural breaks at the end of the sample period.

Chapter 4. Computational efficiency in Bayesian model and variable selection[3]

One of the most significant developments in recent years, along with the general growth of computing power, has been the growth of data. It is now common to search through massive datasets and compute summary statistics from various items that may indicate relationships previously not recognized. The potential benefits of model averaging as a tool for extracting the relevant information from a large set of predictor variables come at the cost of considerable computational complexity. In Bayesian model averaging or model selection we are beginning to encounter model sets of size 10^{30} and larger. It can be quickly determined that a process of evaluating all models would take roughly 32 billion years to complete at a rate of 10^{12} evaluated models per second.

To avoid evaluating all the models, several approaches have been developed to simulate from the posterior distributions. Markov chain Monte Carlo methods can be used to directly draw from the model posterior distributions. It is desirable that the chain moves well through the model space and takes draws from regions with high probabilities. Several computationally efficient sampling schemes, either one at a time or in blocks, have been proposed for speeding up convergence. There is a natural trade-off between local moves, which make use of the current parameter values to propose plausible values for model parameters, and more global transitions, which potentially allow to faster exploration of the distribution of interest, but may be much harder to implement efficiently. Local model moves enable use of fast updating schemes, where it is unnecessary to completely reestimate the new, slightly modified, model to obtain an updated solution.

This chapter investigates the possibilities of increasing computational efficiency by using alternative algorithms to obtain estimates of model parameters as well as keeping track of their numerical accuracy. Also various samplers that explore the model space are presented and compared based on the output of the Markov chain.

[3]This is a joint work with Sune Karlsson.

Bibliography

Artis, M. J., A. Banerjee, and M. Marcellino (2005): "Factor forecasts for the UK," *Journal of Forecasting*, 24(4), 279–298.

Barbieri, M. M., and J. O. Berger (2004): "Optimal Predictive Model Selection," *The Annals of Statistics*, 32(3), 870–897.

Bates, J., and C. Granger (1969): "The Combination of Forecasts," *Operational Research Quarterly*, 20, 451–468.

Elliott, G., and A. Timmermann (2004): "Optimal Forecast Combinations Under General Loss Functions and Forecast Error Distributions," *Journal of Econometrics*, 122(1), 47–79.

Fernández, C., E. Ley, and M. F. Steel (2001): "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100(2), 381–427.

Hendry, D. F. (1995): *Dynamic Econometrics*. Oxford University Press, Oxford.

Hoover, K., and S. J. Perez (2001): "Data Mining Reconsidered: Encompassing and the General-to-specific Approach to Specification Search," *Econometrics Journal*, 2(2), 167–191.

Jacobson, T., and S. Karlsson (2004): "Finding Good Predictors for Inflation: A Bayesian Model Averaging Approach," *Journal of Forecasting*, 23(7), 479–496.

Kapetanios, G., V. Labhard, and S. Price (2006): "Forecasting Using Predictive Likelihood Model Averaging," *Economics Letters*, 91(3), 373 – 379.

Koop, G., and S. Potter (2004): "Forecasting in Dynamic Factor Models Using Bayesian Model Averaging," *Econometrics Journal*, 7(2), 550–565.

Krolzig, H., and D. F. Hendry (2001): "Computer Automation of General-to-Specific Model Selection Procedures," *Journal of Economic Dynamics and Control*, 25(6-7), 831–866.

Madigan, D., and A. E. Raftery (1994): "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89(428), 1535–1546.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler (1982): "The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition," *Journal of Forecasting*, 1(2), 111–153.

Makridakis, S., C. Chatfield, M. Hibon, M. Lawrence, T. Mills, K. Ord, and L. F. Simmons (1993): "The M2-Competition: A Real-Time Judgmentally-Based Forecasting Study," *International Journal of Forecasting*, 9(1), 5–23.

Makridakis, S., and M. Hibon (2000): "The M3-Competition: Results, Conclusions and Implications," *International Journal of Forecasting*, 16(4), 451–476.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997): "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92(437), 179–191.

Raftery, A. E., D. Madigan, and C. Volinsky (1995): "Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance (with Discussion)," in *Bayesian Statistics 5*, ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, pp. 323–349. Oxford University Press, Oxford.

Shintani, M. (2003): "Nonlinear Analysis of Business Cycles Using Diffusion Indexes: Applications to Japan and the U.S," Discussion paper, UCLA Department of Economics, available at http://ideas.repec.org/p/cla/levrem/506439000000000168.html.

Stock, J. H., and M. W. Watson (1999): "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series," in *Cointegration, Causality, and Forecasting A Festschrift in Honour of Clive W.J. Granger*, ed. by R. F. Engle, and H. White, pp. 1–44. Oxford University Press, Oxford.

——— (2002): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics*, 20(2), 147 – 162.

——— (2003a): "Forecasting Output and Inflation: The Role of Asset Prices," *Journal of Economic Literature*, 41(3), 788–829.

——— (2003b): "Has the Business Cycle Changed and Why?," in *NBER Macroeconomics Annual 2002*, ed. by M. Gertler, and K. S. Rogoff. The MIT Press.

Part II

Chapters

Chapter 1

An Embarrassment of Riches: Forecasting Using Large Panels

Acknowledgement: An earlier version of this paper has been presented at the 2003 International Symposium on Forecasting, Merida, and The Time Series Analysis Methods and Application Workshop, Linz, 2003. We wish to thank the participants for their comments.

1.1 Introduction

The number of potential predictors for macroeconomic variables can easily count in the hundreds, e.g. Stock and Watson (2002b) collect some 200 predictor variables for the US economy. Paradoxically, having more information in the form of more predictor variables makes forecasting more difficult. Simply put, the task of formulating a forecasting model and extracting the relevant information from the predictors becomes more complex as the number of possible predictors increases. The increasing availability of data is thus creating new challenges for the forecaster. There are, essentially, two different approaches to address this problem. The first approach builds forecast models based on summaries of the predictor variables, such as principal components, and the second approach is analogous to forecast combination, where the forecasts from a multitude of possible models are averaged.

Each attempts to overcome the shortcomings of the traditional approach of selecting a single forecasting model based on a few predictors. Clearly, using a single model, which by necessity can only incorporate a small subset of the variables, will fail to take account of all information in the data. In addition, by being based on a single model the forecast does not take account of model uncertainty. Basing the forecast model on data summaries in the form of principal components, as in Stock and Watson (2002b), allows information from all the predictors to enter into the forecasts, but not necessarily in an optimal fashion since the summaries of the predictors are created without a reference to the predicted variable. Model averaging, on the other hand, summarizes the different possible *relationships* between the predicted variable and the predictor variables. With appropriately chosen weights, this should lead to more efficient extraction of information. Model averaging also has the advantage of providing robustness against misspecification, and model uncertainty can easily be accounted for, if the model averaging is conducted in a Bayesian setting, i.e. the weights are the posterior probabilities of the models. In addition the models and their averages are more easily interpreted than principal components and inference on the importance of individual predictors is available.

The potential benefits of model averaging as a tool for extracting the relevant information from a large set of predictor variables come at the cost of considerable computational complexity. With 100 predictor variables one obtains more than 10^{30} different models just by considering the different possible combinations of the variables, and it is clearly impossible to include all of them in a model averaging exercise. Recent advances in Bayesian computing, utilized by e.g. Wright(2003a, 2003b), and Jacobson and Karlsson (2004),

provide one way forward by identifying the subset of important models as measured by their posterior probability, i.e. the set of models which would receive a non-negligible weight in the forecast combination.

Koop and Potter (2004) apply Bayesian model averaging (BMA) to dynamic factor models by orthogonalizing the predictors, using a transformation to principal components. Koop and Potter conclude that models containing factors do outperform autoregressive models in forecasting, but only narrowly and at short horizons. Also the gains provided by using BMA over forecasting methods based on a single model are more appreciable relative to the small forecasting gains from factor-based models.

The purpose of this chapter is to evaluate the forecasting performance of the factor model approach of Stock and Watson (2002b), the Bayesian model averaging approach of Jacobson and Karlsson (2004), and the combined approach of Koop and Potter (2004). Any forecast evaluation is dependent on the choice of variable to forecast and the dataset used. To protect against this, we use three different datasets with two different frequencies, and forecast both inflation and GDP.

In all three cases the forecasts are based on a simple linear model,

$$y_{t+h} = \mathbf{x}_t \boldsymbol{\beta}_h + \varepsilon_{t+h}, \tag{1.1}$$

where $\mathbf{x}_t$, in the case of Stock and Watson, consists of the first few principal components, possibly augmented with lags of these and lagged values of y_t. In the Bayesian model averaging approach of Jacobson and Karlsson, $\mathbf{x}_t$ is a subset of the regressor variables, possibly including lags of the predictors and y_t, and the forecasts are obtained by averaging over the forecasts from the different models. In the combined approach of Koop and Potter, $\mathbf{x}_t$ contains a selected subset of orthogonalized regressors. There are two features worth noting about this setup, the forecast model depends on the forecast horizon, h, and the forecasts are static, i.e. there is no need to forecast $\mathbf{x}_t$.

The remainder of the chapter is organized as follows. Section 1.2 presents the forecasting approaches in large panels, Section 1.3 compares the different forecast methods, and Section 1.4 concludes.

1.2 Forecasting methods

1.2.1 Factor models

To forecast the series of interest by extracting the information in many variables while keeping the dimension of the forecasting model small can be achieved by

using factors models. The forecasts are constructed using a two-step procedure. First, the method of principal components is used to estimate factors from the predictors $\mathbf{x}_t$. In the second step the estimated factors are used to forecast the time series y_{t+h}.

In particular, let y_{t+h} be a scalar series that is being forecasted h-periods ahead, and let $\mathbf{x}_t$ be a N-dimensional multiple time series of variables serving as predictors. Now consider the forecasting equation

$$y_{t+h} = \boldsymbol{\beta}(\mathrm{L})\,\mathbf{f}_t + \boldsymbol{\gamma}\mathbf{y}_t + \varepsilon_{t+h}, \qquad t = 1, \ldots, T, \tag{1.2}$$

where $\mathbf{f}_t$ is a vector of q unobservable common factors and $\mathbf{y}_t$ is a smaller set of $p+1$ variables, such as lags of y_t. Furthermore $\boldsymbol{\beta}(\mathrm{L})$ is a vector lag polynomial and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_p)'$. Suppose that the observed series $\mathbf{x}_t$ and y_{t+h} in (1.2) allow for a dynamic factor model with q common dynamic factors $\mathbf{f}_t$

$$x_{it} = \boldsymbol{\lambda}_i(\mathrm{L})\,\mathbf{f}_t + e_{it}, \qquad i = 1, \ldots, N;\ t = 1, \ldots, T, \tag{1.3}$$

where $\boldsymbol{\lambda}_i(\mathrm{L})$ is a lag polynomial vector, and e_{it} is an idiosyncratic disturbance. In addition, it is assumed that

$$\mathrm{E}\left(\varepsilon_{t+h} | \, \mathbf{f}_t, y_t, \mathbf{x}_t, \mathbf{f}_{t-1}, y_{t-1}, \mathbf{x}_{t-1}, \ldots\right) = 0. \tag{1.4}$$

Assume that the lag polynomial vectors are of order s. The dynamic factor model (1.2) – (1.3) can be then restated as

$$y_{t+h} = \mathbf{B}'\mathbf{F}_t + \boldsymbol{\gamma}\mathbf{y}_t + \varepsilon_{t+h}, \tag{1.5}$$

$$\mathbf{x}_t = \boldsymbol{\Lambda}\mathbf{F}_t + \mathbf{e}_t, \tag{1.6}$$

where $\mathbf{B} = (\boldsymbol{\beta}_0', \ldots, \boldsymbol{\beta}_s')', \mathbf{F}_t = \left(\mathbf{f}_t', \mathbf{f}_{t-1}', \ldots, \mathbf{f}_{t-s}'\right)'$ is a $((s+1)\,q \times 1)$ vector, and the i-th row of $\boldsymbol{\Lambda}$ is $(\lambda_{i0}, \ldots, \lambda_{is})$.

The $(s+1)\,q$ factors $\mathbf{F}_t$ in (1.6) are estimated using principal components, denoted by $\widetilde{\mathbf{F}}_t$. In the second step, after regressing y_{t+h} on a constant, $\widetilde{\mathbf{F}}_t$, w possible lags of $\widetilde{\mathbf{F}}_t$, p lags of y_t, the general forecasting function becomes

$$\hat{y}_{T+h|T} = \hat{\alpha}_h + \sum_{j=0}^{w} \hat{\mathbf{B}}_{hj}'\widetilde{\mathbf{F}}_{T-j} + \sum_{j=0}^{p} \hat{\gamma}_{hj} y_{T-j}, \tag{1.7}$$

where $\hat{y}_{T+h|T}$ is the h-step ahead forecast.

Stock and Watson (2002b) use factor models to forecast macroeconomic variables, measuring both real economic activity and prices. The factor model forecast is compared with other forecasting models, such as autoregressive

forecast (AR), vector autoregressive forecast and multivariate leading indicator forecast. Stock and Watson (2002b) consider U.S. monthly series with the total number of possible predictors being 215. They find that for real variables factor models with two factors, or autoregressive factor models with two factors improve the forecasting performance the most. For price indexes the autoregressive factor models forecasts with one factor are preferred. In a recent paper, Boivin and Ng (2005) point out that two researchers can arrive at different forecasts using factor models, because the factors are estimated differently and/or the forecasting equations are specified differently. Boivin and Ng concentrate on the two leading methods in the literature, the dynamic method of Forni, Hallin, Lippi, and Reichlin (2005) and the static method of Stock and Watson (2002a). Boivin and Ng investigate the sensitivity of the estimates of the factors and the forecasts based on factor models to the dynamics of the factors and the specification of the forecasting equation. Their main findings are that unconstrained modelling of the series of interest tends to give more robust forecasts when the data generating process is unknown, and that the methodology of Stock and Watson (2002a) apparently does have these properties.

1.2.2 Bayesian model averaging

Bayesian model averaging can be used to combine forecasts from the set of models that can be constructed using various combinations of the predictors. The averaging over many different competing models incorporates model as well as parameter uncertainty into conclusions about parameters and predictions. Given a set $\mathfrak{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$ of possible models, prior probabilities of the models, $p(\mathcal{M}_i)$, prior distribution of the parameter in each model, $p(\boldsymbol{\theta}_i|\,\mathcal{M}_i)$ and likelihoods, $L(\mathbf{y}|\,\boldsymbol{\theta}_i, \mathcal{M}_i)$ all quantities of interest for model averaging and selection can be obtained by using Bayes rule. The posterior probabilities of the models are given by

$$p(\mathcal{M}_i|\,\mathbf{y}) = \frac{m(\mathbf{y}|\,\mathcal{M}_i)\,p(\mathcal{M}_i)}{\sum_{j=1}^{M} m(\mathbf{y}|\,\mathcal{M}_j)\,p(\mathcal{M}_j)} = \left[\sum_{j=1}^{M} \frac{m(\mathbf{y}|\,\mathcal{M}_j)\,p(\mathcal{M}_j)}{m(\mathbf{y}|\,\mathcal{M}_i)\,p(\mathcal{M}_i)}\right]^{-1}, \quad (1.8)$$

where $m(\mathbf{y}|\,\mathcal{M}_i)$ is the marginal likelihood

$$m(\mathbf{y}|\,\mathcal{M}_i) = \int L(\mathbf{y}|\,\boldsymbol{\theta}_i, \mathcal{M}_i)\,p(\boldsymbol{\theta}_i|\,\mathcal{M}_i)\,\mathrm{d}\boldsymbol{\theta}_i, \quad (1.9)$$

for model $i = 1, \ldots, M$. The posterior distribution of some quantity of interest, ϕ, when taking account of model uncertainty, is

$$p\left(\phi|\,\mathbf{y}\right) = \sum_{j=1}^{M} p\left(\phi|\,\mathbf{y}, \mathcal{M}_j\right) p\left(\mathcal{M}_j|\,\mathbf{y}\right), \tag{1.10}$$

which is an average of the posterior distribution under each of the models, weighted by their posterior model probabilities. Particularly, the minimum mean squared error forecast is given by

$$y_{T+h|T} = \mathrm{E}\left(y_{T+h}|\,\mathbf{y}\right) = \sum_{j=1}^{M} \mathrm{E}\left(y_{T+h}|\,\mathbf{y},\mathcal{M}_j\right) p\left(\mathcal{M}_j|\,\mathbf{y}\right), \tag{1.11}$$

where $\mathrm{E}\left(y_{T+h}|\,\mathbf{y},\mathcal{M}_j\right)$ is the forecast conditional on model $\mathcal{M}_j$.

The parameter prior and the posterior distributions

Consider a linear model with k regressors

$$y_{t+h} = \mathbf{z}_t\boldsymbol{\gamma}_h + \varepsilon_{t+h}, \tag{1.12}$$

where $\boldsymbol{\gamma}_h = (\alpha_h, \boldsymbol{\beta}_h')'$, is a $k+1$ parameter vector and $\mathbf{z}_t = (\boldsymbol{\iota}, \mathbf{x}_t')$ is a vector of explanatory variables.

A challenging task in BMA and model selection is the specification of the prior distribution for the parameters $\boldsymbol{\gamma}$ in the different models. The posterior model probabilities (1.8) depend on the prior for the model parameters. Due to the large number of models it is desirable to use priors in an automated fashion. The priors should be relatively uninformative and also robust in the sense that conclusions are qualitatively insensitive to reasonable changes in the priors. A common choice in BMA for the class of the normal linear model is a g-prior of Zellner (1986) for the regression parameters,

$$p\left(\boldsymbol{\gamma}\,|\sigma^2, \mathcal{M}\right) \sim \mathrm{N}_{k+1}\left(\mathbf{0}, c\sigma^2\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\right) \tag{1.13}$$

that is, the prior mean is set to zero indicating shrinkage of the posterior towards zero and the prior variance is proportional to the data information. Improper priors can be used on the parameters that have identical interpretation across all models. In the case of a linear regression model we can use the usual uninformative prior for the variance,

$$p\left(\sigma^2\right) \propto 1/\sigma^2. \tag{1.14}$$

These priors lead to a proper posterior on the regression parameters, which are t-distributed with T degrees of freedom,

$$p(\boldsymbol{\gamma}|\,\mathbf{y}) \sim \mathrm{t}_{k+1}(\boldsymbol{\gamma}_1, S, \mathbf{M}, T), \tag{1.15}$$

where

$$\boldsymbol{\gamma}_1 = \frac{c}{c+1}\hat{\boldsymbol{\gamma}}, \tag{1.16}$$

is a scaled down version of the least squares estimate, and

$$S = \frac{c}{c+1}(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\gamma}})'(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\gamma}}) + \frac{1}{c+1}\mathbf{y}'\mathbf{y}, \tag{1.17}$$

$$\mathbf{M} = \frac{c+1}{c}\mathbf{Z}'\mathbf{Z}. \tag{1.18}$$

The marginal likelihood is a multivariate t-distribution

$$m(\mathbf{y}|\,\mathcal{M}) \propto (c+1)^{-(k+1)/2} S^{-T/2}. \tag{1.19}$$

The prior (1.13) requires only specification of the hyperparameter c. Our prior is similar to the one advocated by Fernández, Ley, and Steel (2001), the essential difference being that they use an improper prior for both the constant term and the variance. In a rather extensive study Fernández, Ley, and Steel (2001) investigate various choices of c. Their recommendation is to set the hyperparameter to

$$c = \begin{cases} N^2 & \text{if } T \leq N^2, \\ T & \text{if } T > N^2. \end{cases} \tag{1.20}$$

The model prior and the model space

A second challenge arises with the size of the model space. All possible combinations of N potential predictors result in 2^N models. Traversing the complete model space, calculating the posterior probabilities, BMA forecasts and the posterior inclusion probabilities of the variables is thus impractical. A convenient method to identify a set of models with non-negligible posterior model probabilities without examining the full model space, is the reversible jump Markov chain Monte Carlo algorithm, see Green (1995). The details of the algorithm are given as Algorithm 1.1.

This Markov chain converges to the posterior model probabilities under quite general conditions and provides one way of estimating $p(\mathcal{M}|\,\mathbf{y})$. The estimated posterior model probabilities (1.8) are (for obvious reasons) conditional on the set of models visited by the chain. To verify that the Markov chain

Algorithm 1.1 Reversible jump Markov chain Monte Carlo

Suppose that the Markov chain is at model $\mathcal{M}$, having parameters $\boldsymbol{\theta}_{\mathcal{M}}$, where $\boldsymbol{\theta}_{\mathcal{M}}$ has dimension $\dim(\boldsymbol{\theta}_{\mathcal{M}})$.

1. Propose a jump from model $\mathcal{M}$ to a new model $\mathcal{M}'$ with probability $j(\mathcal{M}'|\mathcal{M})$.

2. Generate vector $\mathbf{u}$ (which can have different dimension than $\boldsymbol{\theta}_{\mathcal{M}'}$) from a specified proposal density $q(\mathbf{u}|\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}, \mathcal{M}')$.

3. Set $(\boldsymbol{\theta}_{\mathcal{M}'}, \mathbf{u}') = g_{\mathcal{M},\mathcal{M}'}(\boldsymbol{\theta}_{\mathcal{M}}, \mathbf{u})$, where $g_{\mathcal{M},\mathcal{M}'}$ is a specified invertible function. Hence $\dim(\boldsymbol{\theta}_{\mathcal{M}}) + \dim(\mathbf{u}) = \dim(\boldsymbol{\theta}_{\mathcal{M}'}) + \dim(\mathbf{u}')$. Note that $g_{\mathcal{M},\mathcal{M}'} = g^{-1}_{\mathcal{M}',\mathcal{M}}$.

4. Accept the proposed move with probability

$$\alpha = \min\left\{1, \frac{L(\mathbf{y}|\boldsymbol{\theta}_{\mathcal{M}'}, \mathcal{M}')\, p(\boldsymbol{\theta}_{\mathcal{M}'}|\mathcal{M}')\, p(\mathcal{M}')\, j(\mathcal{M}|\mathcal{M}')}{L(\mathbf{y}|\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M})\, p(\boldsymbol{\theta}_{\mathcal{M}}|\mathcal{M})\, p(\mathcal{M})\, j(\mathcal{M}'|\mathcal{M})} \times \frac{q(\mathbf{u}'|\boldsymbol{\theta}_{\mathcal{M}'}, \mathcal{M}', \mathcal{M})}{q(\mathbf{u}|\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}, \mathcal{M}')} \left|\frac{\partial g_{\mathcal{M},\mathcal{M}'}(\boldsymbol{\theta}_{\mathcal{M}}, \mathbf{u})}{\partial(\boldsymbol{\theta}_{\mathcal{M}}, \mathbf{u})}\right|\right\}. \quad (1.21)$$

5. Set $\mathcal{M} = \mathcal{M}'$ if the move is accepted.

If all parameters of the proposed model are generated directly from a proposal distribution, then $(\boldsymbol{\theta}_{\mathcal{M}'}, \mathbf{u}') = (\mathbf{u}, \boldsymbol{\theta}_{\mathcal{M}})$ with $\dim(\boldsymbol{\theta}_{\mathcal{M}}) = \dim(\mathbf{u}')$ and $\dim(\boldsymbol{\theta}_{\mathcal{M}'}) = \dim(\mathbf{u})$, and the Jacobian is unity. If, in addition, the proposal $q(\mathbf{u}|\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}, \mathcal{M}')$ is the posterior $p(\boldsymbol{\theta}_{\mathcal{M}'}|\mathbf{y}, \mathcal{M}')$ then (1.21) simplifies to

$$\alpha = \min\left\{1, \frac{m(\mathbf{y}|\mathcal{M}')\, p(\mathcal{M}')\, j(\mathcal{M}|\mathcal{M}')}{m(\mathbf{y}|\mathcal{M})\, p(\mathcal{M})\, j(\mathcal{M}'|\mathcal{M})}\right\}. \quad (1.22)$$

This implies that we do not need to perform steps 2 and 3 of the algorithm. Two types of model changing moves are considered:

1. Draw a variable at random and exclude it from the model if it is already in the model otherwise add it. This step is attempted with probability p_A.

2. Swap a randomly selected variable in the model for a randomly selected variable outside the model. This step is attempted with probability $1 - p_A$.

captures most of the total posterior probability mass the method suggested by George and McCulloch (1997) can be used. This method utilizes two separate Markov chains, each starting at a random model. The secondary chain is run for a predetermined number of steps and is then used to provide a capture-recapture type estimate of the total visited probability for the primary chain.

The number of models that enter the model averaging can be further reduced by imposing restriction on the high-dimensional model space. Being uninformative about the model space results in all models having equal probability and then unrealistically large models are included in the average. Instead, a model prior that downweights models containing large number of predictors can be used

$$p\left(\mathcal{M}_i\right) \propto \delta^{k_i}\left(1-\delta\right)^{N-k_i}, \tag{1.23}$$

where k_i is number of predictors in a model $\mathcal{M}_i$. Setting $\delta = 0.5$ is equivalent to a constant model prior

$$p\left(\mathcal{M}_i\right) = p_i = \frac{1}{M}, \quad i = 1, 2, \ldots, M. \tag{1.24}$$

Jacobson and Karlsson (2004) specify forecasting models for Swedish consumer price index inflation using 86 quarterly indicators. Their conclusion is that combining forecasts over all indicator models yields a robust forecast with, in general, smaller root mean squared errors than the included models individually.

1.2.3 Bayesian model averaging with factor models

Koop and Potter (2004) use Bayesian techniques to select factors in dynamic factor models as well as BMA to average forecasts over model specifications. They consider a number of different model priors and evaluate their forecasting performance, and the in-sample and out-of-sample performance of model selection, model averaging and factor models. A common set of variables, the constant term and lags of the dependent variable, are included in each model. These are assigned flat priors and marginalized out in the same way as the constant is treated by Fernández, Ley, and Steel (2001). The basic model (1.12) still applies with suitably transformed dependent and explanatory variables. The priors (1.13) and (1.14) are used for the reduced model. To apply the BMA approach on the factor models, the regressors $\mathbf{z}_t$ are transformed to principal components using the orthogonal transformation $\mathbf{W} = \mathbf{ZE}$, where $\mathbf{E}$ is the matrix of eigenvectors of $\mathbf{Z}'\mathbf{Z}$. The model with orthogonal regressors is then

$$y_{t+h} = \mathbf{w}_t \boldsymbol{\zeta}_h + \varepsilon_{t+h}, \tag{1.25}$$

with $\boldsymbol{\zeta}_h = \mathbf{E}^{-1}\boldsymbol{\gamma}_h$. The prior for the regression coefficients becomes

$$p\left(\boldsymbol{\zeta}\left|\sigma^2, \mathcal{M}\right.\right) \sim \mathrm{N}_{k+1}\left(\mathbf{0}, c\sigma^2\left(\mathbf{E}'\mathbf{Z}'\mathbf{Z}\mathbf{E}\right)^{-1}\right) \tag{1.26}$$

yielding the posterior

$$p\left(\boldsymbol{\zeta}\left|\mathbf{y}\right.\right) \sim \mathrm{t}_{k+1}\left(\boldsymbol{\zeta}_1, S, \mathbf{M}, T\right), \tag{1.27}$$

with $\mathbf{Z}$ and $\boldsymbol{\gamma}$ replaced accordingly by $\mathbf{W}$ and $\boldsymbol{\zeta}$, respectively in the equation (1.16).

Koop and Potter (2004) focus on forecasting the growth rates of US GDP and inflation using a set of 162 predictors. They conclude that BMA forecasts improve on an AR(2) benchmark forecasts at short, but not at longer horizons and only by a small margin. These findings are attributed to the presence of structural instability and the fact that lags of dependent variable seem to contain most of the information relevant for forecasting. Koop and Potter investigate also the forecasting performance of several model priors. They found that priors, which focus on principal components explaining 99.9% of the variance of the predictors, provide the best results, and that the non-informative prior (1.24) performs very poorly.

1.2.4 Median probability model

In addition we take the opportunity to apply a method recently proposed by Barbieri and Berger (2004). This method does not directly deal with the situation of having many predictors, but it is of some interest and is easily implemented since the posterior model probabilities are available from the BMA exercise.

Barbieri and Berger show that for selection among linear models the optimal predictive model is often the median probability model, which is defined as the model consisting of variables that have overall posterior inclusion probability of at least 1/2. The posterior inclusion probability for a variable i is given by

$$p\left(x_i\left|\mathbf{y}\right.\right) = \sum_{j=1}^{M} \mathit{1}\left(x_i \in \mathcal{M}_j\right) p\left(\mathcal{M}_j\left|\mathbf{y}\right.\right), \tag{1.28}$$

where $\mathit{1}\left(x_i \in \mathcal{M}_j\right)$ equals one if x_i is included in model $\mathcal{M}_j$ and zero otherwise. It is possible that no variable has a posterior inclusion probability exceeding 1/2. The median probability model is, however, assured to exist in two important cases, one is the problem of variable selection, when any variable

can be included or excluded from the model, and the other case is when the models under consideration follow a graphical model structure, for example a sequence of nested models. Barbieri and Berger show that the median probability model will frequently coincide with the highest posterior probability model. One obvious situation is when there is a model with posterior probability higher than 1/2. Other situations include the problem of variable selection under an orthogonal design matrix, certain prior structures and known variance σ^2. In Barbieri and Berger's (2004) experience the median probability model outperforms the maximum probability model in terms of predictive performance. They suggest that the median probability model should routinely be determined and reported as a complement to the maximum probability model.

1.3 Forecast comparison

We explore the performance of the methods mentioned in the previous section on three different datasets, and compare their performance through the root mean square forecast error (RMSFE). The first dataset is the U.S. monthly balanced dataset of Stock and Watson (2002b) consisting of 146 series from 1960:01 to 1998:12. The second dataset consists of 161 quarterly U.S. time series from 1959Q1 until 2000Q1 and was used in Stock and Watson (2003) and in Koop and Potter (2004). The last dataset is a Swedish dataset comprising of 77 variables including a wide range of indicators of real and monetary aspects of the Swedish economy ranging from 1983Q1 to 2003Q4.[1] We forecast the CPI or the inflation rate for all three datasets, and for the U.S. quarterly data we forecast the GDP growth rate as well. The forecasting model is

$$y_{t+h} = \alpha_h + \mathbf{x}_t \boldsymbol{\beta}_h + \varepsilon_{t+h}, \tag{1.29}$$

which generalizes (1.7), (1.12) and (1.25) to arbitrary forecasting horizons. The choice of dependent variable as y_{t+h}, instead of y_t, has the great advantage that it does away with the need of forecasting the predictors in $\mathbf{x}_t$ when forecasting y_{t+h}. The obvious disadvantage of this choice of dependent variable is that it leads to a different model for each forecast horizon.

Following Stock and Watson the dependent variable and the predictors in the U.S. datasets are transformed into stationary series. In particular, GDP is modelled as being $I(1)$ in logarithms and CPI as $I(2)$ in logarithms. This

[1] See the cited works for the list of the predictors for the US data and Appendix A for the list of the Swedish series and their transformation.

implies that y_{t+h} for GDP and CPI are transformed as

$$y_{t+h} = a/h \cdot \ln\left(GDP_{t+h}/GDP_t\right), \tag{1.30}$$
$$y_{t+h} = a/h \cdot \ln\left(CPI_{t+h}/CPI_t\right) - a\ln\left(CPI_t/CPI_{t-1}\right), \tag{1.31}$$

where $a = 400$ for quarterly data and $a = 1200$ for monthly data.

The Swedish inflation rate is measured as the four-quarter percentage change in the consumer price index and the remaining variables in the dataset are with few exceptions 4 quarter growth rates or 4 quarter log differences. The current level of inflation is included in the set of predictor variables for inflation h-periods ahead. A dummy variable d_t, for the Swedish low inflation regime dated to start in 1992Q1, is always included in the model (1.29).

The forecasts are constructed for horizons $h = 6$ and $h = 12$ for the monthly data, and for horizons $h = 4$ and $h = 8$ for the quarterly data, respectively. For the U.S. monthly data we evaluate the forecast performance for the period 1989:01 until 1998:12. To investigate the possibility that some forecast work well in some periods but poor during others, we also calculate the forecast for four 30 months sub-periods.

For the factor model forecasts we consider the following variants proposed by Stock and Watson: the forecasts denoted by FM-AR,Lag, based on equation (1.7), include v estimated factors, w lags of factors and p lags of y_t, where all the lag lengths are determined by the Bayesian information criterion (BIC). The FM-AR forecasts contain no lags of $\hat{\mathbf{F}}_t$, with v and p determined by BIC. Finally, FM forecasts contain only contemporaneous $\hat{\mathbf{F}}_t$, with v selected by BIC. Further, forecasts based on the estimated factors holding the number of factors v fixed are also considered, first determining the lag length of the dependent variable, p, by BIC, and then setting $p = 0$. These are denoted as FM-AR,v, and FM,v, respectively.

Our implementation of BMA in dynamic factor models differs slightly from the implementation in Koop and Potter (2004). We use only contemporary values when forming the principal components and use the components corresponding to 20 largest eigenvalues. This roughly corresponds to the 99.9% prior that Koop and Potter find to work well. In addition to the principal components we allow for p lags of the dependent variable in the set of potential predictors rather than forcing these into all forecasting models. We consider forecasts based on two different sets of predictors. The BMA-FM forecasts use only the 20 first principal components and the BMA-FM-AR augments the potential set of predictions with p lags of y_t.

The forecasts calculated using the Bayesian approach include forecasts based on the forecast combination from all visited models, forecasts from 3

Table 1.1 Summary of datasets used and settings for forecasting

	U.S. monthly dataset	U.S. quarterly dataset	Swedish quarterly dataset
Total no. of variables, N	146	161	77
Forecasted variable, y_t	CPI	CPI, GDP	Inflation
Data span	1960:01-1998:12	1961Q1-2000Q4	1983Q1-2003Q4
Forecasted period	1989:01-1998:12	1981Q1-2000Q4	1999Q1-2003Q4
Forecasting horizons, h	6,12	4,8	4,8
No. of sub-periods	4	4	2
No. of lags of y_t, p	$[0, 5]$	$[0, 3]$	$[0, 3]$
No. of factors, FM, v	$[1, 4]$	$[1, 4]$	$[1, 4]$
No. of factors, FM-AR, v	$[1, 12]$	$[1, 4]$	$[1, 4]$
No. of factor lags, w	$[0, 2]$	$[0, 2]$	$[0, 5]$
δ in prior (1.23) in BMA	0.075	0.05	0.1
No. of factors, BMA-FM	20	20	20
δ in prior (1.23) in BMA-FM	0.5	0.5	0.5
MCMC replicates	5 000 000	5 000 000	5 000 000
Burn-in	50 000	50 000	50 000

models with the highest posterior probabilities, denoted Top1 to Top3, and forecasts based on the median model. For the U.S. monthly data we set the model hyperparameter δ to 0.075 in the usual BMA approach. This corresponds to an expected model size of 11 variables. For the BMA-FM approach, $\delta = 0.5$, giving expected size of 10 variables. The value of c is chosen as in (1.20), i.e. $c = N^2$. The Markov chain is run for 5 000 000 steps with 50 000 steps as burn-in. The parameters defining the forecast experiments are summarized in Table 1.1.

The results from the different approaches are compared to a benchmark, an AR process with the lag length determined by BIC. In addition we calculate forecasts based on the random walk, i.e. when the forecast of y_{t+h} equals the current value of the dependent variable.

1.3.1 Results

The results for the transformed U.S. CPI series are reported in Tables 1.2 and 1.3, Tables 1.4 - 1.7 show the results for the U.S. quarterly data, and the results for the Swedish inflation rate can be found in Tables 1.8 - 1.9. The first data

column in the tables represents results based on the whole forecasting period and the remaining columns contain results for the sub-periods. The values reported in the tables are the relative RMSFEs

$$\frac{\text{RMSFE}\left(\hat{y}_{T+h|T}\right)}{\text{RMSFE}\left(\hat{y}^{AR}_{T+h|T}\right)}. \tag{1.32}$$

In general, there is no method that consistently outperforms other methods across all periods, datasets or forecasting horizons. There are, however, some patterns in the results that merit further investigation.

It is important to include lags of the dependent variable when forecasting inflation. Only then is it possible to outperform an AR(p) process. This conclusion is supported in as much as 2/3 of the different periods. Koop and Potter (2004) find that an AR(2) process outperforms factor-based models for longer forecasting horizons. The better predictive performance of an AR(2) can, according to Koop and Potter, likely be explained by the fact that the relevant predictive information is included in the lags of the dependent variable.

One possible exception to this is the US GDP forecasts, where the predictive performance of the FM forecasts is unaffected by allowing for lags of the dependent variable. The GDP forecasts from the sub-periods and the whole period also indicate that it is not clear what the number of included factors should be. The best performing model contains different number of factors across the subperiods. Also, controlling for the number of factors included in a model often performs better than using BIC for their determination.

Forecast combination using BMA regularly outperforms the forecasts from models selected by the posterior model probabilities.

The results differ substantially between the Swedish dataset, where the BMA-based methods perform poorly, and the two US datasets. Overall BMA-FM does better than the FM forecasts for the US datasets. Allowing for lags of the dependent variable in BMA-FM-AR improves the forecasts somewhat, but the FM-AR forecasts show a larger improvement.

On the issue of selecting predictors from the original variables, or data summaries such as principal components, the evidence is mixed. The BMA forecasts do better for the monthly data and the BMA-FM forecasts better for the quarterly data.

Comparing the median model with the highest posterior probability model fails to prove its superiority for forecasting. The median model produces smaller RMSFE than Top1 model only in 47% of all cases.

Table 1.2 RMSFE relative to an AR(p) for monthly U.S. CPI, 6 months ahead forecast.

	89:1-98:12	89:1-91:6	91:7-93:12	94:1-96:6	96:7-98:12
BMA	1.0911	0.8151	0.8986	1.9560	1.8041
Top 1	1.1247	0.7735	0.8989	2.1858	1.9074
Top 2	1.1300	0.7978	0.9103	2.1334	1.9181
Top 3	1.1148	0.8174	0.8443	2.1195	1.8477
Median model	1.1469	0.7967	0.8989	2.2709	1.8950
BMA-FM-AR	1.2189	1.0005	1.3682	1.7506	1.5350
Top 1	1.2435	1.0155	1.3923	1.8061	1.5742
Top 2	1.2418	1.0065	1.3722	1.8172	1.6308
Top 3	1.2313	0.9941	1.4138	1.7883	1.5304
Median model	1.2348	1.0186	1.3715	1.7645	1.5742
BMA-FM	1.4915	1.3485	1.4502	2.1612	1.7257
Top 1	1.5288	1.3430	1.5996	2.2069	1.7404
Top 2	1.5163	1.3575	1.5199	2.2273	1.6904
Top 3	1.5021	1.3344	1.5422	2.1313	1.7432
Median model	1.5283	1.3430	1.5990	2.2044	1.7404
FM-AR, Lag	0.8859	0.7765	0.8140	1.1341	1.3683
FM-AR	0.8998	0.7764	0.9312	1.1005	1.3103
FM	1.4747	1.3808	1.3082	2.0714	1.7743
FM-AR, $v = 1$	0.9068	0.8074	0.8177	1.1450	1.3864
FM-AR, $v = 2$	0.8998	0.7764	0.9312	1.1005	1.3103
FM-AR, $v = 3$	0.8990	0.7756	0.9278	1.0950	1.3174
FM-AR, $v = 4$	0.9020	0.7766	0.9472	1.0902	1.3065
FM, $v = 1$	1.7330	1.7039	1.4604	2.2050	2.0197
FM, $v = 2$	1.7249	1.6917	1.4193	2.2327	2.0553
FM, $v = 3$	1.6954	1.6390	1.4651	2.2308	1.9917
FM, $v = 4$	1.5412	1.4061	1.2932	2.2986	2.0215
Random walk	2.6683	2.4604	2.8274	3.1155	3.1363
AR, RMSFE	0.0053	0.0083	0.0050	0.0031	0.0031

Table 1.3 RMSFE relative to an AR(p) for monthly U.S. CPI, 12 months ahead forecast.

	89:1-98:12	89:1-91:6	91:7-93:12	94:1-96:6	96:7-98:12
BMA	1.2098	0.9203	0.9186	1.9561	2.0671
Top 1	1.2541	0.9288	1.0022	1.7705	2.1715
Top 2	1.2765	0.9460	1.0617	2.2253	1.9353
Top 3	1.2147	1.0089	0.8740	1.9053	2.1088
Median model	1.2300	0.9074	0.9534	1.8035	2.1640
BMA-FM-AR	1.2489	1.4686	0.8855	2.2020	1.4674
Top 1	1.2830	1.4890	0.8585	2.4392	1.5491
Top 2	1.2734	1.4655	0.8890	2.2842	1.5670
Top 3	1.3079	1.4671	0.9498	2.2635	1.6436
Median model	1.2769	1.4890	0.8917	2.2423	1.5654
BMA-FM	1.4063	1.6903	1.0391	2.4939	1.4448
Top 1	1.3969	1.6589	1.0378	2.5728	1.3888
Top 2	1.4601	1.6750	1.1153	2.7405	1.4501
Top 3	1.3887	1.7042	0.9846	2.4904	1.4453
Median model	1.3940	1.6551	1.0339	2.5703	1.3893
FM-AR, Lag	0.9472	1.1032	0.5940	1.3414	1.4666
FM-AR	0.9500	1.0960	0.6902	1.3328	1.3079
FM	1.4225	1.7191	1.0590	2.4182	1.4948
FM-AR, $v = 1$	0.9383	1.0845	0.6327	1.3495	1.3760
FM-AR, $v = 2$	0.9230	1.0558	0.6621	1.2532	1.3227
FM-AR, $v = 3$	0.9250	1.0439	0.6244	1.3119	1.4021
FM-AR, $v = 4$	0.9229	1.0570	0.6390	1.2492	1.3690
FM, $v = 1$	1.6768	2.0189	1.3410	2.5525	1.6960
FM, $v = 2$	1.6748	1.9993	1.3237	2.6328	1.7292
FM, $v = 3$	1.6424	1.9722	1.2235	2.8380	1.7192
FM, $v = 4$	1.4618	1.7028	1.0418	2.6525	1.6846
Random walk	2.6668	2.5597	2.5628	4.1441	2.4498
AR, RMSFE	0.0106	0.0108	0.0157	0.0051	0.0076

Table 1.4 RMSFE relative to an AR(p) for quarterly U.S. CPI, 4 quarters ahead forecast.

	81:1-00:4	81:1-85:4	86:1-90:4	91:1-95:4	96:1-00:4
BMA	0.8118	0.7582	0.8570	0.7386	1.2992
Top 1	0.8814	0.8313	0.9189	0.8344	1.3409
Top 2	0.9429	0.9030	0.9680	0.8794	1.3878
Top 3	0.8782	0.8369	0.9565	0.6481	1.3371
Median model	0.9376	0.9326	0.8691	0.8896	1.3728
BMA-FM-AR	0.8113	0.7122	0.9516	0.8462	1.1386
Top 1	0.8610	0.7881	0.9754	0.8619	1.1308
Top 2	0.8635	0.7488	1.0365	0.8986	1.1863
Top 3	0.8256	0.7370	0.9370	0.8918	1.1373
Median model	0.8899	0.8135	1.0179	0.8619	1.1833
BMA-FM	0.8037	0.7012	0.9480	0.8463	1.1278
Top 1	0.8301	0.7344	0.9708	0.8593	1.1377
Top 2	0.8420	0.7041	1.0480	0.9001	1.1603
Top 3	0.8195	0.7235	0.9346	0.8819	1.1859
Median model	0.8088	0.6732	1.0084	0.8614	1.1378
FM-AR, Lag	0.7029	0.5199	0.9625	0.7350	1.0822
FM-AR	0.7605	0.6579	0.9277	0.7350	1.0822
FM	0.8852	0.8395	1.0013	0.7454	1.1028
FM-AR, $v = 1$	0.8039	0.7461	0.9546	0.5489	1.1170
FM-AR, $v = 2$	0.8116	0.7544	0.9729	0.5163	1.1152
FM-AR, $v = 3$	0.8080	0.7445	0.9843	0.5076	1.1033
FM-AR, $v = 4$	0.7416	0.6219	0.9277	0.7350	1.0822
FM, $v = 1$	0.9676	0.9040	1.0914	0.9737	1.0905
FM, $v = 2$	0.9042	0.8457	1.0223	0.8615	1.0878
FM, $v = 3$	0.8552	0.8290	0.9620	0.5962	1.0741
FM, $v = 4$	0.8383	0.7562	1.0013	0.7454	1.1136
Random walk	1.4910	1.3339	1.6208	2.0157	1.4262
AR, RMSFE	0.0184	0.0289	0.0180	0.0117	0.0077

Table 1.5 RMSFE relative to an AR(p) for quarterly U.S. CPI, 8 quarters ahead forecast.

	81:1-00:4	81:1-85:4	86:1-90:4	91:1-95:4	96:1-00:4
BMA	0.8469	0.8139	1.0077	0.8152	0.8860
Top 1	1.0033	0.9257	1.2254	1.1174	1.1244
Top 2	0.9542	0.8172	1.3516	1.0963	1.0937
Top 3	0.9726	0.9515	1.1286	0.7770	1.2175
Median model	1.0934	1.0175	1.4188	0.9938	1.3249
BMA-FM-AR	0.7840	0.7343	0.9761	0.7497	0.9600
Top 1	0.8633	0.8264	1.0579	0.7300	1.0673
Top 2	0.8878	0.8926	0.9667	0.6914	0.9918
Top 3	0.8273	0.8023	0.9065	0.8022	1.0280
Median model	0.8194	0.7624	1.0602	0.7175	1.0634
BMA-FM	0.7775	0.7265	0.9689	0.7520	0.9554
Top 1	0.8379	0.7911	1.0564	0.7168	1.0671
Top 2	0.8886	0.8943	0.9611	0.7066	0.9716
Top 3	0.8647	0.8529	0.9557	0.7542	1.0153
Median model	0.8279	0.7899	1.0102	0.7059	1.0671
FM-AR, Lag	0.7368	0.6286	1.1048	0.6695	1.0312
FM-AR	0.7821	0.7276	1.0350	0.6310	1.0310
FM	0.8287	0.7746	1.0765	0.6968	1.0676
FM-AR, $v = 1$	0.8096	0.7539	1.1073	0.5372	1.1077
FM-AR, $v = 2$	0.7942	0.7550	1.0264	0.5092	1.1300
FM-AR, $v = 3$	0.8117	0.7904	1.0201	0.5209	1.0014
FM-AR, $v = 4$	0.7448	0.6703	1.0350	0.6310	1.0310
FM, $v = 1$	0.9387	0.8777	1.1680	0.9187	1.1259
FM, $v = 2$	0.8546	0.8101	1.0164	0.7928	1.1372
FM, $v = 3$	0.7982	0.7801	0.9787	0.5331	1.0014
FM, $v = 4$	0.7814	0.7021	1.0765	0.6968	1.0676
Random walk	1.5620	1.5332	1.6817	1.5093	1.7726
AR, RMSFE	0.0449	0.0757	0.0338	0.0298	0.0173

Table 1.6 RMSFE relative to an AR(p) for quarterly U.S. GDP, 4 quarters ahead forecast.

	81:1-00:4	81:1-85:4	86:1-90:4	91:1-95:4	96:1-00:4
BMA	1.2063	1.1241	2.0225	1.1881	0.9988
Top 1	1.1099	1.0836	1.7165	0.9905	0.9725
Top 2	1.3137	1.2435	2.2510	1.2081	1.0758
Top 3	1.2749	1.1216	2.4049	1.2816	1.1548
Median model	1.3179	1.1917	2.2045	1.4320	0.9725
BMA-FM-AR	1.0100	0.9199	1.6678	1.0699	0.8370
Top 1	1.0570	0.9495	1.8045	1.1128	0.9399
Top 2	1.0084	0.9092	1.6169	1.0855	0.9616
Top 3	1.0483	0.9517	1.7625	1.0890	0.9312
Median model	1.0215	0.9207	1.7204	1.0653	0.9581
BMA-FM	1.0111	0.9480	1.5730	1.0406	0.8266
Top 1	1.0107	0.9573	1.4718	1.0436	0.8778
Top 2	1.0338	0.9575	1.6618	1.0703	0.8599
Top 3	1.0771	0.9822	1.7482	1.1734	0.7916
Median model	1.0098	0.9572	1.4806	1.0386	0.8662
FM-AR, Lag	1.0031	0.9422	1.6103	1.0319	0.6892
FM-AR	0.9486	0.8780	1.4170	1.0319	0.7901
FM	0.9486	0.8780	1.4170	1.0319	0.7901
FM-AR, $v = 1$	0.9828	0.9901	0.9047	0.9879	0.9620
FM-AR, $v = 2$	0.9323	0.8571	1.3808	1.0319	0.7901
FM-AR, $v = 3$	0.9485	0.8473	1.6461	0.9949	0.8564
FM-AR, $v = 4$	0.9527	0.8575	1.6932	0.9422	0.9254
FM, $v = 1$	0.9828	0.9901	0.9047	0.9879	0.9620
FM, $v = 2$	0.9323	0.8571	1.3808	1.0319	0.7901
FM, $v = 3$	0.9485	0.8473	1.6461	0.9949	0.8564
FM, $v = 4$	0.9527	0.8575	1.6932	0.9422	0.9254
Random walk	1.3654	1.4445	1.1553	1.2466	0.9992
AR, RMSFE	0.0198	0.0325	0.0099	0.0179	0.0097

Table 1.7 RMSFE relative to an AR(p) for quarterly U.S. GDP, 8 quarters ahead forecast.

	81:1-00:4	81:1-85:4	86:1-90:4	91:1-95:4	96:1-00:4
BMA	1.0405	0.8242	2.4397	1.1525	1.5025
Top 1	1.0523	0.7986	2.3529	1.2145	1.6886
Top 2	1.1186	1.0393	2.5036	1.0022	1.2828
Top 3	1.2311	0.9005	3.0433	1.5464	1.3403
Median model	1.0401	0.8134	2.4967	1.2126	1.2583
BMA-FM-AR	0.9615	0.7761	2.0866	1.1358	1.1387
Top 1	1.0726	0.8735	2.4169	1.2232	1.3094
Top 2	1.0054	0.8168	2.1116	1.1595	1.3516
Top 3	1.0520	0.8435	2.4308	1.2215	1.2274
Median model	1.0428	0.8499	2.2900	1.2102	1.2358
BMA-FM	0.9603	0.7701	2.0662	1.1547	1.1015
Top 1	1.0562	0.8820	2.2669	1.2049	1.2131
Top 2	0.9937	0.7788	2.3552	1.1686	1.1752
Top 3	1.0467	0.8391	2.2653	1.2503	1.2316
Median model	1.0399	0.8352	2.2969	1.2299	1.2157
FM-AR, Lag	1.0480	0.9597	2.2734	1.0671	0.8531
FM-AR	0.9986	0.9264	1.8114	1.0671	0.8531
FM	0.9986	0.9264	1.8114	1.0671	0.8531
FM-AR, $v = 1$	1.0218	1.0313	0.9943	1.0025	1.0042
FM-AR, $v = 2$	0.9608	0.8950	1.4641	1.0671	0.8531
FM-AR, $v = 3$	0.9704	0.8799	1.8158	1.0645	0.8734
FM-AR, $v = 4$	0.9909	0.9141	1.8340	1.0593	0.8706
FM, $v = 1$	1.0218	1.0313	0.9943	1.0025	1.0042
FM, $v = 2$	0.9608	0.8950	1.4641	1.0671	0.8531
FM, $v = 3$	0.9704	0.8799	1.8158	1.0645	0.8734
FM, $v = 4$	0.9909	0.9141	1.8340	1.0593	0.8706
Random walk	1.3643	1.3344	2.6074	1.2800	0.9730
AR, RMSFE	0.0340	0.0557	0.0120	0.0339	0.0151

Table 1.8 RMSFE relative to an AR(p) for quarterly Swedish inflation rate, 4 quarters ahead forecast.

	99:1-03:4	99:1-01:2	01:3-03:4
BMA	1.8653	1.8252	1.9077
Top 1	2.0525	1.8655	2.2372
Top 2	1.7154	1.5059	1.9164
Top 3	2.2446	2.3128	2.1684
Median model	2.1348	1.9876	2.2833
BMA-FM-AR	3.7170	4.8758	1.7449
Top 1	3.7621	4.9464	1.7314
Top 2	3.9111	5.2129	1.5651
Top 3	3.8674	5.1130	1.6903
Median model	3.8101	5.0274	1.6970
BMA-FM	3.7880	5.0065	1.6604
Top 1	3.9833	5.2662	1.7414
Top 2	4.0304	5.2869	1.8921
Top 3	4.2011	5.5781	1.7559
Median model	4.0192	5.3173	1.7449
FM-AR, Lag	1.2061	1.3896	0.9692
FM-AR	1.2678	1.5859	0.7909
FM	0.7346	0.6250	0.8371
FM-AR, $v = 1$	0.6796	0.5573	0.7909
FM-AR, $v = 2$	0.6678	0.5206	0.7969
FM-AR, $v = 3$	0.6864	0.5148	0.8331
FM-AR, $v = 4$	1.4255	1.7763	0.9038
FM, $v = 1$	0.7346	0.6250	0.8371
FM, $v = 2$	0.6127	0.4593	0.7439
FM, $v = 3$	0.6144	0.4630	0.7442
FM, $v = 4$	0.9607	1.0707	0.8254
Random walk	1.1763	1.2390	1.1046
AR, RMSFE	0.8714	0.8883	0.8541

Table 1.9 RMSFE relative to an AR(p) for quarterly Swedish inflation rate, 8 quarters ahead forecast.

	99:1-03:4	99:1-01:2	01:3-03:4
BMA	2.4462	1.0398	4.6134
Top 1	2.9144	1.4916	5.2969
Top 2	2.4841	1.3079	4.4822
Top 3	2.6555	1.4714	4.7224
Median model	2.5931	0.9828	4.9694
BMA-FM-AR	2.3955	2.0820	3.1810
Top 1	2.9568	2.2670	4.4709
Top 2	3.1632	2.1864	5.1297
Top 3	3.2356	2.3933	5.0265
Median model	2.5212	2.3200	3.0653
BMA-FM	2.8479	1.8076	4.8192
Top 1	3.1357	1.9874	5.3096
Top 2	2.9875	1.7706	5.1955
Top 3	3.1580	1.9899	5.3608
Median model	2.9583	1.9486	4.9207
FM-AR, Lag	1.0504	0.5611	1.8879
FM-AR	1.0504	0.5611	1.8879
FM	1.1689	0.4917	2.2081
FM-AR, $v = 1$	0.5919	0.4178	0.9481
FM-AR, $v = 2$	0.8329	0.6488	1.2429
FM-AR, $v = 3$	0.7756	0.6011	1.1624
FM-AR, $v = 4$	1.2303	0.6737	2.1957
FM, $v = 1$	0.7033	0.4917	1.1330
FM, $v = 2$	0.6381	0.4496	1.0233
FM, $v = 3$	0.6209	0.4560	0.9694
FM, $v = 4$	1.3026	0.7200	2.3181
Random walk	1.0256	0.8284	1.4808
AR, RMSFE	1.3238	1.6292	0.9223

1.4 Conclusions

This chapter compares methods for extracting information relevant for forecasting from a large number of predictors. The factor models, the Bayesian model averaging approach and the combination of the two are evaluated on US and Swedish data at both monthly and quarterly frequencies. We find that none of the methods is uniformly superior and that no method performs better than, or is outperformed by, a simple AR(p) process.

A possible disadvantage of all the methods considered here is that they are based on linear models that forecast h-steps ahead directly. It is quite possible that these simple models fail to capture all the information contained in the data. In future research, more complicated, and thus more realistic functions, will be considered. This could improve forecast accuracy, but comes at the cost of not having closed form expressions for the marginal likelihood and posterior distributions with increased computational complexity as the result.

Bibliography

Barbieri, M. M., and J. O. Berger (2004): "Optimal Predictive Model Selection," *The Annals of Statistics*, 32(3), 870–897.

Boivin, J., and S. Ng (2005): "Understanding and Comparing Factor-Based Forecasts," *International Journal of Central Banking*, 1(3), 117–151.

Fernández, C., E. Ley, and M. F. Steel (2001): "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100(2), 381–427.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005): "The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting," *Journal of the American Statistical Association*, 100(471), 830–840.

George, E. I., and R. E. McCulloch (1997): "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Green, P. J. (1995): "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82(4), 711–732.

Jacobson, T., and S. Karlsson (2004): "Finding Good Predictors for Inflation: A Bayesian Model Averaging Approach," *Journal of Forecasting*, 23(7), 479–496.

Koop, G., and S. Potter (2004): "Forecasting in Dynamic Factor Models Using Bayesian Model Averaging," *Econometrics Journal*, 7(2), 550–565.

Stock, J. H., and M. W. Watson (2002a): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97(460), 1167 – 1179.

——— (2002b): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics*, 20(2), 147 – 162.

——— (2003): "Has the Business Cycle Changed and Why?," in *NBER Macroeconomics Annual 2002*, ed. by M. Gertler, and K. S. Rogoff. The MIT Press.

Wright, J. H. (2003a): "Bayesian Model Averaging and Exchange Rate Forecasts," International Finance Discussion Papers 779, Board of Governors of the Federal Reserve System.

——— (2003b): "Forecasting U.S. Inflation by Bayesian Model Averaging," International Finance Discussion Papers 780, Board of Governors of the Federal Reserve System.

Zellner, A. (1986): "On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, ed. by P. K. Goel, and A. Zellner, pp. 233–243. North-Holland, Amsterdam.

Appendix A

Data

The transformation codes for the time series are

Code	Transformation
1	level
2	4 quarters log difference ($\ln y_t - \ln y_{t-4}$)
3	4 quarters growth rate ($y_t - y_{t-4}$)
4	4 quarters percentage change ($(y_t - y_{t-4})/y_{t-4}$)

Table A.1 Financial variables

	Variable	Description	Transf.
1.	GovDebt	Government debt	2
2.	AFGX	Affärsvärlden stock index	2
3.	REPO	Repo rate	1
4.	DISK	Discount rate	1
5.	R3M	3 month money market rate	1
6.	R5Y	5 year government bond rate	1
7.	R10Y	10 year government bond rate	1
8.	GBor	Central government borrowing requirement	1
9.	RsTCW	Short rate (TCW)	1
10.	RlTCW	Long rate (TCW)	1

Table A.2 Exchange rates

	Variable	Description	Transf.
11.	NFX	Effective exchange rate (TCW)	2
12.	RFX	Effective real exchange rate (TCW)	2
13.	USD	SEK/USD exchange rate	2
14.	DEM	SEK/DEM exchange rate	2

Table A.3 Money supply

	Variable	Description	Transf.
15.	M0	Narrow money	2
16.	M3	Broad money	2

Table A.4 Labor costs

	Variable	Description	Transf.
17.	WCSS	Wages incl. social security	2
18.	WgCst	Wages excl. social security	2
19.	WageMM	Hourly wages, mining and manufacturing	2
20.	HLCInd	Hourly labor cost: total industry	2

Table A.5 Population

	Variable	Description	Transf.
21.	PpTot	Total population	2
22.	Pp1664	Share in ages 16-64	2
23.	Pp014	Share in ages 0-14	2
24.	Pp1529	Share in ages 15-29	2
25.	Pp2534	Share in ages 25-34	2
26.	Pp3049	Share in ages 30-49	2
27.	Pp5064	Share in ages 50-64	2
28.	Pp6574	Share in ages 65-74	2
29.	Pp75+	Share 75 and older	2

Table A.6 Labor market variables

	Variable	Description	Transf.
30.	AvJob	# of available jobs	2
31.	LabFrc	# in labor force	2
32.	NLFrc	# not in labor force	2
33.	RelLF	LabFrc/Pp1664	1
34.	Empld	# employed	2
35.	PrvEmp	# privatly employed	2
36.	PubEmp	# publicly employed	2
37.	Av4Wrk	# available for work	2
38.	NA4Wrk	# not available for work	2
39.	NUnemp	# unemployed	2
40.	Unemp	Unemployment	1
41.	U02W	# unemployed $<$ 2 weeks	3
42.	U314W	# unemployed 3 - 14 weeks	3
43.	U1552W	# unemployed 15 - 52 weeks	3
44.	U52W+	# unemployed more than 52 weeks	3
45.	NewJob	New jobs	3

Table A.7 Real activity and Expectations

	Variable	Description	Transf.
46.	IndProd	Industrial production	4
47.	NewCar	New cars	1
48.	NewHouse	New single family houses	1
49.	HourWork	Hours worked	2
50.	GDP	GDP	2
51.	RGDP	Real GDP	2
52.	NAIRU	NAIRU	1
53.	OutGap	Output gap	1
54.	ProdGap	Production gap	1
55.	BCI	Business confidence indicator	1
56.	HExpSWE	Household exp. Swedish economy	1
57.	HExpOwn	Household exp. own economy	1
58.	GDPTCW	TCW-weighted GDP	2

Table A.8 Prices

	Variable	Description	Transf.
59.	InfFor	Foreign CPI (TCW)	4
60.	InfRel	Relative CPI	4
61.	PPP	Real exchange rate	4
62.	Infla	Swedish CPI	4
63.	InfNet	Swedish NPI	4
64.	InfHse	House price index	4
65.	MrtWgh	Weight of mortgage interest in CPI	1
66.	InfUnd	Underlying inflation	4
67.	InfFd	Food component of CPI	4
68.	InfFl	Housing fuel and electricity comp. of CPI	4
69.	InfHWg	Factor price index, housing incl. wages	4
70.	InfCns	Construction cost index	4
71.	InfPrd	Producer price index	4
72.	InfImpP	Import price index	4
73.	InfExp	Export price index	4
74.	InfTCW	TCW-weighted Swedish CPI	4
75.	ExpInf	Households exp. of inflation 1 year from now	1
76.	POilUSD	Oil price, USD	4
77.	POilSEK	Oil price, SEK	4

Chapter 2

Forecast Combination and Model Averaging Using Predictive Measures

Acknowledgement: Earlier versions of this paper have been presented at the 2004 International Symposium on Forecasting in Sydney and the EABCN workshop on Recent Advances in Forecast Combination, Brussels, 2004. We are grateful for useful comments from the participants in these conferences and from two anonymous referees. Financial support from Sveriges Riksbank is gratefully acknowledged.

2.1 Introduction

Following Bates and Granger (1969) forecast combination has proven to be a highly successful forecasting strategy. Examples of formal evaluations of forecast methods, where forecast combination has performed well, include the M-competitions (Makridakis, Andersen, Carbone, Fildes, Hibon, Lewandowski, Newton, Parzen, and Winkler (1982), Makridakis, Chatfield, Hibon, Lawrence, Mills, Ord, and Simmons (1993) and Makridakis and Hibon (2000)), and with a focus on macroeconomic forecasting, Stock and Watson (1999). Much of this success can be attributed to the robustness of forecast combination. By combining forecasts from several models we implicitly acknowledge that more than one model could provide good forecasts and we guard against misspecification by not putting all the weight on one single model. While the literature on forecast combination is extensive, see Clemen (1989) and Timmermann (2006) for reviews, and Hendry and Clements (2004), and Elliott and Timmermann (2004) for recent theoretical contributions, relatively little attention has been given to the use of predictive measures of fit as the base for forecast combination. In this chapter we propose the predictive likelihood as the basis for Bayesian model averaging (BMA) and forecast combination.

We adopt a Bayesian approach, since BMA is an ideal framework for forecast combination. It provides a rigorous statistical foundation, where the weights assigned to the different forecasts arise naturally as posterior probabilities of the models, and the combined forecast has appealing optimality properties given the set of models considered (Min and Zellner (1993), Madigan and Raftery (1994)). In addition BMA accounts for the model uncertainty and it is easy to construct prediction intervals taking account of model uncertainty as well as parameter uncertainty.

The specific forecasting situation we consider is similar to the one studied by Stock and Watson (2002b), i.e. where there is a wealth of potential predictor variables. For computational simplicity we use linear regression models, but in contrast to Stock and Watson we consider the models that arise when taking all possible combinations of the predictor variables. An efficient summary of the forecast content of the predictors is then provided by the model averaged forecast from these models. In previous work Jacobson and Karlsson (2004) find this approach to work well when the forecast combinations are based on the marginal likelihood. In related work Koop and Potter (2004) use BMA for forecasting in large macroeconomic panels using models based on principal components. They conclude that the gain in forecasting performance from the use of principal components is small relative to the gains from BMA. See Stock and Watson (2006) for an overview and additional references.

While the previous studies all apply BMA in a standard fashion using the marginal likelihood, we propose the use of predictive measures of fit[1] and, in particular, the predictive likelihood as a natural basis for forecast combination. In addition to the intuitive appeal, the use of the predictive likelihood relaxes the requirement to specify proper priors for the parameters of each model. In this sense, our work is closely related to the literature on minimally informative priors.

The use of predictive measures leads to some additional practical concerns compared to model averaging based on in-sample measures of fit. In order to calculate the weights for the combined forecast a hold-out sample of l observations is needed for the predictive likelihood. The number of observations available for estimation is thus reduced from T to $T - l$ and there is clearly a trade off involved in the choice of l. The predictive measure becomes less erratic as l increases, which should improve the performance of the procedure. Estimation, on the other hand, is performed without taking the most recent observations into account, which might have a detrimental effect[2].

In general, the weights assigned to the forecasts should have some of the properties of consistent model selection procedures, i.e. if there is a correct model this should receive more weight as the sample evidence accumulates and ultimately all the weight. On the other hand we want the weights to retain the robustness property of forecast combination in finite samples and guard against the overconfidence in a single model that can arise from overfitting the data. We show that the use of the predictive likelihood leads to consistent model selection. In addition we give an intuitively appealing interpretation of the predictive likelihood, indicating that it will have good small sample properties. The latter claim is supported by a simulation study and our empirical application.

The remainder of the chapter is organized as follows. The next section introduces the Bayesian model averaging technique and predictive densities, Section 2.3 presents several Bayes factors and their asymptotics. Section 2.4 studies the small sample properties of the predictive likelihood. Section 2.5 contains a simulation study, Section 2.6 an application to forecasts of the Swedish inflation and Section 2.7 concludes.

[1] See Laud and Ibrahim (1995) for a discussion of different predictive measures in a Bayesian context.

[2] This is an issue only for the calculation of the weights. The forecast from each model used in the forecast combination is based on the full sample.

2.2 Forecast combination using Bayesian model averaging

The standard approach to forecast combination using BMA operates as follows. Let $\mathfrak{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$ be the set of forecasting models under consideration, with a prior probability for each model, $p(\mathcal{M}_i)$, prior distribution of the parameters in each model, $p(\boldsymbol{\theta}_i|\,\mathcal{M}_i)$, and likelihood function $L(\mathbf{y}|\,\boldsymbol{\theta}_i, \mathcal{M}_i)$. The posterior probabilities of the models after observing the data $\mathbf{y}$ follow from Bayes rule

$$p(\mathcal{M}_i|\,\mathbf{y}) = \frac{m(\mathbf{y}|\,\mathcal{M}_i)\,p(\mathcal{M}_i)}{\sum_{j=1}^{M} m(\mathbf{y}|\,\mathcal{M}_j)\,p(\mathcal{M}_j)}, \tag{2.1}$$

where

$$m(\mathbf{y}|\,\mathcal{M}_i) = \int L(\mathbf{y}|\,\boldsymbol{\theta}_i, \mathcal{M}_i)\,p(\boldsymbol{\theta}_i|\,\mathcal{M}_i)\,\mathrm{d}\boldsymbol{\theta}_i \tag{2.2}$$

is the marginal likelihood for model $\mathcal{M}_i$. All knowledge about some quantity of interest, ϕ, when taking account of model uncertainty, is summarized in its posterior $p(\phi|\,\mathbf{y})$, which is given by

$$p(\phi|\,\mathbf{y}) = \sum_{j=1}^{M} p(\phi|\,\mathbf{y}, \mathcal{M}_j)\,p(\mathcal{M}_j|\,\mathbf{y}). \tag{2.3}$$

This is simply an average of the posterior distribution under each of the models, weighted by posterior model probabilities. Alternatively, if $g(\phi)$ is a function of ϕ, then by the rules of conditional expectation

$$\mathrm{E}[g(\phi)|\,\mathbf{y}] = \sum_{j=1}^{M} \mathrm{E}[g(\phi)|\,\mathbf{y}, \mathcal{M}_j]\,p(\mathcal{M}_j|\,\mathbf{y}). \tag{2.4}$$

In particular, the minimum mean squared error forecast is given by

$$y_{T+h|T} = \mathrm{E}(y_{T+h}|\,\mathbf{y}) = \sum_{j=1}^{M} \mathrm{E}(y_{T+h}|\,\mathbf{y},\mathcal{M}_j)\,p(\mathcal{M}_j|\,\mathbf{y}), \tag{2.5}$$

where $\mathrm{E}(y_{T+h}|\,\mathbf{y},\mathcal{M}_j)$ is the forecast conditional on model $\mathcal{M}_j$. The optimal forecast is, in other words, given by a forecast combination using the posterior model probabilities as weights.

It is clear from (2.1) that the conversion of prior model probabilities into posterior probabilities is determined by the marginal likelihood. While this leads to optimal forecasts, conditional on the true model being included in

the set of models, it raises the possibility that the forecast combination is adversely affected by in-sample overfitting of the data. The problem of in-sample overfitting of the data might seem counter-intuitive as the marginal likelihood is commonly interpreted as an out-of-sample or predictive measure of fit. The interpretation as a predictive measure relies on the prior having a predictive content, i.e. that the prior is informative. In our application and in large scale model selection or model averaging exercises in general it is not possible to provide well thought out priors for all models. Instead default, uninformative, priors such as the prior suggested by Fernández, Ley, and Steel (2001) are used, and the marginal likelihood essentially reduces to an in-sample measure of fit. In our case, with an uninformative g-type prior similar to the prior of Fernández, Ley, and Steel (2001), the marginal likelihood is a function of the residual sum of squares from a least squares fit and can be viewed as a pure in-sample measure of fit.

A natural remedy for the problem of in-sample overfitting is to explicitly consider the out-of-sample, or predictive, performance of the models. Split the sample $\mathbf{y} = (y_1, y_2, \ldots, y_T)'$ into two parts with m and l observations, with $T = m + l$. That is, let

$$\mathbf{y}_{T\times 1} = \begin{bmatrix} \mathbf{y}^*_{m\times 1} \\ \widetilde{\mathbf{y}}_{l\times 1} \end{bmatrix}, \tag{2.6}$$

where the first part of the data is used to convert the parameter priors $p(\boldsymbol{\theta}_i|\,\mathcal{M}_i)$ into the posterior distributions, and the second part of the sample is used for evaluating the model performance.

In particular, the *posterior* predictive density of $\widetilde{\mathbf{y}} = (y_{m+1}, y_{m+2}, \ldots, y_T)'$, conditional on $\mathbf{y}^* = (y_1, y_2, \ldots, y_m)'$ and model $\mathcal{M}_i$, is

$$p(\widetilde{\mathbf{y}}|\,\mathbf{y}^*,\mathcal{M}_i) = \int_{\boldsymbol{\theta}_i} L(\widetilde{\mathbf{y}}|\,\boldsymbol{\theta}_i, \mathbf{y}^*, \mathcal{M}_i)\, p(\boldsymbol{\theta}_i|\,\mathbf{y}^*, \mathcal{M}_i)\, \mathrm{d}\boldsymbol{\theta}_i, \tag{2.7}$$

where $p(\boldsymbol{\theta}_i|\,\mathbf{y}^*, \mathcal{M}_i)$ is the posterior distribution of the parameters and $L(\widetilde{\mathbf{y}}|\,\boldsymbol{\theta}_i, \mathbf{y}^*, \mathcal{M}_i)$ is the likelihood. The density of the data is averaged with respect to the *posterior* knowledge of the parameters. The predictive density gives the distribution of future observations, $y_{m+1}, y_{m+2}, \ldots, y_T$, conditional on the observed sample $\mathbf{y}^*$. After observing $\widetilde{\mathbf{y}}$, the expression (2.7) is a real number, the predictive likelihood. It indicates how well model $\mathcal{M}_i$ accounted for the realizations $y_{m+1}, y_{m+2}, \ldots, y_T$. A good model will have a large value of $p(\widetilde{\mathbf{y}}|\,\mathbf{y}^*,\mathcal{M}_i)$.

By replacing the marginal likelihood in (2.1) with the posterior predictive

density (2.7), the *predictive weight*[3] for model $\mathcal{M}_i$ can be expressed as

$$w\left(\mathcal{M}_i|\,\widetilde{\mathbf{y}},\mathbf{y}^*\right)=\frac{p\left(\widetilde{\mathbf{y}}|\,\mathbf{y}^*,\mathcal{M}_i\right)p\left(\mathcal{M}_i\right)}{\sum_{j=1}^{M}p\left(\widetilde{\mathbf{y}}|\,\mathbf{y}^*,\mathcal{M}_j\right)p\left(\mathcal{M}_j\right)}. \tag{2.8}$$

The forecast combination based on predictive likelihood weights is then obtained by substituting $w\left(\mathcal{M}_j|\,\widetilde{\mathbf{y}},\mathbf{y}^*\right)$ for $p\left(\mathcal{M}_j|\,\mathbf{y}\right)$ in (2.5). Note that the forecast from a single model is still based on the full sample posterior distribution of the parameters.

The partitioning of the data in a *training* sample $\mathbf{y}^*$ and a *hold-out* sample $\widetilde{\mathbf{y}}$ in (2.6) is natural for time series data. This is obviously not the only way to partition the data and other approaches may be more appropriate at times, Gelfand and Dey (1994) provide a typology of the various forms the predictive likelihood can take:

1. $\widetilde{\mathbf{y}}=\mathbf{y}$, $\mathbf{y}^*=\varnothing$, which yields the marginal density, $m\left(\mathbf{y}\right)$, of the data.
2. $\widetilde{\mathbf{y}}=\{y_r\}$, $\mathbf{y}^*=\mathbf{y}_{\backslash r}=\left(y_1,y_2,\ldots,y_{r-1},y_{r+1},\ldots,y_T\right)$, which yields the cross-validation density $p\left(y_r|\,\mathbf{y}_{\backslash r},\mathcal{M}_i\right)$, as in Stone (1974) or Geisser (1975).
3. $\widetilde{\mathbf{y}}$ contains usually two or three observations, $\mathbf{y}^*=\mathbf{y}-\widetilde{\mathbf{y}}$, extending the point 2, as in Peña and Tiao (1992).
4. $\widetilde{\mathbf{y}}=\mathbf{y}$, $\mathbf{y}^*=\mathbf{y}$, which yields the posterior predictive density defined in Aitkin (1991).
5. $\widetilde{\mathbf{y}}=\mathbf{y}\backslash\mathbf{y}^*$, $\mathbf{y}^*=\left(y_1,y_2,\ldots,y_{\lfloor\rho T\rfloor}\right)$, where $\backslash$ denotes set difference and $\lfloor\,\rfloor$ denotes the greatest integer function; here a proportion ρ of the observation is set aside for prior updating with the remainder used for model determination, as suggested by O'Hagan (1991).
6. $\widetilde{\mathbf{y}}=\mathbf{y}\backslash\mathbf{y}^*$, $\mathbf{y}^*$ is a minimal subset, i.e. the least number of data points such that $p\left(\boldsymbol{\theta}_i|\,\mathbf{y}^*,\mathcal{M}_i\right)$ is a proper density. The problem of selecting a particular training sample is avoided by averaging over all possible minimal subsets, as suggested by Berger and Pericchi (1996).

The main motivation for these alternatives is that they can be used with improper priors on the parameters. An adequate choice of $\mathbf{y}^*$ removes the impropriety of $p\left(\boldsymbol{\theta}_i|\,\mathbf{y}^*,\mathcal{M}_i\right)$ and therefore the posterior predictive density

[3]We use the notation $w\left(\mathcal{M}_i|\,\widetilde{\mathbf{y}},\mathbf{y}^*\right)$ to emphasize that these are not posterior model probabilities and should only be interpreted as a measure of predictive performance.

$p(\widetilde{\mathbf{y}}|\,\mathbf{y}^*, \mathcal{M}_i)$ does not diverge and can be calculated. As such, any of them can be expected to overcome most of the difficulties associated with the use of the marginal likelihood when the parameter priors are uninformative. In a forecasting setting we find the simple sample split (2.6) most natural. The first part of the data, the training sample $\mathbf{y}^*$, is used to obtain posterior distributions $p(\boldsymbol{\theta}_i|\,\mathbf{y}^*, \mathcal{M}_i)$. The updated prior distributions are then used for assessing the fit of the model to the data $\widetilde{\mathbf{y}}$.

As the hold-out sample size, l, increases, that is the size of the training sample m decreases, the predictive measure will be more stable and should perform better up to a point, where the predictive distribution becomes diffuse for all models and is unable to discriminate. Berger and Pericchi (1996) favor minimal training samples in order to devote as much data as possible to the model comparison.

Related procedures of interest include the intrinsic Bayes factor of Berger and Pericchi (1996), based on the considerations in item 6, and the fractional Bayes factor of O'Hagan (1995), which does not rely on dividing the data into training and hold-out samples. Bernardo and Smith (1994, section 6.1.6) build on the cross-validation idea of item 2 to suggest a procedure for model choice when the true data generating process is not assumed to be included in the model set. The idea is to directly evaluate the posterior expectation loss function under the different models via cross-validation. In our setting, with a very large number of possible models and in a forecasting context, the procedure suffers from two problems. Firstly, cross-validation relies on an exchangeability argument that is unreasonable with time series data. Secondly, this becomes extremely computationally demanding as the number of models increases. In addition, the procedure does not lend itself easily to model averaging or forecast combination. An exhaustive evaluation of these alternative approaches is beyond the scope of this chapter and is left for future research.

2.3 Model choice and large sample properties

Ideally, the weights assigned to the forecasts should act as consistent model selection criteria. The weight, or posterior probability, of the true model should approach unity as the sample size increases. As any non-dogmatic prior over the models is irrelevant asymptotically, it suffices to study the Bayes factor of model $\mathcal{M}_i$ against model $\mathcal{M}_j$

$$BF_{ij} = \frac{P(\mathcal{M}_i|\mathbf{y})}{P(\mathcal{M}_j|\mathbf{y})} \Big/ \frac{P(\mathcal{M}_i)}{P(\mathcal{M}_j)} = \frac{m(\mathbf{y}|\,\mathcal{M}_i)}{m(\mathbf{y}|\,\mathcal{M}_j)}. \tag{2.9}$$

Correspondingly, the predictive odds ratio in favor of model $\mathcal{M}_i$ versus model $\mathcal{M}_j$ for the future observations $m+1$ through T is

$$\frac{P(\mathcal{M}_i)}{P(\mathcal{M}_j)} PBF_{ij}(\widetilde{\mathbf{y}}|\,\mathbf{y}^*), \tag{2.10}$$

where

$$PBF_{ij}(\widetilde{\mathbf{y}}|\,\mathbf{y}^*) = \frac{p(\widetilde{\mathbf{y}}|\,\mathbf{y}^*,\mathcal{M}_i)}{p(\widetilde{\mathbf{y}}|\,\mathbf{y}^*,\mathcal{M}_j)} = \frac{m(\mathbf{y}|\,\mathcal{M}_i)}{m(\mathbf{y}|\,\mathcal{M}_j)} \Big/ \frac{m(\mathbf{y}^*|\,\mathcal{M}_i)}{m(\mathbf{y}^*|\,\mathcal{M}_j)}. \tag{2.11}$$

O'Hagan (1995) defines (2.11) as the *partial* Bayes factor (PBF) and points out that the PBF is less sensitive to the choice of the prior distribution than the Bayes factor (2.9), and that the PBF does not depend on arbitrary constants when improper priors are used.

Gelfand and Dey (1994) show that Bayes factor (2.9) is a consistent model selection criterion under the assumption that the true model is included in the model set. In a recent paper Fernández-Villaverde and Rubio-Ramírez (2004) extended this to the case when all the models are misspecified and show that, asymptotically, the Bayes factor will select the model that minimizes the Kullback-Leibler distance to the true data density. That is, the best approximation to the true model is selected.

O'Hagan (1995) establishes that model choice based on the partial Bayes factor (2.11) is consistent provided that $l/m \to \infty$. That is, when the hold-out sample grows faster than the training sample or the training sample is fixed. We conjecture that the results of Fernández-Villaverde and Rubio-Ramírez (2004) apply to the partial Bayes factor as well, possibly with a rate condition on the limiting behavior of l/m. Consequently, the partial Bayes factor will select the best approximation to the true model out of a set of misspecified models.

2.4 Small sample properties

Turning to the small sample properties, we concentrate the analysis on the linear regression models we use in the forecasting exercises. Consider a linear regression model with an intercept α and k regressors

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{2.12}$$

where $\boldsymbol{\gamma} = (\alpha, \boldsymbol{\beta}')'$, $\mathbf{Z} = (\boldsymbol{\iota}, \mathbf{X})$ and $\boldsymbol{\varepsilon}$ is a vector of $\mathrm{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$ disturbances. Partitioning $\mathbf{Z}$ conformably with (2.6) into the training and hold-out samples

we use a g-prior for the regression parameters

$$\boldsymbol{\gamma}|\,\sigma^2 \sim \mathrm{N}\left(\mathbf{0}, c\sigma^2\left(\mathbf{Z}^{*\prime}\mathbf{Z}^{*}\right)^{-1}\right), \tag{2.13}$$

with a large value of c. That is, the prior mean is set to zero indicating modest shrinkage of the posterior towards zero and the prior variance is proportional to the information in the training sample. For the variance the usual uninformative prior is used

$$p\left(\sigma^2\right) \propto 1/\sigma^2. \tag{2.14}$$

This gives the predictive density for $\widetilde{\mathbf{y}}$

$$\begin{aligned} p\left(\widetilde{\mathbf{y}}|\,\mathbf{y}^{*}\right) \propto{}& \left(S^{*}\right)^{\frac{m}{2}}\left|\mathbf{I}_l+\widetilde{\mathbf{Z}}\left(\mathbf{M}^{*}\right)^{-1}\widetilde{\mathbf{Z}}\right|^{-\frac{1}{2}} \\ &\times\left[S^{*}+\left(\widetilde{\mathbf{y}}-\widetilde{\mathbf{Z}}\boldsymbol{\gamma}_1\right)'\left(\mathbf{I}_l+\widetilde{\mathbf{Z}}\left(\mathbf{M}^{*}\right)^{-1}\widetilde{\mathbf{Z}}\right)^{-1}\left(\widetilde{\mathbf{y}}-\widetilde{\mathbf{Z}}\boldsymbol{\gamma}_1\right)\right]^{-(m+l)/2} \end{aligned} \tag{2.15}$$

with $S^{*}=\frac{c}{c+1}\left(\mathbf{y}^{*}-\mathbf{Z}^{*}\hat{\boldsymbol{\gamma}}^{*}\right)'\left(\mathbf{y}^{*}-\mathbf{Z}^{*}\hat{\boldsymbol{\gamma}}^{*}\right)+\frac{1}{c+1}\mathbf{y}^{*\prime}\mathbf{y}^{*}$ and $\mathbf{M}^{*}=\frac{c+1}{c}\mathbf{Z}^{*\prime}\mathbf{Z}^{*}$. See Appendix A for further details.

A slight reformulation of the predictive density (2.15) is quite revealing,

$$\begin{aligned} p\left(\widetilde{\mathbf{y}}|\,\mathbf{y}^{*}\right) \propto{}& \left(\frac{S^{*}}{m}\right)^{-l/2}\frac{\left|\mathbf{M}^{*}\right|^{\frac{1}{2}}}{\left|\mathbf{M}^{*}+\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}\right|^{\frac{1}{2}}} \\ &\times\left[m+\frac{1}{S^{*}/m}\left(\widetilde{\mathbf{y}}-\widetilde{\mathbf{Z}}\boldsymbol{\gamma}_1\right)'\left(\mathbf{I}+\widetilde{\mathbf{Z}}\left(\mathbf{M}^{*}\right)^{-1}\widetilde{\mathbf{Z}}'\right)^{-1}\left(\widetilde{\mathbf{y}}-\widetilde{\mathbf{Z}}\boldsymbol{\gamma}_1\right)\right]^{-(m+l)/2}, \end{aligned} \tag{2.16}$$

and shows that the predictive likelihood can be decomposed into three components.

1. The in-sample fit over the training sample is measured by $\left(\frac{S^{*}}{m}\right)^{-l/2}$. Comparing the in-sample fit of two models by this criterion, $\left(S_i^{*}/S_j^{*}\right)^{-l/2}$, it is clear that the effect of differences in fit is increasing in l, the size of the hold-out sample.

2. A penalty for the size of the model is provided by $\left|\mathbf{M}^{*}\right|^{1/2}\Big/\left|\mathbf{M}^{*}+\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}\right|^{1/2}$. We have $\mathbf{M}^{*}=\frac{c+1}{c}\mathbf{Z}^{*\prime}\mathbf{Z}^{*}$ and $\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}\approx\frac{l}{m}\mathbf{Z}^{*\prime}\mathbf{Z}^{*}$, which gives

$$\left|\mathbf{M}^{*}\right|=\left|\frac{c+1}{c}\mathbf{Z}^{*\prime}\mathbf{Z}^{*}\right|=\left(\frac{c+1}{c}\right)^{k+1}\left|\mathbf{Z}^{*\prime}\mathbf{Z}^{*}\right| \tag{2.17}$$

and

$$\left|\mathbf{M}^* + \widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}\right| \approx \left|\frac{c+1}{c}\mathbf{Z}^{*\prime}\mathbf{Z}^* + \frac{l}{m}\mathbf{Z}^{*\prime}\mathbf{Z}^*\right| = \left(\frac{1+c\left(1+\frac{l}{m}\right)}{c}\right)^{k+1}\left|\mathbf{Z}^{*\prime}\mathbf{Z}^*\right|. \tag{2.18}$$

For large values of c we can then approximate the ratio of determinants by

$$\frac{|\mathbf{M}^*|^{1/2}}{\left|\mathbf{M}^* + \widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}\right|^{1/2}} \approx \left(\frac{c+1}{1+c\left(1+\frac{l}{m}\right)}\right)^{\frac{k+1}{2}} \approx \left(1+\frac{l}{m}\right)^{-\frac{k+1}{2}} = \left(\frac{m+l}{m}\right)^{-\frac{k+1}{2}}. \tag{2.19}$$

This penalty for size is relatively modest and increasing in l.

3. The out-of-sample forecasting accuracy is measured by

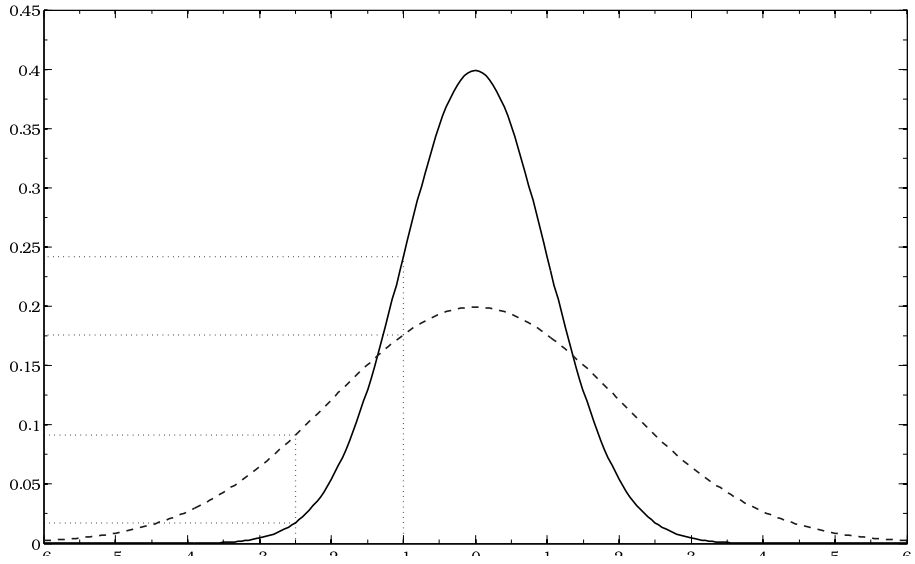
$$\left[m + \frac{1}{S^*/m}\left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}_1\right)'\left(\mathbf{I} + \widetilde{\mathbf{Z}}\left(\mathbf{M}^*\right)^{-1}\widetilde{\mathbf{Z}}'\right)^{-1}\left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}_1\right)\right]^{-(m+l)/2}. \tag{2.20}$$

It is especially noteworthy that the forecast error is relative to the forecast error variance implied by the model. In this sense the predictive likelihood is quite different from, say, ranking the models according to the root mean square forecast error (RMSFE).

Figure 2.1 illustrates the overall behavior for a model with good in-sample fit and corresponding small forecast error variance, and a model with poor in-sample fit and large forecast error variance. If the forecast error is modest, as can be expected from a model with small forecast variance, the model with smaller forecast error variance is preferred. If, on the other hand, the forecast error is larger than can be expected from the model with good in-sample fit this indicates that the model is overfitting the data and the model with relatively poor fit but a realistic prediction interval is preferred. The apparent in-sample overfitting and poor out-of-sample forecast may also be due to breaks in the parameters of the model. The predictive likelihood will thus penalize models with unstable parameters and give preference to models that are stable over time. This penalty is obvious when a break is close to the split between training and hold-out samples, but the penalty may also be substantial when the break occurs in the hold-out or training samples. This issue will be investigated in more detail in the Monte Carlo experiments.

The contribution of all three components of the predictive likelihood increase with l, the size of the hold-out sample, given a fixed total sample size

Figure 2.1 Predictive likelihood for models with small and large prediction error variance.



$T = m + l$. They do, however, increase with different rates and it is not clear what the appropriate finite sample trade off between m and l is. Asymptotic arguments indicate that l should be large relative to m.

2.5 Monte Carlo study

We use a Monte Carlo study to investigate some aspects of the small sample performance of forecast combinations based on the predictive likelihood as well as the traditional in-sample marginal likelihood. In particular we aim to shed some light on two issues. The appropriate choice of m and l for common sample sizes and how the procedures cope with the likely case that the true model is not included in the set of considered models. The second issue is investigated in two ways. First by assuming that some of the variables in the true model are unavailable to the investigator and secondly by introducing a shift in the parameters of the true model while only considering constant parameter models.

The design of the experiment is based on Fernández, Ley, and Steel (2001). We generate a matrix of 15 predictors $\mathbf{X}_{(T\times 15)}$, where the first 10 random variables, $\mathbf{x}_1, \ldots, \mathbf{x}_{10}$, are i.i.d. standard normal and then construct the

additional five variables according to

$$(\mathbf{x}_{11}, \ldots, \mathbf{x}_{15}) = (\mathbf{x}_1, \ldots, \mathbf{x}_5) \begin{pmatrix} 0.3 & 0.5 & 0.7 & 0.9 & 1.1 \end{pmatrix}' \boldsymbol{\iota} + \mathbf{E}, \qquad (2.21)$$

where $\boldsymbol{\iota}$ is a (1×5) vector of ones and $\mathbf{E}$ is a $(T \times 5)$ matrix of i.i.d. standard normals. This produces a correlation between the first five and the last five predictors. The dependent variable is generated as

$$y_t = 4 + 2x_{1,t} - x_{5,t} + 1.5x_{7,t} + x_{11,t} + 0.5x_{13,t} + \sigma\varepsilon_t, \qquad (2.22)$$

where the disturbances ε_t are i.i.d. standard normal and $\sigma = 2.5$. We consider three sample sizes, $T = 100$, 230 and 380, corresponding to roughly 25 years of quarterly data, 20 years of monthly data and 30 years of monthly data. In the remainder we will refer to these as the small, medium and large datasets. In each case we generate additional 20 observations that are set aside for the forecast evaluation.

The forecasts used in the evaluation are true out-of-sample forecasts, where the dataset is sequentially updated. That is, the first forecast for $t = 101$ is based on the first 100 observations, which are split into training and hold-out samples. The training sample is used to convert the prior into a posterior, the predictive likelihood is calculated for the hold-out sample and predictive weights are calculated as in (2.8), or posterior model probabilities as in (2.1) for the marginal likelihood based on the full sample. The model averaged forecasts are then formed using (2.5), where the forecast from each model is based on the posterior from the full sample of 100 observations. For the next forecast for $t = 102$, observation 101 is added to the data and the procedure is repeated. Note that the size of the hold-out sample is held constant for the 20 forecasted observations. This means that the training sample size increases as t increases.

The first set of simulation experiments are executed with all predictors available for variable selection. Here, as in all other cases of the simulation study, the constant term is always included in the model. This corresponds to the $\mathfrak{M}-$closed view of Bernardo and Smith (1994), when the true model is assumed to be part of the model set.

For the medium and the large datasets we conduct additional experiments where two of the variables, x_1 and x_7, in the true model (2.22) are excluded from the set of potential predictors. The true model is not in the model set and we can only hope to find a good approximation.[4] This corresponds to the $\mathfrak{M}-$open view of Bernardo and Smith.

[4]The best approximation, given a squared error loss, is of course the expectation of y conditional on the remaining variables.

Finally, we conduct one experiment for the medium sample size, $T = 230$, where all the variables are retained. Instead the coefficient of x_7 changes from 1.5 to -1.5 at the beginning ($t = 60$), middle ($t = 125$) or end of the data ($t = 190$). This again corresponds to the $\mathfrak{M}-$open view, but with the added complication that no constant parameter model will provide a good approximation both before and after the break.

For each sample size we generate 100 independent samples of the explanatory variables $\mathbf{X}$ and the dependent variable $\mathbf{y}$ in order to avoid sample dependent results. For each dataset 20 forecasts are calculated using the individual models and the forecast combinations. The estimated RMSFE of the different procedures is the average of the RMSFE from the 100 datasets.

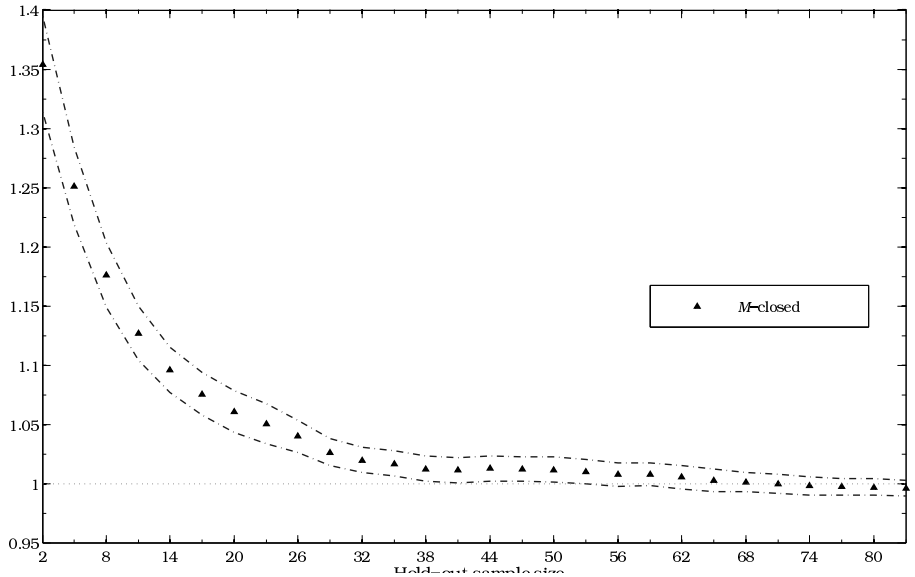
The prior specification is the same for all experiments. The prior on the models is given by

$$p(\mathcal{M}_i) \propto \delta^{k_i} (1-\delta)^{k'-k_i}, \tag{2.23}$$

where k_i is the number of variables included in model $\mathcal{M}_i$, the maximum number of variables in a model is $k' = 15$ (or 13 when x_1 and x_7 are dropped) and we set $\delta = 0.2$ corresponding to a prior expected model size of 3. The constant c in (2.13) is set to $(k')^3$. This choice is somewhat arbitrary, but motivated by two concerns. First, we want similar priors in the simulation study and the application. Secondly, we try to avoid making the parameter prior too uninformative. In the application $(k')^3$ yields a slightly more informative prior than the choice k^2 recommended by Fernández, Ley, and Steel (2001) while maintaining comparability with the simulation study.

For a large number of possible predictor variables it is too time consuming to actually calculate the predictive or marginal likelihood for every model. Instead we use a Markov chain to explore the model space. The chain is based on the reversible jump Markov chain Monte Carlo (RJMCMC) of Green (1995) and is designed to have the posterior model probabilities (2.1) or the predictive weights (2.8) as its stationary distribution. The details of the algorithm are given in Appendix B. While the chain will provide a simulation consistent estimate of the posterior probabilities or predictive weights we use it primarily as a device for identifying the set of practically relevant models. That is, models with sufficiently large weight to enter into the forecast combination in a meaningful way. To this end we take the set of relevant models to be the set of models visited by the chain, and exact weights for the forecast combinations are calculated conditional on this set of models using expression (2.1) or expression (2.8). A practically relevant issue in this context is that we run the chain long enough to account for most of the total posterior probability mass. We use the algorithm of George and McCulloch (1997) to estimate the

Figure 2.2 Ratio of RMSFE for predictive likelihood and marginal likelihood as a function of l for the simulated small dataset.



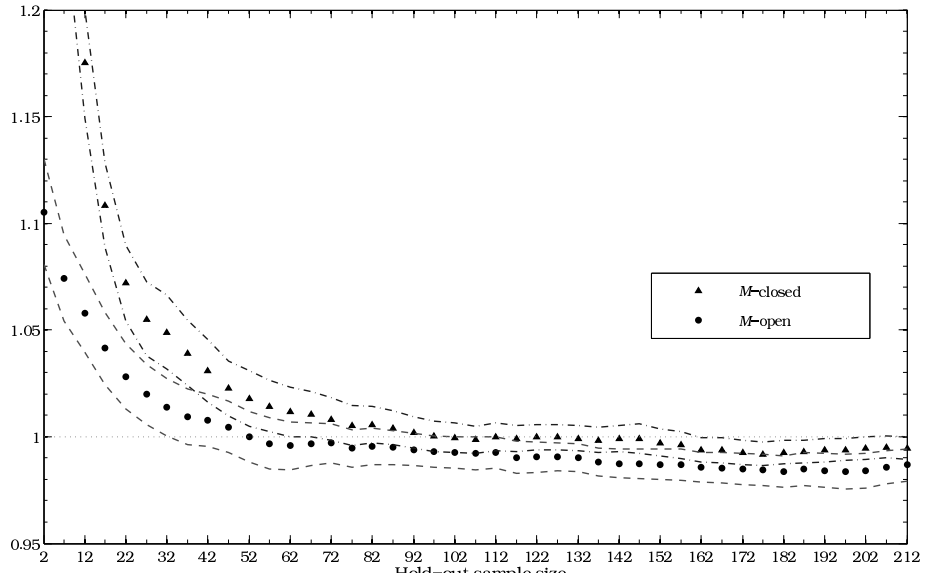
probability coverage of the chain. In the simulation experiments we run the chain for 70 000 replicates and discard the first 20 000 draws as burn-in.

2.5.1 Results for $\mathfrak{M}-$closed view and $\mathfrak{M}-$open view with constant parameters

For the simulated small dataset the simulations include different sizes of the hold-out sample, from $l = 2$, to $l = 83$ with increments of 3. For the medium dataset the hold-out sample size varies from $l = 2$, to $l = 212$ with an increment of 5. In the large dataset the hold-out sample size starts at $l = 2$ and ends at $l = 362$, with step 10.

The impact of l on the forecast accuracy for the $\mathfrak{M}$-closed view is presented in Tables C.1 - C.3, and for the $\mathfrak{M}$-open view in Tables C.4 and C.5 in Appendix C. Figures 2.2 - 2.4 plot the ratio of the average RMSFE (over the datasets) for the predictive likelihood to the average RMSFE for the marginal likelihood for the three sample sizes with pointwise 95% confidence intervals based on the Delta method. All the results show that the predictive likelihood RMSFE decreases as the size of the hold-out sample increases. For the small dataset the predictive likelihood provides a small but insignificant improvement on the

Figure 2.3 Ratio of RMSFE for predictive likelihood and marginal likelihood as a function of l for the simulated medium dataset.

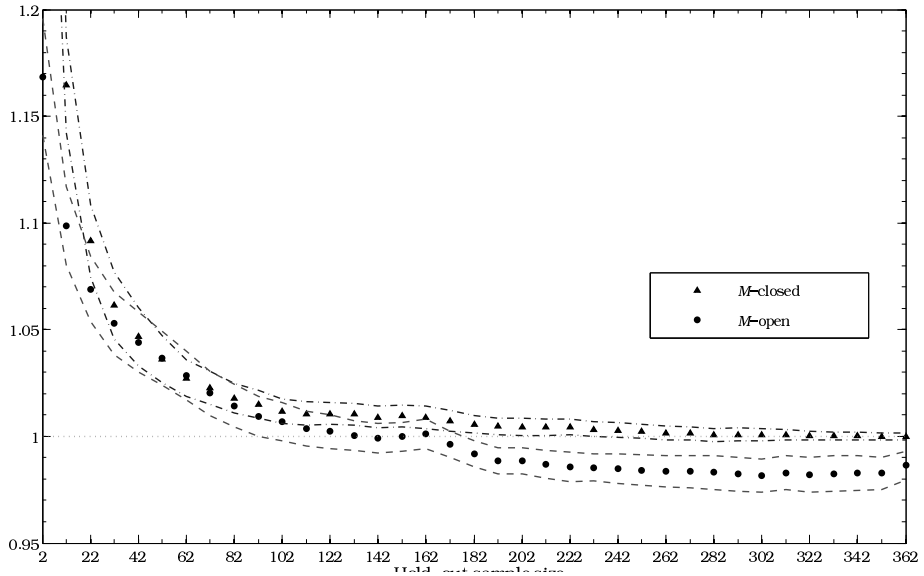


marginal likelihood for $l \geq 74$, indicating that at least 70% of the data should be left for model comparison.

In the case of the medium dataset the predictive likelihood outperforms the marginal likelihood for $l \geq 102$ (the RMSFE is significantly smaller for $l \geq 162$), with some indication of a minimum around $l = 177$ for the $\mathfrak{M}$-closed view. About 70% of the data for the hold-out sample seems to be appropriate in this case as well. For the large dataset the differences between the RMSFE of the marginal likelihood and predictive likelihood are moderate, but still indicating that 75% of the data is needed for the hold-out sample. The gains from using the predictive likelihood are modest in the $\mathfrak{M}$-closed case since the forecast combination based on the marginal likelihood is close to the best possible forecast in each case. The RMSFE is 2.5 when the true model with known parameters is used for forecasting. There is thus little room for improvement when the RMSFEs for the marginal likelihood are 2.61, 2.49 and 2.50 in the three experiments in the $\mathfrak{M}$-closed view.

The results for the large data in set also confirm the consistency results for the marginal and predictive likelihoods. The marginal likelihood assigns the highest posterior probability to the true model, on average the probability is 79.08% over the replicates. Similarly for the predictive likelihood, for $l = 362$,

Figure 2.4 Ratio of RMSFE for predictive likelihood and marginal likelihood as a function of l for the simulated large dataset.



the average probability of selecting the true model is 50.24%. With this large dataset, the forecast combinations are dominated by the forecast from the true model. In addition there is little posterior parameter uncertainty. As a result, both forecast combinations are close to the best possible forecast and there is a little to choose between them.

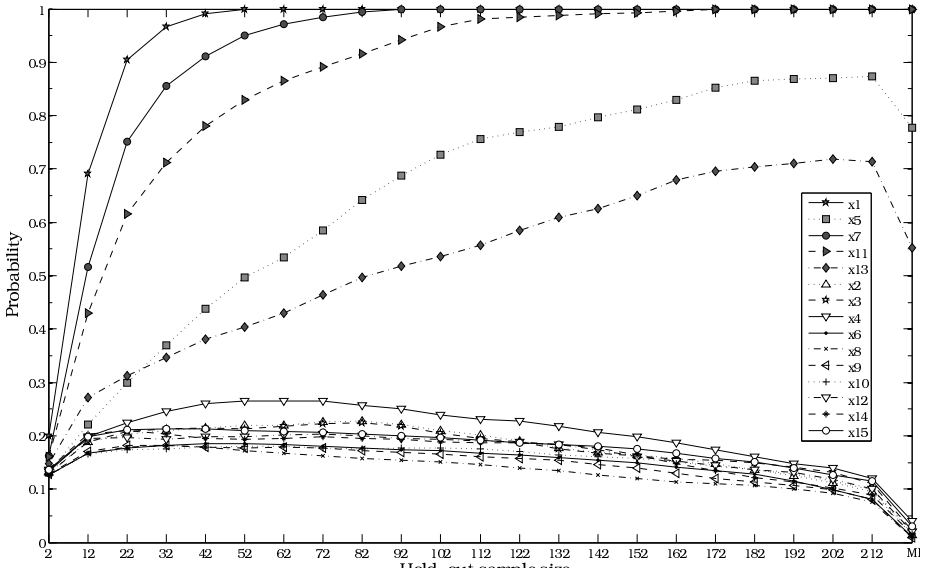
For the $\mathfrak{M}$-open view, when x_1 and x_7 are dropped from the data, the predictive likelihood outperforms the marginal likelihood by a greater margin and over a larger range of hold-out sample sizes. The predictive likelihood improves on the marginal likelihood for $l \geq 52$ (the reduction in RMSFE is significant for $l \geq 117$) with the medium dataset and $l \geq 172$ (significantly for $l \geq 182$) with the large dataset, with minima around $l = 182$ and $l = 302$, respectively. Again, using 70% of the data for the hold-out sample seems about right.

The better performance of predictive likelihood in the $\mathfrak{M}-$open view can be explained by the variable inclusion probabilities plotted in Figures 2.5 and 2.6. The posterior probability of a variable i being in the model is given by

$$p\left(x_i|\,\mathbf{y}\right) = \sum_{j=1}^{M} 1\left(x_i \in \mathcal{M}_j\right) p\left(\mathcal{M}_j|\,\mathbf{y}\right), \tag{2.24}$$

Figure 2.5 Variable inclusion probabilities (average) for medium dataset.

(a) $\mathfrak{M}-$closed view



(b) $\mathfrak{M}-$open view

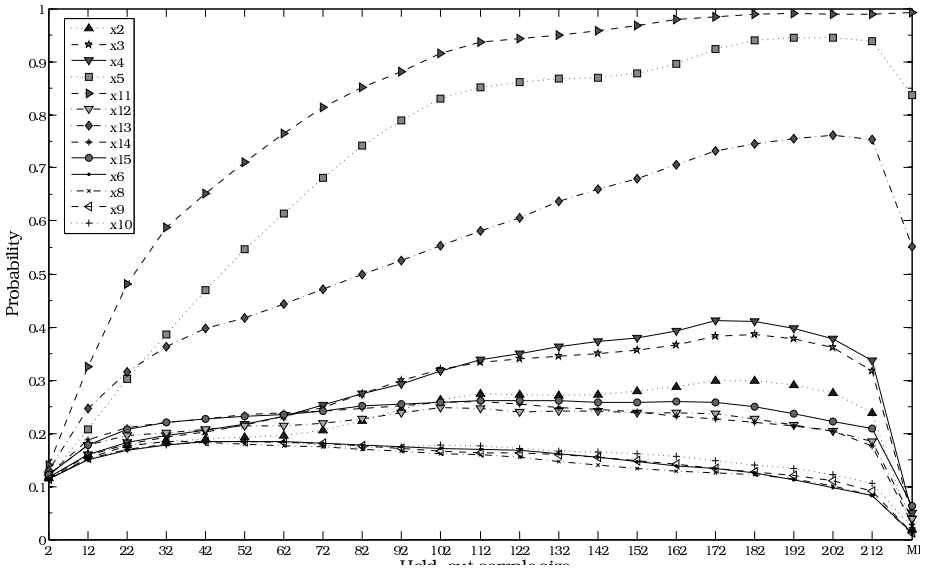
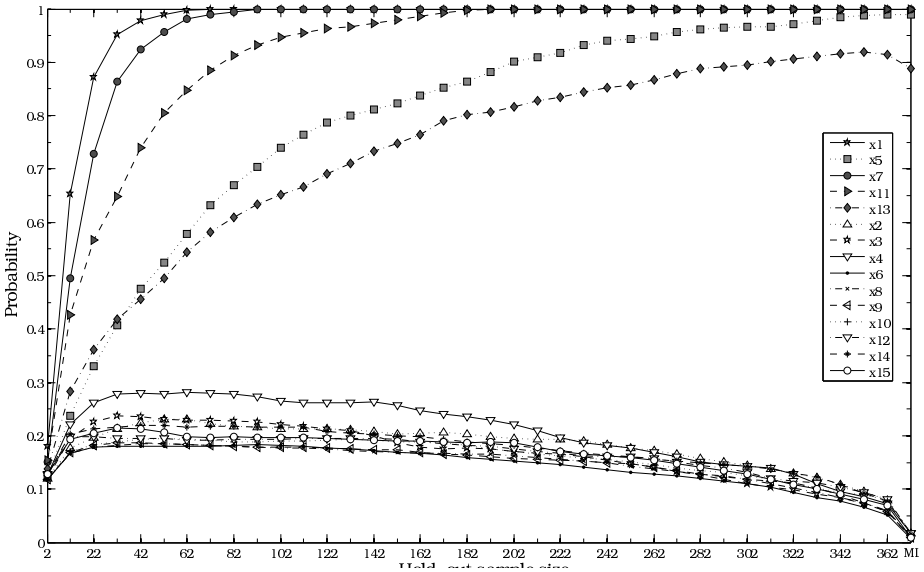
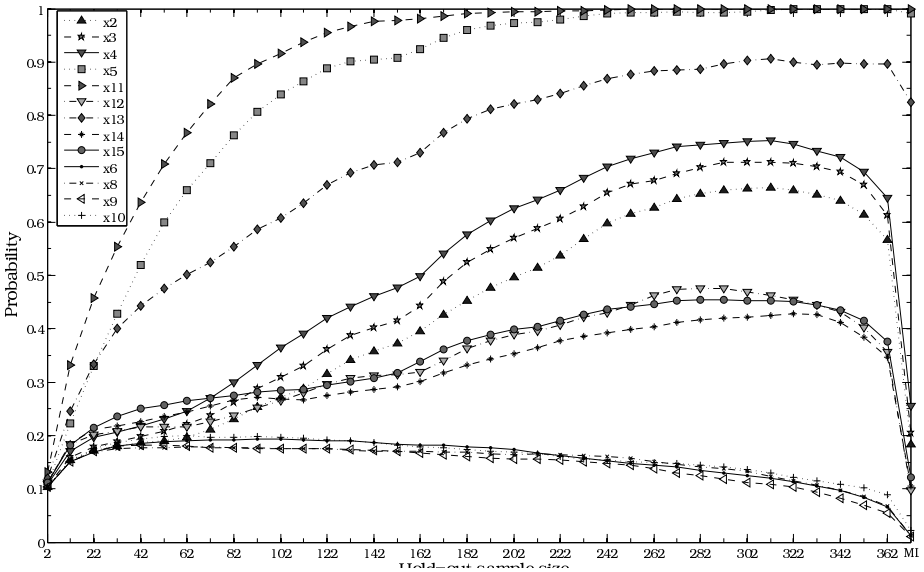


Figure 2.6 Variable inclusion probabilities (average) for large dataset.

(a) $\mathfrak{M}-$closed view



(b) $\mathfrak{M}-$open view



where *1* $(x_i \in \mathcal{M}_j)$ equals one if x_i is included in model j and zero otherwise. When the combinations are formed using predictive weights we substitute $w(\mathcal{M}_j|\widetilde{\mathbf{y}}, \mathbf{y}^*)$ for $p(\mathcal{M}_j|\mathbf{y})$ in (2.24). With a slight abuse of terminology we refer to both as posterior variable probabilities.

In the case when the true model is not in the model set, the predictive likelihood by and large finds the approximation given by the conditional expectation,

$$\begin{aligned} y_t|\,\mathbf{x}_{\setminus 1,7} = &-1.034x_{2,t} - 1.448x_{3,t} - 1.862x_{4,t} - 3.276x_{5,t} + 1.414x_{11,t} \\ &+ 0.414x_{12,t} + 0.914x_{13,t} + 0.414x_{14,t} + 0.414x_{15,t}. \end{aligned} \tag{2.25}$$

In contrast, the marginal likelihood in general only selects from the variables originally in the model.[5] Note that the standard deviation of the prediction error from the conditional model (2.25) is 3.355 compared to 2.5 for the true model.

In the experiments the set of models visited by the chain accounted for $95\% - 98\%$ of the posterior mass for the different forecast observations. The Markov chain visits many more models when using the predictive likelihood, indicating that the model probabilities are much less concentrated than with the marginal likelihood.

2.5.2 Results for $\mathfrak{M}-$open view with shifting parameters

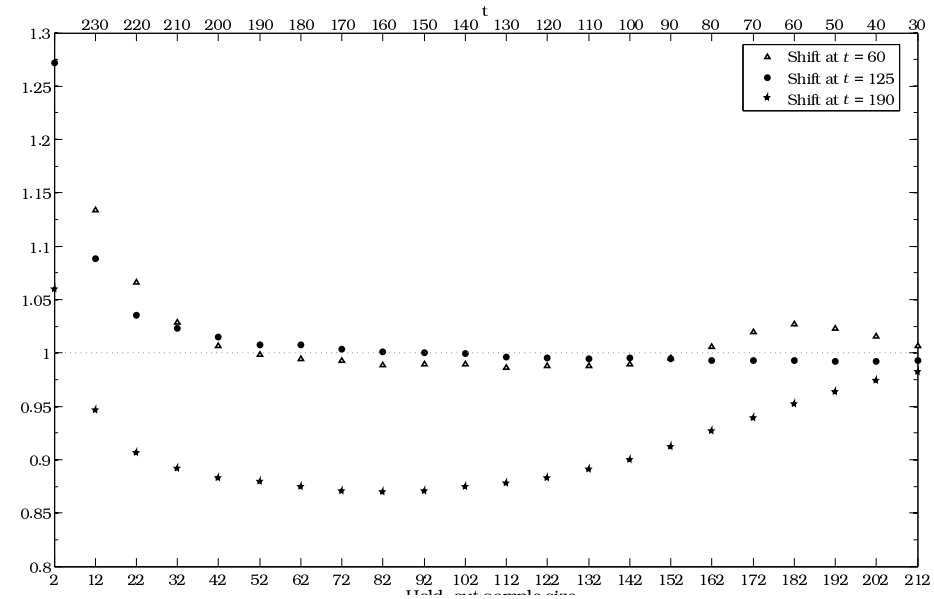
In these experiments, executed only for the medium dataset, we let the size of the hold-out sample vary from $l = 2$ to $l = 212$ with increments of 10. The RMSFEs are reported in Tables C.6 - C.8 and the ratio of the RMSFE for the predictive likelihood to the marginal likelihood RMSFE is graphically represented in Figure 2.7(a). The variable inclusion probabilities for the shifting variable x_7 are plotted in Figure 2.7(b).

The behavior of the variable inclusion probability of x_7 depends on whether the break is in the training sample or in the hold-out sample, and on its position in the sample. In general, when the break is close to the split between the training and hold-out samples, the variable inclusion probabilities for x_7 are at their minimum. When the shift is in the training sample and at the beginning of the data ($t = 60$, $l < 182$) the posterior for the parameters is not heavily influenced by the presence of the break and the hold-out sample fit of the model agrees with the results from the training sample. (Both training and hold-out samples indicate a negative value for the parameter associated with

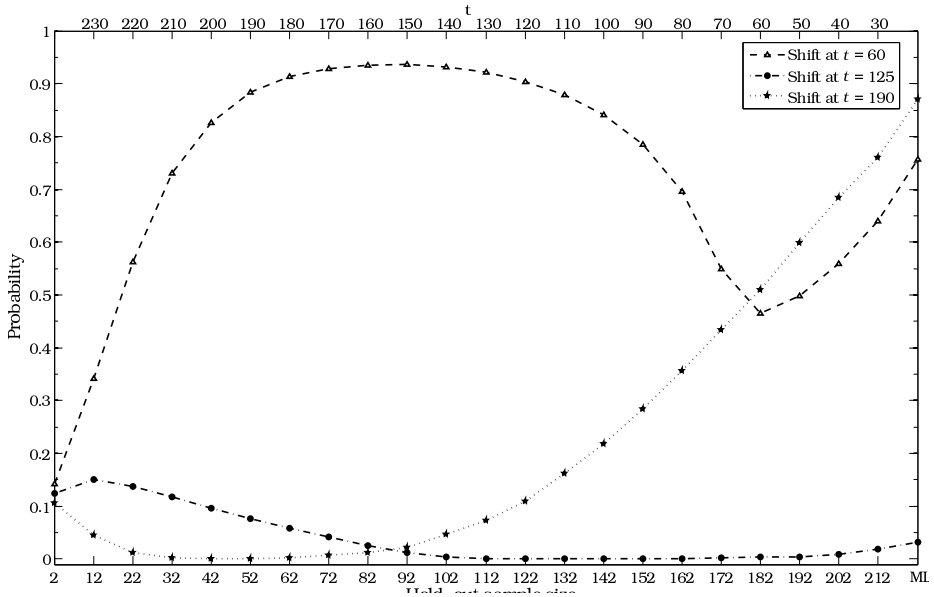
[5] This might be an example of Lindley's paradox. With flat priors the Bayes factor selects asymptotically the smallest model in any nested comparison.

Figure 2.7 Results for the medium dataset with a shifting parameter.

(a) Ratio of RMSFE for predictive likelihood and marginal likelihood as a function of l



(b) Variable inclusion probabilities for x_7, predictive and marginal likelihoods



x_7.) When the break is at the end of the training sample ($t = 190,\ l < 52$) the posterior for the parameters is again relatively unaffected by the break, but the out-of-sample forecasts performance over the hold-out sample is poor as a result of the sign change. Finally, when the break is in the middle of the sample, none of the models performs well with a high model uncertainty as consequence.

When the shift is located in the hold-out sample all models, and in particular models containing x_7, will have problems with prediction after the break. However, as the size of the hold-out sample increases the problem diminishes since the number of pre-break observations grows and it is natural that the inclusion probability for x_7 increases.

The actual forecasts are calculated using posterior distributions for the parameters that are based on the full sample up to the date of the forecast. That is, the forecasts from a given model are the same for the marginal and predictive likelihoods and the difference in forecasting performance is due to the different weights assigned to the models. The difference in performance between the marginal and predictive likelihood is thus due to the ability of the predictive likelihood weights to adapt to structural changes. It does not offer complete protection as the posterior distribution of the parameters is still based on the full sample, but this can be viewed as striking a balance between quick adaption to structural change and efficient use of the data in the no change case.

Roughly speaking, the forecasting problem can be divided into a relatively easy case when the break occurs at the beginning of the data, and a more challenging problem when the break occurs at the end of the data. In the first case, the performance of the marginal and predictive likelihoods are similar and close to what we observe for the no-break case (with a break at $t = 60$ the best RMSFE for the predictive likelihood is 2.69 compared to 2.47 for the no-break, $\mathfrak{M}-$closed case). In the second, more challenging case, with a break at $t = 190$ the smallest RMSFE for the predictive likelihood is 3.17, but this is still a substantial improvement on the 3.65 RMSFE for the marginal likelihood. The pattern of the relative performance depends on the location of the break. For the base, no-break, $\mathfrak{M}-$closed case the predictive likelihood improves significantly on the marginal likelihood for $l \geq 162$. In contrast, when the break occurs at $t = 60$ we have a significantly smaller RMSFE for $82 \leq l \leq 142$ and significantly larger RMSFE for $172 \leq l \leq 202$. In the intermediate case with a break at $t = 125$ the predictive likelihood gives a smaller RMSFE for $l \geq 162$. Finally, for the break at $t = 190$ the predictive likelihood improves significantly on the marginal likelihood except for the smallest, $l = 2$, hold-out

sample. Overall, the largest improvements occur for relatively small hold-out samples, about 40% of the data. This runs counter to the no-break case when the largest improvement occurred with roughly 70% of the data left for the hold-out sample.

2.6 Forecasting the Swedish inflation rate

Our primary goal is forecasting and evaluating forecast performance and we do not attempt to develop models for the inflation rate with causal interpretations. We concentrate on simple regression model of the form

$$y_{t+h} = \alpha + \omega d_{t+h} + \mathbf{x}_t\boldsymbol{\beta} + \varepsilon_{t+h}, \tag{2.26}$$

with the aim to forecast h-periods ahead. The constant term α and a dummy variable, d_t, capturing the low inflation regime assumed to start in 1992Q1, are always included in the model[6], whereas the members of $\mathbf{x}_t$ are selected from the set of potential predictors. While this might seem an overly simplistic and static model formulation at first, there is nothing preventing us from including lags of variables in $\mathbf{x}_t$. The model can thus allow for quite complicated dynamics in the inflation rate. Another feature of the model class is the use of the h-period lead, y_{t+h}, instead of y_t as the dependent variable. This choice of dependent variable has the great advantage that it abolishes the need of forecasting the predictors in $\mathbf{x}_t$ when forecasting y_{t+h}. The obvious alternative is an autoregressive distributed lag specification or a VAR model. See Chevillon and Hendry (2005) and Marcellino, Stock, and Watson (forthcoming) for an in-depth discussion of the relative merits of direct forecasts models like (2.26) and the more traditional dynamic model with forecasts based on the chain rule of forecasting. With dynamic models of this type the predictive density will in general not be available in closed form beyond lead times of $h = 1$. This implies that the predictive distribution must be simulated as part of the MCMC exercise. In practice, this limits the predictive likelihood approach to cases where the number of possible models is relatively small. This would typically be the case with VAR models.

[6]The inclusion of the dummy variable creates a technical difficulty in that this leads to a singular $\mathbf{Z}^{*\prime}\mathbf{Z}^{*}$ matrix for the training sample with improper priors and posteriors as result. We solve this by demeaning both the explanatory and dependent variables separately for the periods before and after 1992Q1. This removes α and ω from the model, which is then estimated without an intercept. This corresponds to using an improper uniform prior on α and ω.

In essence we view (2.26) as the reduced form of a joint model for y_t and $\mathbf{x}_t$. The obvious disadvantage of this choice of dependent variable is that it leads to a different model for each forecast horizon.

The simplicity of the model class allows us to consider a wide range of explanatory variables and possible forecasting models. For the application at hand we have quarterly data for the period 1983Q1 to 2003Q4 on the $N = 77$ predictor variables listed in Appendix A. This set of variables includes a wide range of indicators of real and monetary aspects of the Swedish economy and is close to an exhaustive set of potential predictors for the inflation rate. Note that we include (the current level of) inflation in the set of predictor variables for inflation h-periods ahead. Inflation is measured as the 4-quarter percentage change in the consumer price index and the remaining variables are with few exceptions 4-quarter growth rates or 4-quarter log differences.

We evaluate the performance of the predictive likelihood by producing 4-quarter ahead forecasts for the period 1999Q1 to 2003Q4.

We use the model prior (2.23) with the maximum number of variables in a single model set to $k' = 15$ and $\delta = 0.1$, corresponding to a prior expected model size of 7.7. If substantive prior knowledge about the predictive strength of the variables is available, this can be utilized in the specification of the model prior. Let π_j be the prior probability that variable j is a "good" predictor and included in the correct model. A model prior can then be constructed as

$$p\left(\mathcal{M}_i\right) = \prod_{j=1}^{N} \pi_j^{\tau_{ij}} \left(1 - \pi_j\right)^{1-\tau_{ij}}, \tag{2.27}$$

where $\tau_{ij} = 1$ if x_j is included in $\mathcal{M}_i$ and zero otherwise.[7] We refrain from using domain knowledge when specifying the model priors since our primary aim is to evaluate the relative performance of forecast combinations based on the marginal and predictive likelihoods, not the quality of the prior information.

For the regression parameters of each model we use the g-type prior (2.13) with $c = (k')^3$combined with a Jeffreys prior on the error variance. For each of the point forecasts, we run a preliminary variable selection RJMCMC run with all predictors included in the dataset. After this run we add one lag to the 20 predictors with the highest posterior probabilities of being included in the model and run a final RJMCMC run, which selects models from the new set of 40 variables, keeping the same prior hyper parameters. The prior expected model size in the second run is then 4. See Jacobson and Karlsson (2004)

[7]The model probabilities should be adjusted accordingly if, as in our application with $k' < N$, some models are excluded a priori.

Table 2.1 RMSFE of the Swedish inflation 4-quarter ahead forecast, for $l = 44$.

	Predictive likelihood	*Marginal likelihood*
Forecast combination	0.9429	1.5177
Top 1	1.0323	1.5376
Top 2	0.9036	1.7574
Top 3	0.9523	1.6438
Top 4	1.0336	1.4828
Top 5	0.9870	2.0382
Top 6	0.9661	1.6441
Top 7	1.0534	1.5755
Top 8	1.1758	1.2905
Top 9	1.0983	1.8356
Top 10	1.0999	1.7202
Random walk	1.0251	1.0251

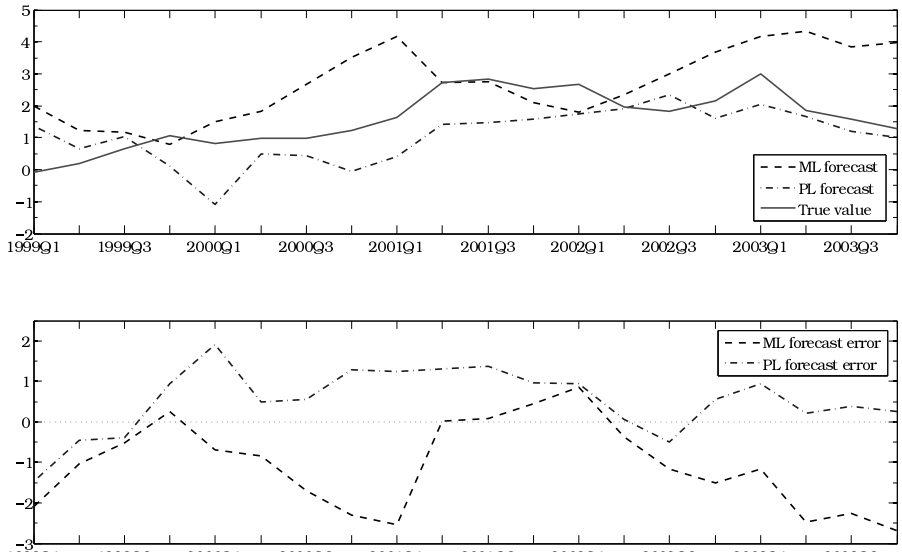
for further details on the variable selection procedure. The chain is run for 5 000 000 replicates in each run.

2.6.1 Results

As the dataset is rather short, starting in 1983Q1, we are restricted in our choice of hold-out sample size. The maximum size of l is given by $T-k'-h-1$, since all the model parameters need to be identified. For the first set of forecast in 1999Q1 we only have 64 observations available and with $k' = 15$ the largest possible hold-out sample size is $l = 44$. This is a little bit short of the 70% found in the simulation study, but might offer good protection against structural breaks at the end of the data. The results for the predictive likelihood (with $l = 44$) and the marginal likelihood are presented in Table 2.1. The table also includes forecasts from the 10 models with the largest weights and the forecast assuming the process is a random walk, i.e. the forecast for y_{t+h} is y_t. The top panel of Figure 2.8 plots the actual values of the inflation, including the forecasts based on the predictive likelihood and the marginal likelihood. In the lower panel the errors from both methods are depicted.

For the inflation forecasts the gain from using the predictive likelihood is quite substantial. For the forecast combination the RMSFE is reduced by 37% compared to the marginal likelihood. In addition all ten models with

Figure 2.8 Swedish inflation rate 4-quarter ahead forecasts and forecast errors, $l = 44$.



the largest weight in the combination outperform the top 10 models for the marginal likelihood. The gain from forecast combination is clear with the predictive likelihood where the combined forecast does better than selecting a single model by the predictive likelihood criterion.

For this dataset the Markov chain accounted for about $88\% - 96\%$ of the posterior mass when the predictive likelihood is used and about $96\% - 99\%$ for the marginal likelihood. The Markov chain visits approximately 4.5 times more models, when using the predictive likelihood, than when using the marginal likelihood, suggesting that the predictive likelihood does not discriminate between models to the same extent as the marginal likelihood. This is confirmed by Table 2.2, which gives the average of the variable inclusion probabilities over the 20 forecasts. The marginal likelihood clearly favors three variables, the population share in two age groups and housing prices, including them in essentially all models. The inclusion probabilities are much more dispersed for the predictive likelihood, with current inflation having probability 1/2 of being included. One factor contributing to this difference is that the marginal likelihood consistently picks the same three variables for all the forecasts, whereas the predictive likelihood favors different sets of variables for different time periods.

Table 2.2 Variables with highest posterior inclusion probabilities (average).

	Predictive likelihood		*Marginal likelihood*	
	Variable	Post. prob.	Variable	Post. prob.
1.	`Infla`	0.5528	`Pp1664`	0.9994
2.	`InfRel`	0.4493	`Pp1529`	0.9896
3.	`U314W`	0.3271	`InfHWg`	0.9456
4.	`REPO`	0.2871	`AFGX`	0.8104
5.	`IndProd`	0.2459	`PpTot`	0.4996
6.	`ExpInf`	0.2392	`PrvEmp`	0.4804
7.	`R5Y`	0.1947	`InfCns`	0.4513
8.	`InfFl`	0.1749	`InfPrd`	0.4105
9.	`M0`	0.1533	`R3M`	0.4048
10.	`InfUnd`	0.1473	`Pp75+`	0.3927
11.	`LabFrc`	0.1409	`ExpInf`	0.3829
12.	`NewHouse`	0.1245	`InfFor`	0.3786
13.	`InfImpP`	0.1225	`M0`	0.1793
14.	`PrvEmp`	0.1219	`POilSEK`	0.1702
15.	`PPP`	0.1134	`USD`	0.1170

As an example consider Table 2.3, which reports the models with the largest weights for the 1999Q1 forecast. While not always the case, the predictive likelihood favors smaller models for this forecast. Note that the marginal likelihood clearly favors one model with a posterior probability of 0.13, thrice that of the second best model, while the predictive likelihood indicates much more model uncertainty with a predictive weight of 0.05 for the best model. Effectively, the predictive likelihood will thus include more models in the forecast combination and provide greater robustness against in-sample overfitting.

The variables selected by the marginal and predictive likelihoods can in general be expected to have predictive content for inflation. The population variables, which at first sight might be more difficult to motivate, are a possible exception. Lindh and Malmberg (1998) argue from a Wicksellian perspective that the demographic age structure influences investment and savings decisions and should consequently be one of the determinants of the inflation process. It is never the less interesting to note that the population variables are ranked much lower by the predictive likelihood. In addition, the variables related to current inflation and real activity are ranked higher by the predictive likelihood.

Table 2.3 Model weights and posterior probabilities, 4-quarter ahead Swedish inflation forecast for 1999Q1.

(a) Predictive likelihood, $l = 44$

	Model				
Variable	1	2	3	4	5
InfRel	×	×	×		×
$\texttt{InfRel}_{-1}$				×	×
ExpInf	×	×	×	×	×
R5Y	×	×	×		×
InfFl	×	×		×	×
$\texttt{InfFl}_{-1}$			×		
InfUnd	×	×	×	×	×
USD	×	×	×		×
GDPTCW		×			×
$\texttt{GDPTCW}_{-1}$				×	
Pred. weights	0.0538	0.0301	0.0218	0.0187	0.0184

(b) Marginal likelihood

	Model				
Variable	1	2	3	4	5
Pp1664	×	×	×	×	×
Pp1529	×	×	×	×	×
InfHWg	×	×	×	×	×
$\texttt{AFGX}_{-1}$	×	×	×		
PpTot	×	×	×		×
$\texttt{PpTot}_{-1}$				×	
$\texttt{R3M}_{-1}$	×	×	×	×	×
InfFor	×				
$\texttt{InfFor}_{-1}$		×			
POilSEK			×		
$\texttt{NewJob}_{-1}$	×	×			
PP2534	×	×			
Post. prob	0.1316	0.0405	0.0347	0.0264	0.0259

2.7 Conclusions

This chapter proposes the use of the out-of-sample predictive likelihood in Bayesian forecast combination. We show that the forecast weights based on the predictive likelihood have desirable asymptotic properties, i.e. they will consistently select the correct model. Our analysis indicates that the weights based on the predictive likelihood will have better small sample properties than the traditional in-sample marginal likelihood when uninformative priors are used. The improved small sample performance is due to the predictive likelihood considering both in-sample fit and out-of-sample predictive performance, where the latter protects against in-sample overfitting of the data. The analytical results are supported by a simulation study and an application to forecasting the Swedish inflation rate. Forecast combination based on the predictive likelihood outperforms forecast combination based on the marginal likelihood in both cases.

In practice, we can not expect the true model or data generating process to be included in the set of considered models. The simulation experiments indicate that this is also when we can expect the largest gains from the use of the predictive likelihood. When there is a true model, the predictive likelihood will select the true model asymptotically, but will converge slower to the true model than the marginal likelihood. It is this slower convergence coupled with the protection against overfitting provided by explicitly considering out-of-sample predictive ability that drives the better performance of the predictive likelihood when the true model is not in the model set. The superior performance of the predictive likelihood in the $\mathfrak{M}$-open case is also a likely explanation of the results for the Swedish inflation forecasts.

Bibliography

Aitkin, M. (1991): "Posterior Bayes Factors," *Journal of the Royal Statistical Society B*, 53(1), 111–142.

Bates, J., and C. Granger (1969): "The Combination of Forecasts," *Operational Research Quarterly*, 20, 451–468.

Bauwens, L., M. Lubrano, and J.-F. Richard (1999): *Bayesian Inference in Dynamic Econometric Models*. Oxford University Press, New York.

Berger, J. O., and L. R. Pericchi (1996): "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91(433), 109–122.

Bernardo, J. M., and A. F. Smith (1994): *Bayesian Theory*. John Wiley & Sons, Chichester.

Chevillon, G., and D. F. Hendry (2005): "Non-Parametric Direct Multi-Step Estimation for Forecasting Economic Processes," *International Journal of Forecasting*, 21(2), 201–218.

Clemen, R. T. (1989): "Combining Forecasts: A Review and an Annotated Bibliography," *International Journal of Forecasting*, 5(4), 559–583.

Elliott, G., C. W. J. Granger, and A. Timmermann (eds.) (2006): *Handbook of Economic Forecasting*, vol. 1. Elsevier.

Elliott, G., and A. Timmermann (2004): "Optimal Forecast Combinations Under General Loss Functions and Forecast Error Distributions," *Journal of Econometrics*, 122(1), 47–79.

Fernández, C., E. Ley, and M. F. Steel (2001): "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100(2), 381–427.

Fernández-Villaverde, J., and J. F. Rubio-Ramírez (2004): "Comparing Dynamic Equilibrium Models to Data: A Bayesian Approach," *Journal of Econometrics*, 123(1), 153–187.

Geisser, S. (1975): "The Predictive Sample Reuse Method with Applications," *Journal of the American Statistical Association*, 70(350), 320–328.

Gelfand, A. E., and D. K. Dey (1994): "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society B*, 56(3), 501–514.

George, E. I., and R. E. McCulloch (1997): "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Green, P. J. (1995): "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82(4), 711–732.

Hendry, D. F., and M. P. Clements (2004): "Pooling of Forecasts," *Econometrics Journal*, 7(1), 1–31.

Jacobson, T., and S. Karlsson (2004): "Finding Good Predictors for Inflation: A Bayesian Model Averaging Approach," *Journal of Forecasting*, 23(7), 479–496.

Koop, G., and S. Potter (2004): "Forecasting in Dynamic Factor Models Using Bayesian Model Averaging," *Econometrics Journal*, 7(2), 550–565.

Laud, P. W., and J. G. Ibrahim (1995): "Predictive Model Selection," *Journal of the Royal Statistical Society B*, 57(1), 247–262.

Lindh, T., and B. Malmberg (1998): "Age Structure and Inflation - a Wicksellian Interpretation of the OECD Data," *Journal of Economic Behavior & Organization*, 36, 19–37.

Madigan, D., and A. E. Raftery (1994): "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89(428), 1535–1546.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler (1982): "The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition," *Journal of Forecasting*, 1(2), 111–153.

Makridakis, S., C. Chatfield, M. Hibon, M. Lawrence, T. Mills, K. Ord, and L. F. Simmons (1993): "The M2-Competition: A Real-Time Judgmentally-Based Forecasting Study," *International Journal of Forecasting*, 9(1), 5–23.

Makridakis, S., and M. Hibon (2000): "The M3-Competition: Results, Conclusions and Implications," *International Journal of Forecasting*, 16(4), 451–476.

Marcellino, M., J. H. Stock, and M. W. Watson (forthcoming): "A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series," *Journal of Econometrics*.

Min, C.-K., and A. Zellner (1993): "Bayesian and Non-Bayesian Methods for Combining Models and Forecasts with Applications to Forecasting International Growth Rates," *Journal of Econometrics*, 56(1-2), 89–118.

O'Hagan, A. (1991): "Discussion on Posterior Bayes Factors (by M. Aitkin)," *Journal of the Royal Statistical Society B*, 53(1), 136.

——— (1995): "Fractional Bayes Factors for Model Comparison," *Journal of the Royal Statistical Society B*, 57(1), 99–138.

Peña, D., and G. C. Tiao (1992): "Bayesian Robustness Functions for Linear Models," in *Bayesian Statistics 4*, ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, pp. 147–167. Oxford University Press, Oxford.

Stock, J. H., and M. W. Watson (1999): "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series," in *Cointegration, Causality, and Forecasting A Festschrift in Honour of Clive W.J. Granger*, ed. by R. F. Engle, and H. White, pp. 1–44. Oxford University Press, Oxford.

——— (2002): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics*, 20(2), 147 – 162.

——— (2006): "Forecasting with Many Predictors," in Elliott, Granger, and Timmermann (2006), chap. 10.

Stone, M. (1974): "Cross-Validatory Choice and Assessment of Statistical Predictions (with Discussion)," *Journal of the Royal Statistical Society B*, 36(2), 111–147.

Timmermann, A. (2006): "Forecast Combinations," in Elliott, Granger, and Timmermann (2006), chap. 4.

Appendix A

Marginal and predictive likelihoods

A.1 Marginal likelihood

Consider a linear regression model for $\mathbf{y}^* = (y_1, y_2, \ldots, y_m)'$

$$\mathbf{y}^* \sim \mathrm{N}_m \left(\mathbf{Z}^* \boldsymbol{\gamma}^*, \sigma^2 \mathbf{I}_m\right), \tag{A.1}$$
$$\boldsymbol{\gamma}^* \sim \left(\alpha, \boldsymbol{\beta}'\right)', \quad (k+1) \times 1, \tag{A.2}$$
$$\mathbf{Z}^* = (\boldsymbol{\iota}, \mathbf{X}^*), \tag{A.3}$$

with following priors for the parameters

$$p\left(\boldsymbol{\gamma}^* | \sigma^2\right) \sim \mathrm{N}_{k+1} \left(0, c\sigma^2 \left(\mathbf{Z}^{*\prime} \mathbf{Z}^*\right)^{-1}\right), \tag{A.4}$$
$$p\left(\sigma^2\right) \propto \frac{1}{\sigma^2}. \tag{A.5}$$

This yields the Normal-Inverted Gamma-2 posterior density

$$\boldsymbol{\gamma}^* | \mathbf{y}^*, \sigma^2 \sim \mathrm{N}_{k+1} \left(\boldsymbol{\gamma}_1, \sigma^2 \left(\mathbf{M}^*\right)^{-1}\right), \tag{A.6}$$
$$\sigma^2 | \mathbf{y}^* \sim \mathrm{IG}_2 \left(S^*, m\right), \tag{A.7}$$
$$\mathbf{M}^* = \frac{c+1}{c} \mathbf{Z}^{*\prime} \mathbf{Z}^*, \tag{A.8}$$
$$\boldsymbol{\gamma}_1 = \frac{c}{c+1} \hat{\boldsymbol{\gamma}}^*, \tag{A.9}$$
$$S^* = \frac{c}{c+1} \left(\mathbf{y}^* - \mathbf{Z}^* \hat{\boldsymbol{\gamma}}^*\right)' \left(\mathbf{y}^* - \mathbf{Z}^* \hat{\boldsymbol{\gamma}}^*\right) + \frac{1}{c+1} \mathbf{y}^{*\prime} \mathbf{y}^*, \tag{A.10}$$

The marginal likelihood is then

$$m(\mathbf{y}) \propto \frac{\left|\frac{1}{c}\mathbf{Z}^{*\prime}\mathbf{Z}^{*}\right|^{\frac{1}{2}}}{\left|\frac{c+1}{c}\mathbf{Z}^{*\prime}\mathbf{Z}^{*}\right|^{\frac{1}{2}}}\left(S^{*}\right)^{-m/2} = (c+1)^{(-k+1)/2}\left(S^{*}\right)^{-m/2}. \tag{A.11}$$

A.2 Predictive likelihood

The predictive density of $\widetilde{\mathbf{y}} = (y_{m+1}, y_{m+2}, \ldots, y_T)'$ is

$$\widetilde{\mathbf{y}}|\,\widetilde{\mathbf{Z}}, \mathbf{Z}^{*}, \mathbf{y}^{*}, \boldsymbol{\gamma}^{*}, \sigma^{2} \sim \mathrm{N}_{l}\left(\widetilde{\mathbf{Z}}\boldsymbol{\gamma}^{*}, \sigma^{2}\mathbf{I}_{l}\right), \tag{A.12}$$

where $\widetilde{\mathbf{Z}}$ is a $(l \times (k+1))$ matrix of observations of the future exogenous variables. The joint density of $\boldsymbol{\gamma}^{*}$ and $\widetilde{\mathbf{y}}$ conditionally on $\sigma^{2}, \widetilde{\mathbf{Z}}, \mathbf{Z}^{*}, \mathbf{y}^{*}$ is Normal

$$\left(\begin{array}{c}\boldsymbol{\gamma}^{*}\\ \widetilde{\mathbf{y}}\end{array}\right)\Bigg|\,\widetilde{\mathbf{Z}}, \mathbf{Z}^{*}, \mathbf{y}^{*}, \sigma^{2} \sim \mathrm{N}_{k+1+l}\left(\left(\begin{array}{c}\boldsymbol{\gamma}_{1}\\ \widetilde{\mathbf{Z}}\boldsymbol{\gamma}_{1}\end{array}\right), \sigma^{2}\left[\begin{array}{cc}\left(\mathbf{M}^{*}\right)^{-1} & \left(\mathbf{M}^{*}\right)^{-1}\widetilde{\mathbf{Z}}'\\ \widetilde{\mathbf{Z}}\left(\mathbf{M}^{*}\right)^{-1} & \mathbf{I}_{l} + \widetilde{\mathbf{Z}}\left(\mathbf{M}^{*}\right)^{-1}\widetilde{\mathbf{Z}}'\end{array}\right]\right). \tag{A.13}$$

See Bauwens, Lubrano, and Richard (1999) for further details.

As $\sigma^{2}\big|\,\mathbf{y}^{*}, \mathbf{Z}^{*} \sim \mathrm{IG}_{2}\left(S^{*}, m\right)$ it follows that the predictive density of $\widetilde{\mathbf{y}}$ is multivariate Student and defined by

$$\widetilde{\mathbf{y}}|\,\widetilde{\mathbf{Z}}, \mathbf{Z}^{*}, \mathbf{y}^{*} \sim \mathrm{t}_{l}\left(\widetilde{\mathbf{Z}}\boldsymbol{\gamma}_{1}, S^{*}, \left(\mathbf{I}_{l} + \widetilde{\mathbf{Z}}\left(\mathbf{M}^{*}\right)^{-1}\widetilde{\mathbf{Z}}'\right)^{-1}, m\right) \tag{A.14}$$

with the density function

$$\begin{aligned} p\left(\widetilde{\mathbf{y}}|\,\widetilde{\mathbf{Z}}, \mathbf{Z}^{*}, \mathbf{y}^{*}\right) = {} & \frac{\Gamma\left(\frac{m+l}{2}\right)\left(S^{*}\right)^{m/2}}{\pi^{l/2}\Gamma\left(\frac{m}{2}\right)\left|\mathbf{I} + \widetilde{\mathbf{Z}}\left(\mathbf{M}^{*}\right)^{-1}\widetilde{\mathbf{Z}}'\right|^{1/2}} \\ & \times\left[S^{*} + \left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}_{1}\right)'\left(\mathbf{I} + \widetilde{\mathbf{Z}}\left(\mathbf{M}^{*}\right)^{-1}\widetilde{\mathbf{Z}}'\right)^{-1}\left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}_{1}\right)\right]^{-T/2} \end{aligned} \tag{A.15}$$

Appendix B

MCMC algorithms

B.1 Predictive likelihood

Algorithm B.1 Reversible jump Markov chain Monte Carlo

Suppose that the Markov chain is at model $\mathcal{M}$, having parameters $\boldsymbol{\theta}_{\mathcal{M}}$, where $\boldsymbol{\theta}_{\mathcal{M}}$ has dimension $\dim(\boldsymbol{\theta}_{\mathcal{M}})$.

1. Propose a jump from model $\mathcal{M}$ to a new model $\mathcal{M}'$ with probability $j(\mathcal{M}'|\mathcal{M})$.

2. Generate a vector $\mathbf{u}$ (which can have a different dimension than $\boldsymbol{\theta}_{\mathcal{M}'}$) from a specified proposal density $q(\mathbf{u}|\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}, \mathcal{M}')$.

3. Set $(\boldsymbol{\theta}_{\mathcal{M}'}, \mathbf{u}') = g_{\mathcal{M},\mathcal{M}'}(\boldsymbol{\theta}_{\mathcal{M}}, \mathbf{u})$, where $g_{\mathcal{M},\mathcal{M}'}$ is a specified invertible function. Hence $\dim(\boldsymbol{\theta}_{\mathcal{M}}) + \dim(\mathbf{u}) = \dim(\boldsymbol{\theta}_{\mathcal{M}\prime}) + \dim(\mathbf{u}')$. Note that $g_{\mathcal{M},\mathcal{M}'} = g^{-1}_{\mathcal{M}',\mathcal{M}}$.

4. Accept the proposed move with probability

$$\alpha = \min\left\{1, \frac{L(\widetilde{\mathbf{y}}|\mathbf{y}^*, \boldsymbol{\theta}_{\mathcal{M}'}, \mathcal{M}')\, p(\boldsymbol{\theta}_{\mathcal{M}'}|\mathbf{y}^*, \mathcal{M}')\, p(\mathcal{M}')\, j(\mathcal{M}|\mathcal{M}')}{L(\widetilde{\mathbf{y}}|\mathbf{y}^*, \boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M})\, p(\boldsymbol{\theta}_{\mathcal{M}}|\mathbf{y}^*, \mathcal{M})\, p(\mathcal{M})\, j(\mathcal{M}'|\mathcal{M})} \times \frac{q(\mathbf{u}'|\boldsymbol{\theta}_{\mathcal{M}'}, \mathcal{M}', \mathcal{M})}{q(\mathbf{u}|\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}, \mathcal{M}')} \left|\frac{\partial g_{\mathcal{M},\mathcal{M}'}(\boldsymbol{\theta}_{\mathcal{M}}, \mathbf{u})}{\partial(\boldsymbol{\theta}_{\mathcal{M}}, \mathbf{u})}\right|\right\}. \quad \text{(B.1)}$$

5. Set $\mathcal{M} = \mathcal{M}'$ if the move is accepted.

The target distribution of the chain is proportional to

$$L\left(\widetilde{\mathbf{y}}|\mathbf{y}^*,\boldsymbol{\theta}_{\mathcal{M}},\mathcal{M}\right)p\left(\boldsymbol{\theta}_{\mathcal{M}}|\mathbf{y}^*,\mathcal{M}\right)p\left(\mathcal{M}\right)=p\left(\widetilde{\mathbf{y}},\boldsymbol{\theta}_{\mathcal{M}}|\mathbf{y}^*,\mathcal{M}\right)p\left(\mathcal{M}\right). \tag{B.2}$$

Discarding the draws of $\boldsymbol{\theta}_{\mathcal{M}}$, yields draws from the marginal distribution of $\mathcal{M}$, which is proportional to $p\left(\widetilde{\mathbf{y}}|\mathbf{y}^*,\mathcal{M}\right)p\left(\mathcal{M}\right)$, the numerator of (2.8).

If all parameters of the proposed model are generated directly from a proposal distribution, then $\left(\boldsymbol{\theta}_{\mathcal{M}'},\mathbf{u}'\right)=\left(\mathbf{u},\boldsymbol{\theta}_{\mathcal{M}}\right)$ with $\dim\left(\boldsymbol{\theta}_{\mathcal{M}}\right)=\dim\left(\mathbf{u}'\right)$ and $\dim\left(\boldsymbol{\theta}_{\mathcal{M}'}\right)=\dim\left(\mathbf{u}\right)$, and the Jacobian is unity. If, in addition, the proposal $q\left(\mathbf{u}|\boldsymbol{\theta}_{\mathcal{M}},\mathcal{M},\mathcal{M}'\right)$ is the posterior $p\left(\boldsymbol{\theta}_{\mathcal{M}'}|\widetilde{\mathbf{y}},\mathbf{y}^*,\mathcal{M}'\right)$ then (B.1) simplifies to

$$\alpha=\min\left\{1,\frac{p\left(\widetilde{\mathbf{y}}|\mathbf{y}^*,\mathcal{M}'\right)p\left(\mathcal{M}'\right)j\left(\mathcal{M}|\mathcal{M}'\right)}{p\left(\widetilde{\mathbf{y}}|\mathbf{y}^*,\mathcal{M}\right)p\left(\mathcal{M}\right)j\left(\mathcal{M}'|\mathcal{M}\right)}\right\} \tag{B.3}$$

since

$$p\left(\widetilde{\mathbf{y}}|\mathbf{y}^*,\mathcal{M}'\right)=\frac{L\left(\widetilde{\mathbf{y}}|\mathbf{y}^*,\boldsymbol{\theta}_{\mathcal{M}'},\mathcal{M}'\right)p\left(\boldsymbol{\theta}_{\mathcal{M}'}|\mathbf{y}^*,\mathcal{M}'\right)}{p\left(\boldsymbol{\theta}_{\mathcal{M}'}|\widetilde{\mathbf{y}},\mathbf{y}^*,\mathcal{M}'\right)}. \tag{B.4}$$

Note that this implies that we do not need to sample the parameters since the acceptance probability depends only on the predictive likelihood. This is the form of the algorithm we use where steps 2 and 3 are omitted. Two types of model changing moves are considered:

1. Draw a variable at random and drop it if it is in the model or add it to the model (if $k_{\mathcal{M}}<k'$). This step is attempted with probability p_A.
2. Swap a randomly selected variable in the model for a randomly selected variable outside the model (if $k_{\mathcal{M}}>0$). This step is attempted with probability $1-p_A$.

Note that for these two moves $j\left(\mathcal{M}|\mathcal{M}'\right)=j\left(\mathcal{M}'|\mathcal{M}\right)$ and the acceptance ratio simplifies further.

B.2 Marginal likelihood

The same basic algorithm is used with the marginal likelihood. The only difference is that we substitute $L\left(\mathbf{y}|\boldsymbol{\theta}_{\mathcal{M}'},\mathcal{M}'\right)p\left(\boldsymbol{\theta}_{\mathcal{M}'}|\mathcal{M}'\right)$ for $L\left(\widetilde{\mathbf{y}}|\mathbf{y}^*,\boldsymbol{\theta}_{\mathcal{M}'},\mathcal{M}'\right)\cdot p\left(\boldsymbol{\theta}_{\mathcal{M}'}|\mathbf{y}^*,\mathcal{M}'\right)$ in (B.1). A similar simplifications of the acceptance ratio is available here by taking the posterior as the proposal distribution for the parameters and the acceptance ratio simplifies to

$$\alpha=\min\left(1,\frac{m\left(\mathbf{y}|\mathcal{M}'\right)p\left(\mathcal{M}'\right)}{m\left(\mathbf{y}|\mathcal{M}\right)p\left(\mathcal{M}\right)}\right)$$

and it is not necessary to sample the parameters.

Appendix C

Simulation results

l is the size of the hold-out sample, ML is the marginal likelihood

Table C.1 RMSFE for simulated small data set, $\mathfrak{M}-$closed view.

l	RMSFE	l	RMSFE	l	RMSFE	l	RMSFE
2	3.5348*	23	2.7428*	44	2.6445*	65	2.6184
5	3.2670*	26	2.7155*	47	2.6429*	68	2.6144
8	3.0705*	29	2.6807*	50	2.6421*	71	2.6106
11	2.9421*	32	2.6631*	53	2.6373	74	2.6060
14	2.8620*	35	2.6557*	56	2.6312	77	2.6048
17	2.8089*	38	2.6442*	59	2.6320	80	2.6041
20	2.7697*	41	2.6406*	62	2.6252	83	2.6008
						ML	2.6100

* significantly different from the ML RMSFE at the 5% level

Table C.2 RMSFE for simulated medium data set, $\mathfrak{M}-$closed view.

l	RMSFE	l	RMSFE	l	RMSFE	l	RMSFE
2	3.6146*	57	2.5305*	112	2.4943	167	2.4790*
7	3.1886*	62	2.5236	117	2.4927	172	2.4761*
12	2.9322*	67	2.5207	122	2.4943	177	2.4743*
17	2.7654*	72	2.5152	127	2.4940	182	2.4764*
22	2.6741*	77	2.5077	132	2.4927	187	2.4774*
27	2.6323*	82	2.5085	137	2.4907	192	2.4788*
32	2.6170*	87	2.5049	142	2.4921	197	2.4796*
37	2.5926*	92	2.4993	147	2.4926	202	2.4815
42	2.5714*	97	2.4950	152	2.4876	207	2.4826
47	2.5510*	102	2.4931	157	2.4848	212	2.4814*
52	2.5394*	107	2.4909	162	2.4791*	ML	2.4945

* significantly different from the ML RMSFE at the 5% level

Table C.3 RMSFE for simulated large data set, $\mathfrak{M}-$closed view.

l	RMSFE	l	RMSFE	l	RMSFE	l	RMSFE
2	3.6407*	102	2.5282*	202	2.5103*	302	2.5008
12	2.9105*	112	2.5258*	212	2.5099*	312	2.5011
22	2.7280*	122	2.5257*	222	2.5101*	322	2.4996
32	2.6526*	132	2.5248*	232	2.5074*	332	2.4997
42	2.6158*	142	2.5213*	242	2.5063	342	2.4996
52	2.5895*	152	2.5229*	252	2.5046	352	2.4986
62	2.5670*	162	2.5211*	262	2.5029	362	2.4990
72	2.5561*	172	2.5174*	272	2.5028		
82	2.5439*	182	2.5132*	282	2.5007	ML	2.4989
92	2.5365*	192	2.5106*	292	2.5012		

* significantly different from the ML RMSFE at the 5% level

Table C.4 RMSFE for simulated medium data set, $\mathfrak{M}-$open view.

l	RMSFE	l	RMSFE	l	RMSFE	l	RMSFE
2	3.9913*	57	3.5990	112	3.5839	167	3.5581*
7	3.8793*	62	3.5956	117	3.5758*	172	3.5563*
12	3.8202*	67	3.5983	122	3.5760*	177	3.5551*
17	3.7602*	72	3.5998	127	3.5763*	182	3.5512*
22	3.7130*	77	3.5910	132	3.5751*	187	3.5558*
27	3.6821*	82	3.5944	137	3.5681*	192	3.5538*
32	3.6603*	87	3.5923	142	3.5656*	197	3.5519*
37	3.6451	92	3.5890	147	3.5651*	202	3.5535*
42	3.6385	97	3.5852	152	3.5640*	207	3.5587*
47	3.6272	102	3.5842	157	3.5634*	212	3.5629*
52	3.6106	107	3.5828	162	3.5588*	ML	3.6107

* significantly different from the ML RMSFE at the 5% level

Table C.5 RMSFE for simulated large data set, $\mathfrak{M}-$open view.

l	RMSFE	l	RMSFE	l	RMSFE	l	RMSFE
2	3.9788*	102	3.4292	202	3.3663*	302	3.3428*
12	3.7416*	112	3.4175	212	3.3610*	312	3.3473*
22	3.6408*	122	3.4130	222	3.3569*	322	3.3436*
32	3.5858*	132	3.4065	232	3.3557*	332	3.3460*
42	3.5559*	142	3.4020	242	3.3544*	342	3.3472*
52	3.5307*	152	3.4052	252	3.3513*	352	3.3466*
62	3.5029*	162	3.4097	262	3.3495*	362	3.3589*
72	3.4741*	172	3.3924	272	3.3491*		
82	3.4541*	182	3.3774*	282	3.3477*	ML	3.4053
92	3.4376*	192	3.3663*	292	3.3449*		

* significantly different from the ML RMSFE at the 5% level

Table C.6 RMSFE for simulated medium data set, break at $t = 60$.

l	RMSFE	l	RMSFE	l	RMSFE	l	RMSFE
2	3.6211*	62	2.7106	122	2.6927*	182	2.8003*
12	3.0914*	72	2.7057	132	2.6918*	192	2.7891*
22	2.9072*	82	2.6956*	142	2.6981*	202	2.7673*
32	2.8034*	92	2.6974*	152	2.7135	212	2.7443
42	2.7436	102	2.6971*	162	2.7417		
52	2.7218	112	2.6891*	172	2.7799*	ML	2.7255

* significantly different from the ML RMSFE at the 5% level

Table C.7 RMSFE for simulated medium data set, break at $t = 125$.

l	RMSFE	l	RMSFE	l	RMSFE	l	RMSFE
2	3.6013*	62	2.8533*	122	2.8186	182	2.8100*
12	3.0812*	72	2.8421	132	2.8167	192	2.8089*
22	2.9318*	82	2.8345	142	2.8170	202	2.8084*
32	2.8972*	92	2.8330	152	2.8146	212	2.8108*
42	2.8741*	102	2.8289	162	2.8112*		
52	2.8523	112	2.8201	172	2.8117*	ML	2.8312

* significantly different from the ML RMSFE at the 5% level

Table C.8 RMSFE for simulated medium data set, break at $t = 190$.

l	RMSFE	l	RMSFE	l	RMSFE	l	RMSFE
2	3.8659*	62	3.1902*	122	3.2197*	182	3.4720*
12	3.4526*	72	3.1737*	132	3.2506*	192	3.5134*
22	3.3066*	82	3.1715*	142	3.2817*	202	3.5520*
32	3.2510*	92	3.1735*	152	3.3276*	212	3.5815*
42	3.2209*	102	3.1907*	162	3.3815*		
52	3.2064*	112	3.2010*	172	3.4259*	ML	3.6470

* significantly different from the ML RMSFE at the 5% level

Chapter 3

Forecasting GDP with Factor Models and Bayesian Forecast Combination

Acknowledgement: I wish to thank Chris Jeffery for his valuable comments.

3.1 Introduction

In recent years there has been increasing interest in forecasting methods that utilize large datasets. Standard econometric techniques are not well suited to exploit the huge quantity of information available within such datasets. There are, essentially, two different methodologies for extracting the relevant information from possible predictors: factor modelling, which summarizes a proportion of the variation in all the data in a limited number of factors, and forecast combination, where the information in many possible models is combined.

Several studies have used dynamic factor models for forecasting when a large quantity of potential regressors is available, e.g. Stock and Watson (2002b), Stock and Watson (2003), Shintani (2003), Artis, Banerjee, and Marcellino (2005) and others. A common conclusion from these studies is that factor-based models outperform forecasts from simple AR processes. The magnitude of the improvements is, however, varying with the dependent variable and the forecasting horizon.

The applications of Bayesian model averaging (BMA) in large macroeconomic panels include Jacobson and Karlsson (2004), Koop and Potter (2004) and Wright(2003a, 2003b). They conclude that forecast combination using BMA improves upon forecasts based on single models. These studies use the traditional marginal likelihood as weights in combining the forecasts. Chapter 2 suggests using an out-of-sample measure when combining forecasts, showing that the forecast weights based on the predictive likelihood have better small sample properties than the traditional marginal likelihood-based weights. Recently, Kapetanios, Labhard, and Price (2006) use an out-of-sample measure of fit in standard information criteria when constructing weights for forecast combination in a frequentist information theoretic approach. They find that the proposed method performs well and, in some respects, outperforms other averaging methods considered.

The goal of this chapter is to apply the idea of forecast combination using predictive likelihood-based weights to forecast GDP. The out-of-sample weights have the ability to adapt to structural changes and this could be superior to existing methods used for GDP growth forecasting. The forecast combination technique based on both in-sample and out-of-sample weights is compared to forecasts based on factor models. Output growth is forecast for six countries, using quarterly datasets, containing between 25 and 43 variables. In addition, confidence intervals of the forecasts are calculated.

This chapter is organised as follows; Section 3.2 describes the Bayesian approach to forecast combination, Section 3.3 introduces the factor model,

and the forecast comparison is presented in Section 3.4. Finally, Section 3.5 concludes.

3.2 Bayesian combination of forecasts

Model averaging reflects the need to account for model uncertainty in carrying out statistical analysis. Consider a set of models $\mathfrak{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$. Denote the prior probability of a model being true as $p(\mathcal{M}_i)$ and the prior distribution of the parameters in each model as $p(\boldsymbol{\theta}_i|\,\mathcal{M}_i)$. The posterior probabilities of the models after observing the data $\mathbf{y}$ follow from Bayes rule

$$p(\mathcal{M}_i|\,\mathbf{y}) = \frac{m(\mathbf{y}|\,\mathcal{M}_i)\,p(\mathcal{M}_i)}{\sum_{j=1}^{M} m(\mathbf{y}|\,\mathcal{M}_j)\,p(\mathcal{M}_j)}, \tag{3.1}$$

where

$$m(\mathbf{y}|\,\mathcal{M}_i) = \int L(\mathbf{y}|\,\boldsymbol{\theta}_i, \mathcal{M}_i)\,p(\boldsymbol{\theta}_i|\,\mathcal{M}_i)\,\mathrm{d}\boldsymbol{\theta}_i \tag{3.2}$$

is the marginal likelihood for model $\mathcal{M}_i$ and $L(\mathbf{y}|\,\boldsymbol{\theta}_i, \mathcal{M}_i)$ is the likelihood function.

When taking into account model uncertainty, all knowledge about some quantity of interest, ϕ, is summarized in its posterior distribution. In forecasting, the quantity of interest is the minimum mean squared error forecast,

$$y_{T+h|T} = \mathrm{E}(y_{T+h}|\,\mathbf{y}) = \sum_{j=1}^{M} \mathrm{E}(y_{T+h}|\,\mathbf{y},\mathcal{M}_j)\,p(\mathcal{M}_j|\,\mathbf{y}), \tag{3.3}$$

where $\mathrm{E}(y_{T+h}|\,\mathbf{y},\mathcal{M}_j)$ is the forecast conditional on model $\mathcal{M}_j$. The optimal prediction is thus given by a forecast combination using the posterior model probabilities as weights .

The weights in the forecast combination (3.3) are based on the marginal likelihood, which is often interpreted as a measure of out-of-sample performance. The marginal likelihood is a predictive density based on the prior distribution $p(\boldsymbol{\theta}_i|\,\mathcal{M}_i)$ as a summary of parameter uncertainty. In large scale model selection or model averaging exercises, uninformative priors are used, causing the marginal likelihood to be a function of the residual sum of squares from a least squares fit. The prior predictive density can therefore be seen as an in-sample measure of fit. To account for possible in-sample overfitting, in Chapter 2 the marginal likelihood is replaced by the predictive likelihood.

The posterior predictive density of $\widetilde{\mathbf{y}} = (y_{m+1}, y_{m+2}, \ldots, y_T)'$, conditional on $\mathbf{y}^* = (y_1, y_2, \ldots, y_m)'$ and model $\mathcal{M}_i$, is

$$p(\tilde{\mathbf{y}}|\,\mathbf{y}^*,\mathcal{M}_i) = \int_{\boldsymbol{\theta}_i} L(\tilde{\mathbf{y}}|\,\boldsymbol{\theta}_i, \mathbf{y}^*, \mathcal{M}_i)\, p(\boldsymbol{\theta}_i|\,\mathbf{y}^*, \mathcal{M}_i)\, \mathrm{d}\boldsymbol{\theta}_i, \tag{3.4}$$

where $p(\boldsymbol{\theta}_i|\,\mathbf{y}^*, \mathcal{M}_i)$ is the posterior distribution of the parameters and $L(\tilde{\mathbf{y}}|\,\boldsymbol{\theta}_i, \mathbf{y}^*, \mathcal{M}_i)$ is the likelihood. The density of the data is averaged with respect to the posterior knowledge of the parameters. The predictive density gives the distribution of future observations, $\widetilde{\mathbf{y}}$, conditional on the observed sample $\mathbf{y}^*$. The predictive likelihood indicates how well model $\mathcal{M}_i$ accounted for the realizations $y_{m+1}, y_{m+2}, \ldots, y_T$.

The use of predictive likelihood leads to some additional practical concerns compared to model averaging based on in-sample measures of fit. A hold-out sample of l observations is needed for the predictive likelihood, in order to calculate the weights for the combined forecast. There is clearly a trade off involved in the choice of l. The number of observations available for estimation is reduced from T to $T-l$, which might have a detrimental effect. On the other hand, as the hold-out sample size increases, the predictive measure will be more stable and performs better up to the point where the predictive distribution becomes diffuse for all models and is unable to discriminate between them.

The use of the predictive likelihood leads to predictive weights $w(\mathcal{M}_i|\,\widetilde{\mathbf{y}}, \mathbf{y}^*)$ defined as

$$w(\mathcal{M}_i|\,\widetilde{\mathbf{y}}, \mathbf{y}^*) = \frac{p(\widetilde{\mathbf{y}}|\,\mathbf{y}^*,\mathcal{M}_i)\, p(\mathcal{M}_i)}{\sum_{j=1}^{M} p(\widetilde{\mathbf{y}}|\,\mathbf{y}^*,\mathcal{M}_j)\, p(\mathcal{M}_j)}. \tag{3.5}$$

These weights replace the posterior model probabilities in the forecast combination. Note, that the forecast from a each model used in the combination is still based on the full sample posterior distribution of the parameters.

3.2.1 The model, the prior and the posterior distributions

Consider a linear regression model with an intercept α and k regressors

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{3.6}$$

where $\boldsymbol{\gamma}$ is a vector of parameters and $\mathbf{Z} = (\boldsymbol{\iota}, \mathbf{X})$ is the corresponding design matrix. The disturbances $\boldsymbol{\varepsilon}$ are assumed to be $\mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Divide the T observations into two parts having m and l observations, $T = m + l$. The model can be then partitioned as

$$\mathbf{y}_{T\times 1} = \begin{bmatrix} \mathbf{y}^*_{m\times 1} \\ \widetilde{\mathbf{y}}_{l\times 1} \end{bmatrix}, \quad \mathbf{Z}_{T\times(k+1)} = \begin{bmatrix} \mathbf{Z}^*_{m\times(k+1)} \\ \widetilde{\mathbf{Z}}_{l\times(k+1)} \end{bmatrix}. \tag{3.7}$$

The first part of the data is used to convert the parameter priors $p(\boldsymbol{\theta}_i|\,\mathcal{M}_i)$ into the posterior distributions, and the second part of the sample is used for evaluating the model performance.

Implementation of BMA requires specification of the prior distribution of the parameters in the various models. It is convenient to avoid specifying informative priors for the parameters of the model. However, improper priors on the parameters that are common to all models and have identical interpretation can be used. In the case of a linear regression model the usual uninformative prior for the variance can be used

$$p\left(\sigma^2\right) \propto 1/\sigma^2. \tag{3.8}$$

Both for reasons of computational simplicity and for the interpretability of theoretical results, the most obvious choice for the prior distribution of the regression parameters is a g-prior (Zellner (1986))

$$p\left(\boldsymbol{\gamma}\,|\sigma^2, \mathcal{M}\right) \sim \mathrm{N}\left(\mathbf{0}, c\sigma^2\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\right) \tag{3.9}$$

that is, the prior mean is set to zero indicating shrinkage of the posterior towards zero and the prior variance is proportional to the data information. Fernández, Ley, and Steel (2001) recommend setting the hyperparameter c according to

$$c = \begin{cases} N^2 & \text{if } T \leq N^2, \\ T & \text{if } T > N^2, \end{cases} \tag{3.10}$$

where N is the total number of available variables.

These priors lead to a proper posterior on the regression parameters that is multivariate t-distributed with T degrees of freedom,

$$p\left(\boldsymbol{\gamma}|\,\mathbf{y}\right) \sim \mathrm{t}_{k+1}\left(\boldsymbol{\gamma}_1, S, \mathbf{M}, T\right), \tag{3.11}$$

where

$$\boldsymbol{\gamma}_1 = \frac{c}{c+1}\hat{\boldsymbol{\gamma}}, \tag{3.12}$$

is a scaled down version of the least squares estimate, and

$$S = \frac{c}{c+1}\left(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\gamma}}\right)'\left(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\gamma}}\right) + \frac{1}{c+1}\mathbf{y}'\mathbf{y}, \tag{3.13}$$

$$\mathbf{M} = \frac{c+1}{c}\mathbf{Z}'\mathbf{Z}. \tag{3.14}$$

The prior predictive density, or marginal likelihood, is then defined as

$$m\left(\mathbf{y}|\,\mathcal{M}\right) \propto (c+1)^{-(k+1)/2}\, S^{-T/2}, \tag{3.15}$$

where $\mathbf{y}^*$ and $\mathbf{Z}^*$ are empty sets and $\mathbf{y} = \widetilde{\mathbf{y}}$ and $\mathbf{Z} = \widetilde{\mathbf{Z}}$.

The posterior predictive density, the predictive likelihood, is then defined as

$$p\left(\widetilde{\mathbf{y}}|\,\mathbf{y}^*\right) \propto \left(S^*\right)^{\frac{m}{2}} \left|\mathbf{I}_l + \widetilde{\mathbf{Z}}\left(\mathbf{M}^*\right)^{-1}\widetilde{\mathbf{Z}}\right|^{-\frac{1}{2}} \times \left[S^* + \left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}_1\right)' \left(\mathbf{I}_l + \widetilde{\mathbf{Z}}\left(\mathbf{M}^*\right)^{-1}\widetilde{\mathbf{Z}}\right)^{-1} \left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}_1\right)\right]^{-(m+l)/2}, \quad (3.16)$$

where

$$S^* = \frac{c}{c+1}\left(\mathbf{y}^* - \mathbf{Z}^*\hat{\boldsymbol{\gamma}}^*\right)'\left(\mathbf{y}^* - \mathbf{Z}^*\hat{\boldsymbol{\gamma}}^*\right) + \frac{1}{c+1}\mathbf{y}^{*\prime}\mathbf{y}^*, \quad (3.17)$$

$$\mathbf{M}^* = \frac{c+1}{c}\mathbf{Z}^{*\prime}\mathbf{Z}^*. \quad (3.18)$$

See Appendix A in Chapter 2 for further details.

3.2.2 The model space

All possible combinations of the N potential predictors result in 2^N models, which prohibits evaluating the posterior model probabilities $p\left(\mathcal{M}_i|\,\mathbf{y}\right)$ for every model. A convenient method to identify a set of models $\mathfrak{M}$, without examining the full model space, is the reversible jump Markov chain Monte Carlo algorithm, see Green (1995). The details of the algorithm are described in Appendix B in Chapter 2. This Markov chain converges to the posterior model probabilities under quite general conditions and provides one way of estimating $p\left(\mathcal{M}|\,\mathbf{y}\right)$. To verify that the Markov chain captures most of the total posterior probability mass the method suggested by George and McCulloch (1997) can be implemented. This method utilizes two separate Markov chains, each starting at a random model, where the secondary chain is used to provide the estimate of the total visited probability for the primary chain.

3.2.3 Prediction intervals

Let $p\left(\phi|\mathbf{y}\right)$ and $P\left(\phi|\mathbf{y}\right)$ be the marginal posterior density function and the marginal posterior cumulative distribution function (cdf) of the parameter of interest ϕ, respectively.

A $100\left(1-\alpha\right)\%$ credible interval for a scalar function of $g\left(\phi\right)$ is any interval $[\phi_L, \phi_U]$ such that

$$\int_{\phi_L}^{\phi_U} p\big(g\left(\phi\right)\big)\,\mathrm{d}g\left(\phi\right) = P\big(g\left(\phi_U\right)\big) - P\big(g\left(\phi_L\right)\big) = 1-\alpha. \quad (3.19)$$

The interval $[\phi_L, \phi_U]$ can be chosen in different ways. Combining two $100\,(1-\alpha/2)\,\%$ intervals puts equal probability in each tail. When the interval consist of all values of $g\,(\phi)$ with $p\big(g\,(\phi)\,\big) > c$ and c is chosen so that (3.19) holds, it is the highest posterior density (HPD) interval. HPD is more plausible when $p\,(\phi|\mathbf{y})$ is not symmetric.

For the model and the priors defined in Section 3.2.1 the optimal prediction $y_{T+h|T}$ has a univariate t-distribution

$$y_{T+h|T} \sim \mathrm{t}_1\left(y_{T+h}; \mathbf{z}_{T+h}\boldsymbol{\gamma}_1, S, \left(1+\mathbf{z}_{T+h}\,(\mathbf{M})^{-1}\,\mathbf{z}'_{T+h}\right)^{-1}, T\right). \tag{3.20}$$

This implies that for the forecast combination the predictive density is a mixture of t-distributions

$$\begin{aligned} p\left(y_{T+h|T}\right) = &\sum_{i=1}^{M} p\left(\mathcal{M}_i|\mathbf{y}\right) \\ &\times \mathrm{t}_{1,i}\left(\mathbf{z}_{i,T+h}\boldsymbol{\gamma}_{1,i}, S_i, \left(1+\mathbf{z}_{i,T+h}\,(\mathbf{M}_i)^{-1}\,\mathbf{z}'_{i,T+h}\right)^{-1}, T\right). \end{aligned} \tag{3.21}$$

An equal tail interval can be constructed using

$$\alpha/2 = \int_{-\infty}^{\phi_L} p\left(y_{T+h|T}\right)\mathrm{d}y_{T+h} \tag{3.22}$$

and

$$\alpha/2 = \int_{\phi_U}^{\infty} p\left(y_{T+h|T}\right)\mathrm{d}y_{T+h}. \tag{3.23}$$

Formulas (3.22) and (3.23) are valid for both the marginal and the predictive likelihood, since the posterior distribution of the parameters is based on the whole sample. Only the weights in the mixture (3.21) are functions of marginal or predictive likelihoods.

3.3 Forecasts based on factor models

Another approach that systematically handles the relevant information contained in large datasets are factor models. Factors estimated from a large panel of data can help forecast the series of interest, so that information in a large number of variables can be used while keeping the dimension of the forecasting model small. The forecasts are constructed using a two-step procedure. At each time point $t = t_0$, data available up to t_0 are used to estimate the common

factors. The otherwise standard regression of y_{t+h} on its own lags is then augmented by contemporaneous and possibly lagged values of the estimated factors. In particular, consider

$$y_{t+h} = \alpha + \boldsymbol{\beta}'\mathbf{f}_t + \boldsymbol{\gamma}'\mathbf{y}_t + \varepsilon_{t+h}, \tag{3.24}$$

where y_{t+h} is a scalar series being forecasted h-periods ahead, $\mathbf{f}_t$ is a $(r \times 1)$ vector of common factors, and $\mathbf{y}_t$ contains y_t and possibly p lags of y_t.

If $\mathbf{f}_t, \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}$ were known, and assuming the mean of ε_t conditional on the past is zero, the minimum mean square error forecast of y_{t+h} is the conditional mean given by

$$y_{T+h|T} = \alpha_h + \boldsymbol{\beta}_h'\mathbf{f}_T + \boldsymbol{\gamma}_h'\mathbf{y}_T = \boldsymbol{\theta}_h'\mathbf{z}_T, \tag{3.25}$$

where $\mathbf{z}_T = \begin{pmatrix} \iota, & \mathbf{f}_T', & \mathbf{y}_T' \end{pmatrix}'$ and $\boldsymbol{\theta}_h = \begin{pmatrix} \alpha, & \boldsymbol{\beta}_h', & \boldsymbol{\gamma}_h'. \end{pmatrix}'$

But the vector of factors $\mathbf{f}_t$ is unobservable, and we instead observe data x_{it} which contain information about $\mathbf{f}_t$

$$x_{it} = \boldsymbol{\lambda}_i'\mathbf{f}_t + e_{it}, \qquad i = 1, \ldots, N;\ t = 1, \ldots, T, \tag{3.26}$$

where $\boldsymbol{\lambda}_i$ is vector of factor loadings and e_{it} is an idiosyncratic disturbance.

Replacing the unknown factors and parameters by their estimates, an h-step ahead prediction of the factor model (FM) is given by

$$\hat{y}_{T+h|T} = \hat{\alpha}_h + \hat{\boldsymbol{\beta}}_h'\tilde{\mathbf{f}}_T + \hat{\boldsymbol{\gamma}}_h'\mathbf{y}_T = \hat{\boldsymbol{\theta}}_h'\hat{\mathbf{z}}_T, \tag{3.27}$$

where $\hat{\mathbf{z}}_T = \begin{pmatrix} \iota, & \tilde{\mathbf{f}}_T', & \mathbf{y}_T' \end{pmatrix}'$. The factors $\mathbf{f}_t$ are estimated from x_{it} using principal components method on data up to period T. The coefficients $\hat{\alpha}_h$, $\hat{\boldsymbol{\beta}}_h$ and $\hat{\boldsymbol{\gamma}}_h$ are obtained by regressing y_{t+h} on $\tilde{\mathbf{f}}_t, \mathbf{y}_t$ and a constant for $t = 1, \ldots, T-h$. Various combinations of $\tilde{\mathbf{f}}_t$ and $\mathbf{y}_t$ give rise to different factor models. See Stock and Watson (2002b) for details.

Stock and Watson (2002a) show that $\hat{y}_{T+h|T}$ is a consistent estimator for $y_{T+h|T}$. Bai and Ng (2006) derive the limiting distributions of the parameter estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$, and the processes $\hat{y}_{T+h|T}$ and $\hat{\varepsilon}_{T+h}$, and clarify the role of N and T in the factor-augmented regression model.

3.3.1 Prediction intervals

Bai and Ng (2006) show that the distribution of the forecasting error, $\hat{\varepsilon}_{T+h} = y_{T+h|T} - \hat{y}_{T+h|T} + \varepsilon_{T+h}$, is

$$\hat{\varepsilon}_{T+h} \sim \mathrm{N}\left(0, \sigma_\varepsilon^2 + \mathrm{var}\left(\hat{y}_{T+h|T}\right)\right). \tag{3.28}$$

When constructing a confidence interval, σ_ε^2 is replaced by its consistent estimate $\hat{\sigma}_\varepsilon^2 = \frac{1}{T}\hat{\varepsilon}_T'\hat{\varepsilon}_T$. The variance of the estimated conditional mean is calculated as

$$\operatorname{var}\left(\hat{y}_{T+h|T}\right) = \frac{1}{T}\hat{\mathbf{z}}_T' \operatorname{Avar}\left(\hat{\boldsymbol{\theta}}\right)\hat{\mathbf{z}}_T + \frac{1}{N}\hat{\boldsymbol{\beta}}' \operatorname{Avar}\left(\tilde{\mathbf{f}}_T\right)\hat{\boldsymbol{\beta}}, \tag{3.29}$$

where $\operatorname{Avar}(\cdot)$ denotes the asymptotic variance. Consistent estimators, robust to heteroscedasticity, of the parameter variance and the variance of the estimated factors are

$$\widehat{\operatorname{Avar}}\left(\hat{\boldsymbol{\theta}}\right) = \left(\frac{1}{T}\sum_{t=1}^{T-h}\hat{\mathbf{z}}_t\hat{\mathbf{z}}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T-h}\hat{\varepsilon}_{t+h}^2\hat{\mathbf{z}}_t\hat{\mathbf{z}}_t'\right)\left(\frac{1}{T}\sum_{t=1}^{T-h}\hat{\mathbf{z}}_t\hat{\mathbf{z}}_t'\right)^{-1} \tag{3.30}$$

$$\widehat{\operatorname{Avar}}\left(\tilde{\mathbf{f}}_T\right) = \widetilde{\mathbf{V}}^{-1}\widetilde{\boldsymbol{\Gamma}}_T\widetilde{\mathbf{V}}^{-1}, \tag{3.31}$$

where $\widetilde{\boldsymbol{\Gamma}}_T = \frac{1}{N}\sum_{i=1}^N \tilde{e}_{iT}^2\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i'$, $\tilde{e}_{iT} = x_{iT} - \boldsymbol{\lambda}_i'\tilde{\mathbf{f}}_T$, and $\widetilde{\mathbf{V}}$ is a $(r \times r)$ diagonal matrix consisting of the r largest eigenvalues of $\mathbf{X}\mathbf{X}'/(TN)$.

A $100(1-\alpha)\,\%$ confidence interval for the forecast variable y_{T+h} can be constructed according to

$$\left(\hat{y}_{T+h|T} - z_{\alpha/2}\sqrt{\hat{\sigma}_\varepsilon^2 + \widehat{\operatorname{var}}\left(\hat{y}_{T+h|T}\right)}, \hat{y}_{T+h|T} + z_{\alpha/2}\sqrt{\hat{\sigma}_\varepsilon^2 + \widehat{\operatorname{var}}\left(\hat{y}_{T+h|T}\right)}\right), \tag{3.32}$$

where z_α is the usual α percentile of the standard normal distribution. Bai and Ng show that the analytical formulas (3.29) - (3.31) for the prediction interval (3.32) are valid regardless of the magnitude of N/T.

3.4 Forecast evaluation

The main focus of the comparison is in evaluating the performance of the forecast combination based on the predictive likelihood relative to the combination based on the marginal likelihood. Further, the performance of predictive likelihood relative to other methods is considered.

The data available for the analysis consist of quarterly time series from six countries: Canada, Germany, Italy, Japan, United Kingdom and USA, ranging from 1960Q1 to 1998Q4. The number of the series varies across countries, but 18 series are available for all countries. The data has been transformed by taking logarithms, and/or first or second differences. Four series have been calculated as a deviation from a stochastic trend and then detrended using a one-sided Hodrick-Prescott filter. Many of the series used in the forecast

exercise are used with more than one transformation. The complete list of the series, originally used in Stock and Watson (2004), and their transformations is given in Appendix C. The dependent variable, annualised quarterly GDP growth (rGDP), is modelled as being $I(1)$ in logarithms

$$y_{t+h} = \frac{400}{h} \cdot \ln\left(GDP_{t+h}/GDP_t\right). \tag{3.33}$$

The forecasting horizons considered in this chapter are $h = 1, 2$ and 4 quarters ahead.

The forecasts denoted by FM-AR in the results tables are based on equation (3.27) and include v estimated factors and p lags of y_t. Both v and p are determined by the Bayesian information criterion (BIC). FM forecasts contain only contemporary $\tilde{\mathbf{f}}_t$, with v selected by BIC. Two forecasts with v fixed are also considered: first determining the lag length of the dependent variable by BIC, and second by setting $p = 0$. These are denoted as FM-AR,v and FM,v, respectively. The maximum number of factors considered is $v = 4$ and the maximum lag length is $p = 3$.

The h-period ahead forecasts used in the forecast combination approach are formed according to

$$\hat{y}_{T+h|T} = \alpha + \boldsymbol{\beta}'\mathbf{x}_T + \sum_{j=0}^{p} \gamma_{j+1} y_{T-j}. \tag{3.34}$$

The model prior used in all the forecasting exercises is

$$p\left(\mathcal{M}_i\right) \propto \delta^{k_i} \left(1-\delta\right)^{N-k_i}, \tag{3.35}$$

where k_i is number of predictors in model $\mathcal{M}_i$. This prior can be used to downweight models containing many variables and adjust the expected model size by appropriately choosing the value for δ. In the application, the value of δ is different for each country, due to the different number of available predictors for each country. For Canada and Germany $\delta = 0.2$, for Japan and the UK $\delta = 0.25$, for Italy $\delta = 0.3$ and $\delta = 0.15$ for the US. This gives an expected number of variables in each model of between 6.5 and 7.5. The constant c in (3.9) is set to N^2, as in (3.10). The hold-out sample size, l, for evaluating the fit of the model to the data is an increasing additive sequence starting at $l = 12 - h$, and finishing at $l = 108 - h$ with increments of 6. The Markov chain is run for 5 000 000 steps, with 50 000 additional steps as burn-in.

The performance of the forecasting models is analysed using the standard forecast pseudo out-of-sample setup. First, the model parameters are estimated

using data up to a time period T, and then the forecast of y_{T+h} is calculated. The estimation sample is then extended to include the observed data at time $T+1$, the model is then re-estimated and the forecast of y_{T+h+1} is computed. This is repeated until data are no longer available to evaluate the h-step ahead forecast. The pseudo out-of-sample forecasts run from 1994Q1 to 1998Q4, giving 20 forecasts in total. The accuracy of the point forecasts is evaluated using the root mean squared forecast error (RMSFE)

$$RMSFE = \sqrt{\frac{1}{H}\sum_{t=T-h+1}^{T+H-1} e_{t+h}^2}, \tag{3.36}$$

where $e_{t+h} = y_{t+h} - \hat{y}_{t+h|t}$ is the forecast error and H is the number of evaluated h-step ahead forecasts.

3.4.1 Results

Tables A.1 - A.3 report the RMSFE relative to an AR(p) benchmark, with the lag length p determined by BIC. Figures B.1(a) - B.1(f) plot the ratio of the RMSFE for forecast combination with weights based on the predictive likelihood to the RMSFE for forecast combination with the marginal likelihood-based weights. The figures show that the predictive likelihood-based forecast combination outperforms the marginal likelihood-based forecasts. The exact location of the largest improvement is varying with the forecasting horizon and the country. The actual forecasts are calculated using full sample parameter posterior distributions up to the date of the forecast. The forecasts from a given model are therefore the same for the marginal and predictive likelihoods and the difference in the forecasting performance of the combinations is due to the different weights assigned to each of the models.

For Canada, the forecast combination with predictive weights always outperforms, with one exception, the combination based on posterior model probabilities. The maximum improvement is for small hold-out sample sizes. In the case of Germany, the predictive likelihood weights perform better for $l < 95$ at the one quarter ahead horizon, for the 2-quarter ahead forecasts for $l < 70$. For one year ahead, the predictive likelihood weights combination always outperforms the marginal weights combination. The better performance for all the horizons in the case of Italy is shown up to one coinciding point at $l + h \leq 42$. For Japan has the predictive likelihood better performance than the marginal only for $h = 4$.

The UK differs from the previous countries for 2- and 4-quarter ahead forecasts. Here, the ratio of RMSFEs exhibits rather sharp turns, and in the

case of $h = 4$ almost oscillates round the value of 1. Only for the shortest horizon is the improvement pattern similar to previously described countries. For the US, the outperformance of predictive likelihood weights for the shortest and longest horizons is for values of $l + h < 54$ and partly for the 2-quarter ahead forecasts.

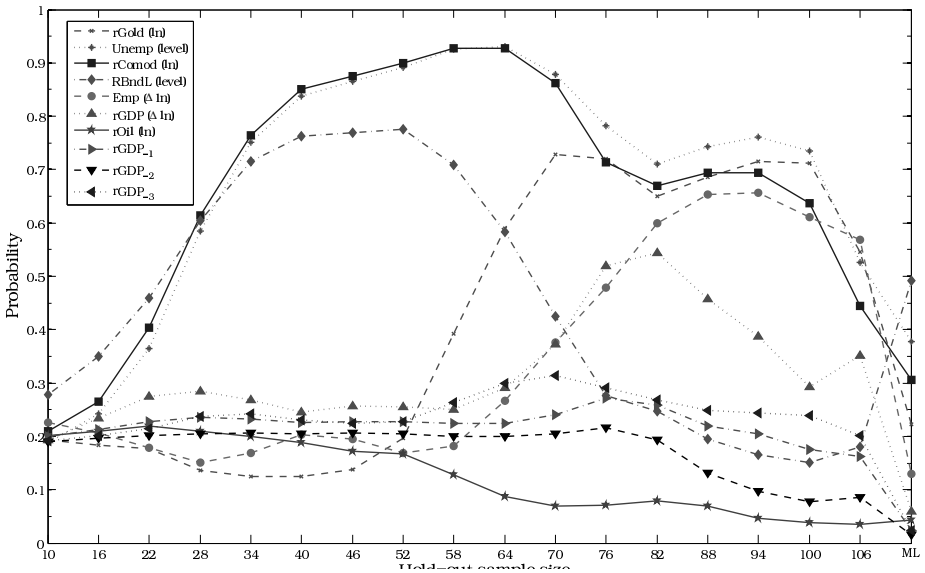
To summarize, in general, for the largest forecasting horizon considered, $h = 4$, the forecast combination with predictive weights exhibits largest improvements over the combination with in-sample weights. The recommendation from Chapter 2 to leave about 40% of the latest data for model comparison seems to apply here. This recommendation was however based on the case with shifting parameters. Ongoing structural change seems a very plausible assumption for GDP growth, a very volatile series with many turning points that is known to be difficult to forecast.

It should be noted that the hold-out sample labelled as $l + h = 108$ in the figures, contains data points from 1969Q4 until 1996Q2 $\pm$ 10 quarters. The oil shocks in 1974-75 are captured in the hold-out samples for $l+h \geq 84$. Similarly the problematic years 1981-82, the second oil shocks as well as changes in the macroeconomic policy frameworks of various countries, are appearing in the hold-out samples for $l + h \geq 54$.

The positions of these instabilities in the different hold-out sample sizes could explain the erratic behaviour of the predictive likelihood in the case of the UK. Great Britain has been subject to marked breaks in the policy regime, including income policy in the 1970s, monetary targeting in the early 1980s, exchange rate targeting in the late 1980s, and inflation targeting in the 1990s. Concentrating only on the case of 2-quarter ahead forecast, certain similarities to the shift case scenario in Chapter 2 can be seen. Figure 3.1 plots the variable inclusion probabilities[1] of several variables in the dataset for United Kingdom. The real gold price, unemployment and the real commodity price index seem to contain a break around 1975. The real gold price has the lowest inclusion probabilities around $l = 34$ and the variable employment growth round $l = 28$. For both variables the inclusion probabilities are gradually rising thereafter, suggesting an instability round 1989. Interestingly, the oil price has very low variable inclusion probabilities. There could be two reasons for this. One is that the oil price has on average no predictive power for GDP growth in UK. The other, more plausible, is that a break occurs in the variable and models containing the oil price variable receive only a negligible proportion of the posterior probabilities.

[1]The variable inclusion probability is defined as $p(x_i|\mathbf{y}) = \sum_{j=1}^{M} 1(x_i \in \mathcal{M}_j)\, p(\mathcal{M}_j|\mathbf{y})$, where $1(x_i \in \mathcal{M}_j)$ equals one if x_i is included in model j and zero otherwise.

Figure 3.1 Variable inclusion probabilities for 2-quarter ahead forecast for United Kingdom



It is also interesting to note, that for five of the six countries, lagged values of GDP are not often selected by the predictive likelihood nor the marginal likelihood. Only in the case of Italy is the second lag of the dependent variable included in the regression, although with the probability of less 1/2 on average.

Comparing the performance of the out-of-sample weights combination to the factor methods the general conclusion is that the forecast combination performs better for $h = 1$, for all countries but Germany. For the remaining horizons the RMSFE of the predictive likelihood combination is smaller that the one of the factor models for only Japan and the US. The case of Japan is the only one where the forecast combination based on predictive likelihood constantly outperforms the benchmark AR process for all horizons considered. This is in contrast to common findings that it is difficult to improve upon an AR process at longer horizons.

Tables A.4 - A.8 show the empirical coverage probabilities calculated over the $H = 20$ periods for the three forecasting horizons. This is calculated as a ratio of the number of forecast intervals that cover the actual data observations to the total number of evaluated forecasts. A good model should of course have a forecast density that is in close correspondence with the actual coverage probability. The theoretical coverage probabilities considered here are 90% and

95%. Since the empirical coverage probabilities are based on the asymptotic variance formulas (3.30) and (3.31), results based on only 20 forecast should therefore be interpreted with caution. The results are very variable. The only country with suitably narrow confidence intervals is Japan. This is also the only country that has experienced a recession over the evaluation window. On the other hand, for Canada for the 4-quarter ahead forecast the predictive likelihood interval does not reach the desired coverage of the actual value. This is caused by the actual forecasts being too close to the lower bound. However, the factor models do perform well in this case. The wider intervals may also reflect the underlying volatility of the data. The mid-late 1990s were a time of global macroeconomic stability, in contrast to the 1970 and 1980 over which models and confidence intervals have been estimated.

3.5 Conclusions

This chapter applies the predictive likelihood approach in forecast combination to GDP growth forecasting. The analysis is performed on six countries with up to 43 possible predictors available.

When using large numbers of predictors, combinations based on the predictive likelihood weights have better forecasting performance compared to existing methods such the factor models or forecast combinations using traditional marginal likelihood-based weights.

The predictive likelihood outperforms the factors models and the AR(p) process at 1-quarter ahead forecasts for all but one country. As the forecast horizon increases, the better performance of the out-of-sample approach remains for a subset of the countries considered. The largest improvement over the in-sample weights is for small values of hold-out sample sizes, which provides protection against structural breaks at the end of the sample period.

Bibliography

Artis, M. J., A. Banerjee, and M. Marcellino (2005): "Factor forecasts for the UK," *Journal of Forecasting*, 24(4), 279–298.

Bai, J., and S. Ng (2006): "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions," *Econometrica*, 74(4), 1133–1150.

Fernández, C., E. Ley, and M. F. Steel (2001): "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100(2), 381–427.

George, E. I., and R. E. McCulloch (1997): "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Green, P. J. (1995): "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82(4), 711–732.

Jacobson, T., and S. Karlsson (2004): "Finding Good Predictors for Inflation: A Bayesian Model Averaging Approach," *Journal of Forecasting*, 23(7), 479–496.

Kapetanios, G., V. Labhard, and S. Price (2006): "Forecasting Using Predictive Likelihood Model Averaging," *Economics Letters*, 91(3), 373 – 379.

Koop, G., and S. Potter (2004): "Forecasting in Dynamic Factor Models Using Bayesian Model Averaging," *Econometrics Journal*, 7(2), 550–565.

Shintani, M. (2003): "Nonlinear Analysis of Business Cycles Using Diffusion Indexes: Applications to Japan and the U.S," Discussion paper, UCLA Department of Economics, available at http://ideas.repec.org/p/cla/levrem/506439000000000168.html.

Stock, J. H., and M. W. Watson (2002a): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97(460), 1167 – 1179.

——— (2002b): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics*, 20(2), 147 – 162.

——— (2003): "Has the Business Cycle Changed and Why?," in *NBER Macroeconomics Annual 2002*, ed. by M. Gertler, and K. S. Rogoff. The MIT Press.

——— (2004): "Combination Forecasts of Output Growth in a Seven-Country Data Set," *Journal of Forecasting*, 23(6), 405 – 430.

Wright, J. H. (2003a): "Bayesian Model Averaging and Exchange Rate Forecasts," International Finance Discussion Papers 779, Board of Governors of the Federal Reserve System.

——— (2003b): "Forecasting U.S. Inflation by Bayesian Model Averaging," International Finance Discussion Papers 780, Board of Governors of the Federal Reserve System.

Zellner, A. (1986): "On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, ed. by P. K. Goel, and A. Zellner, pp. 233–243. North-Holland, Amsterdam.

Appendix A

Tables

Table A.1 Ratio of RMSFEs with an AR(p) as a benchmark for $h = 1$ quarter ahead forecasts.

	Canada	*Germany*	*Italy*	*Japan*	*UK*	*US*
DI-AR	1.1827	1.0000	1.0463	0.8579	1.0000	1.0280
DI	1.1595	1.0489	1.0463	0.8428	1.0370	1.0280
DI-AR, $v = 1$	1.1595	1.0443	1.0152	0.8891	1.0514	0.9834
DI-AR, $v = 2$	1.0813	1.0158	1.0142	0.9153	1.0134	1.0280
DI-AR, $v = 3$	1.0789	1.0294	0.9764	0.8203	1.0759	1.0332
DI-AR, $v = 4$	1.0047	0.9947	1.0300	0.8579	1.0850	1.0150
DI, $v = 1$	1.1595	1.1075	0.9357	0.8891	1.0370	0.9834
DI, $v = 2$	1.1375	1.0475	1.0142	0.9155	1.0720	1.0280
DI, $v = 3$	1.1364	1.0478	0.9764	0.8144	1.0759	1.0332
DI, $v = 4$	1.0127	1.0544	1.0300	0.8428	1.0850	1.0150
ML	1.0651	1.1204	0.9366	0.7713	1.0288	0.9657
PL, $l = 107$	1.0288	1.2293	0.9621	0.7903	1.0979	0.9575
PL, $l = 101$	1.0067	1.1775	0.9718	0.7848	1.0705	1.0225
PL, $l = 95$	1.0044	1.1349	0.9716	0.7886	1.0294	1.0231
PL, $l = 89$	0.9987	1.1028	0.9610	0.8015	0.9890	0.9786
PL, $l = 83$	0.9767	1.0827	0.9605	0.7999	1.0050	0.9733
PL, $l = 77$	0.9759	1.0517	0.9482	0.8004	0.9890	1.0202
PL, $l = 71$	0.9853	1.0655	0.9506	0.8052	1.0181	1.0311
PL, $l = 65$	1.0240	1.0382	0.9559	0.8099	1.0176	1.0317
PL, $l = 59$	1.0139	1.0218	0.9552	0.8101	1.0207	1.0212
PL, $l = 53$	1.0028	1.0957	0.9516	0.8097	0.9883	0.9871
PL, $l = 47$	1.0165	1.0804	0.9523	0.8141	1.0064	0.9479
PL, $l = 41$	0.9902	1.0670	0.9359	0.8114	0.9892	0.9441
PL, $l = 35$	0.9811	1.0568	0.9263	0.8034	0.9914	0.9403
PL, $l = 29$	0.9542	1.0067	0.9171	0.7997	1.0201	0.9223
PL, $l = 23$	0.8902	1.0390	0.9184	0.7830	1.0317	0.9101
PL, $l = 17$	0.9143	1.0190	0.9068	0.7886	1.0472	0.9288
PL, $l = 11$	0.9601	1.0150	0.9016	0.8010	1.0403	0.9324
AR, RMSFE	0.0047	0.0057	0.0071	0.0152	0.0030	0.0045

Table A.2 Ratio of RMSFEs with an AR(p) as a benchmark for $h = 2$ quarters ahead forecasts.

	Canada	*Germany*	*Italy*	*Japan*	*UK*	*US*
DI-AR	1.1914	1.0660	0.9620	0.7715	1.0000	1.0345
DI	1.2479	1.0660	0.9620	0.7982	1.0642	1.1289
DI-AR, $v = 1$	1.2477	1.0660	0.8777	0.8156	1.0564	0.9430
DI-AR, $v = 2$	1.1170	1.0631	0.9620	0.8095	1.0791	1.1257
DI-AR, $v = 3$	1.1256	1.0622	0.9081	0.7130	1.0541	1.0345
DI-AR, $v = 4$	1.0550	1.1724	0.9348	0.7715	1.0639	1.0486
DI, $v = 1$	1.2477	1.0660	0.8106	0.8034	1.0642	0.9712
DI, $v = 2$	1.2469	1.0631	0.9620	0.8284	1.0791	1.0690
DI, $v = 3$	1.2355	1.0622	0.9081	0.7342	1.0541	1.0280
DI, $v = 4$	1.1061	1.1724	0.9348	0.7982	1.0639	0.9742
ML	1.3627	1.2221	0.9754	0.6367	0.9723	0.8410
PL, $l = 106$	1.3537	1.3174	1.0104	0.6649	1.0463	0.9007
PL, $l = 100$	1.3359	1.3169	1.0169	0.6611	1.0605	0.9274
PL, $l = 94$	1.3190	1.2919	1.0180	0.6431	1.0603	0.8938
PL, $l = 88$	1.3213	1.2452	1.0215	0.6449	1.0045	0.8760
PL, $l = 82$	1.3483	1.2258	1.0333	0.6436	1.3562	0.8665
PL, $l = 76$	1.3531	1.2361	1.0219	0.6551	1.2093	0.8034
PL, $l = 70$	1.3491	1.2429	1.0290	0.6724	1.0788	0.8852
PL, $l = 64$	1.4014	1.2092	1.0080	0.6790	1.1063	0.8938
PL, $l = 58$	1.3335	1.1611	1.0132	0.6743	1.0633	0.9679
PL, $l = 52$	1.3468	1.1673	1.0050	0.6740	0.9873	0.9985
PL, $l = 46$	1.2751	1.1620	0.9968	0.6795	0.9926	0.8932
PL, $l = 40$	1.2260	1.1863	0.9740	0.6696	0.9308	0.8376
PL, $l = 34$	1.1910	1.1554	0.9514	0.6602	0.9951	0.8466
PL, $l = 28$	1.1395	1.0748	0.9211	0.6642	1.0447	0.8048
PL, $l = 22$	1.0007	1.1418	0.8910	0.6563	1.0671	0.8311
PL, $l = 16$	0.9686	1.1112	0.8404	0.6337	1.0610	0.8772
PL, $l = 10$	1.0427	1.1048	0.8157	0.6325	1.0809	0.8687
AR, RMSFE	0.0084	0.0085	0.0117	0.0250	0.0052	0.0060

Table A.3 Ratio of RMSFEs with an AR(p) as a benchmark for $h = 4$ quarters ahead forecasts.

	Canada	*Germany*	*Italy*	*Japan*	*UK*	*US*
DI-AR	1.2424	1.3116	0.9132	0.7440	0.9423	1.0089
DI	1.2424	1.3116	0.8937	0.8438	1.0662	0.8901
DI-AR, $v = 1$	1.3843	0.9607	0.9436	0.8439	1.0305	1.0294
DI-AR, $v = 2$	1.4431	1.0806	0.8937	0.7701	1.0662	0.9209
DI-AR, $v = 3$	1.4151	1.0643	0.8780	0.6285	0.9950	1.0089
DI-AR, $v = 4$	1.2424	1.3716	0.8223	0.7440	0.9717	0.8811
DI, $v = 1$	1.3843	0.9607	0.7617	0.8123	1.1914	0.9219
DI, $v = 2$	1.4431	1.0806	0.8937	0.7860	1.0662	0.9209
DI, $v = 3$	1.4444	1.0643	0.8780	0.7001	0.9950	0.8901
DI, $v = 4$	1.2424	1.3716	0.8223	0.8438	0.9717	0.8811
ML	2.0322	1.5795	1.0748	0.6119	1.1870	1.0127
PL, $l = 104$	2.0081	1.3392	1.1028	0.6248	1.2574	0.9882
PL, $l = 98$	1.9836	1.5485	1.1352	0.6063	1.1130	1.0163
PL, $l = 92$	1.9574	1.4780	1.1358	0.5795	1.0309	1.0163
PL, $l = 86$	1.9663	1.4950	1.1866	0.6026	1.3804	1.1336
PL, $l = 80$	1.9461	1.5212	1.1599	0.6009	1.4878	1.2135
PL, $l = 74$	1.9539	1.5076	1.1611	0.5933	1.2477	1.2048
PL, $l = 68$	1.9704	1.5068	1.1483	0.5983	1.0162	1.2755
PL, $l = 62$	2.0075	1.4671	1.1401	0.6101	1.1036	1.1773
PL, $l = 56$	1.9469	1.3308	1.1126	0.6116	1.1760	1.1701
PL, $l = 50$	1.8673	1.3185	1.1211	0.6142	1.1514	1.0960
PL, $l = 44$	1.8284	1.3471	1.1011	0.6138	1.1169	0.9554
PL, $l = 38$	1.7903	1.3032	1.0719	0.6084	1.0332	0.8157
PL, $l = 32$	1.7607	1.2601	1.0309	0.5933	1.1267	0.8060
PL, $l = 26$	1.6247	1.1638	0.9900	0.5733	1.2397	0.8052
PL, $l = 20$	1.2882	1.2184	0.9330	0.5807	1.2923	0.6609
PL, $l = 14$	1.1612	1.2285	0.7716	0.5553	1.1291	0.6320
PL, $l = 8$	1.1058	1.1827	0.6906	0.5441	1.1602	0.7855
AR, RMSFE	0.0138	0.0135	0.0228	0.0449	0.0085	0.0097

Table A.4 Empirical vs. nominal coverage probability, Canada.

	$h=1$		$h=2$		$h=4$	
Nominal coverage	0.90	0.95	0.90	0.95	0.90	0.95
AR	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR	1.00	1.00	0.95	1.00	1.00	1.00
DI	1.00	1.00	0.95	1.00	1.00	1.00
DI-AR, $v=1$	1.00	1.00	0.95	1.00	1.00	1.00
DI-AR, $v=2$	1.00	1.00	1.00	1.00	0.90	1.00
DI-AR, $v=3$	1.00	1.00	1.00	1.00	0.90	1.00
DI-AR, $v=4$	1.00	1.00	1.00	1.00	0.90	1.00
DI, $v=1$	1.00	1.00	0.95	1.00	1.00	1.00
DI, $v=2$	1.00	1.00	0.95	1.00	0.90	1.00
DI, $v=3$	1.00	1.00	0.95	1.00	0.90	1.00
DI, $v=4$	1.00	1.00	1.00	1.00	0.90	1.00
ML	1.00	1.00	1.00	1.00	0.60	0.75
PL, $l=108-h$	1.00	1.00	1.00	1.00	0.45	0.75
PL, $l=102-h$	1.00	1.00	1.00	1.00	0.50	0.65
PL, $l=96-h$	1.00	1.00	1.00	1.00	0.50	0.65
PL, $l=90-h$	1.00	1.00	1.00	1.00	0.50	0.70
PL, $l=84-h$	1.00	1.00	1.00	1.00	0.50	0.70
PL, $l=78-h$	1.00	1.00	1.00	1.00	0.50	0.70
PL, $l=72-h$	1.00	1.00	1.00	1.00	0.50	0.70
PL, $l=66-h$	1.00	1.00	1.00	1.00	0.50	0.70
PL, $l=60-h$	1.00	1.00	1.00	1.00	0.45	0.70
PL, $l=54-h$	1.00	1.00	1.00	1.00	0.55	0.80
PL, $l=48-h$	1.00	1.00	1.00	1.00	0.75	0.90
PL, $l=42-h$	1.00	1.00	1.00	1.00	0.75	0.95
PL, $l=36-h$	1.00	1.00	1.00	1.00	0.80	0.90
PL, $l=30-h$	1.00	1.00	1.00	1.00	0.80	0.90
PL, $l=24-h$	1.00	1.00	1.00	1.00	0.85	0.95
PL, $l=18-h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l=12-h$	1.00	1.00	1.00	1.00	1.00	1.00

Table A.5 Empirical vs. nominal coverage probability, Germany.

	$h = 1$		$h = 2$		$h = 4$	
Nominal coverage	0.90	0.95	0.90	0.95	0.90	0.95
AR	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR	1.00	1.00	1.00	1.00	1.00	1.00
DI	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR, $v = 1$	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR, $v = 2$	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR, $v = 3$	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR, $v = 4$	1.00	1.00	1.00	1.00	1.00	1.00
DI, $v = 1$	1.00	1.00	1.00	1.00	1.00	1.00
DI, $v = 2$	1.00	1.00	1.00	1.00	1.00	1.00
DI, $v = 3$	1.00	1.00	1.00	1.00	1.00	1.00
DI, $v = 4$	1.00	1.00	1.00	1.00	1.00	1.00
ML	1.00	1.00	1.00	1.00	0.90	1.00
PL, $l = 108 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 102 - h$	1.00	1.00	1.00	1.00	0.90	1.00
PL, $l = 96 - h$	1.00	1.00	1.00	1.00	0.95	1.00
PL, $l = 90 - h$	1.00	1.00	1.00	1.00	0.90	1.00
PL, $l = 84 - h$	1.00	1.00	1.00	1.00	0.90	1.00
PL, $l = 78 - h$	1.00	1.00	1.00	1.00	0.90	1.00
PL, $l = 72 - h$	1.00	1.00	1.00	1.00	0.90	1.00
PL, $l = 66 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 60 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 54 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 48 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 42 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 36 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 30 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 24 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 18 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 12 - h$	1.00	1.00	1.00	1.00	1.00	1.00

Table A.6 Empirical vs. nominal coverage probability, Italy.

	$h = 1$		$h = 2$		$h = 4$	
Nominal coverage	0.90	0.95	0.90	0.95	0.90	0.95
AR	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR	1.00	1.00	1.00	1.00	1.00	1.00
DI	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR, $v = 1$	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR, $v = 2$	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR, $v = 3$	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR, $v = 4$	1.00	1.00	1.00	1.00	1.00	1.00
DI, $v = 1$	1.00	1.00	1.00	1.00	1.00	1.00
DI, $v = 2$	1.00	1.00	1.00	1.00	1.00	1.00
DI, $v = 3$	1.00	1.00	1.00	1.00	1.00	1.00
DI, $v = 4$	1.00	1.00	1.00	1.00	1.00	1.00
ML	1.00	1.00	1.00	1.00	0.95	0.95
PL, $l = 108 - h$	1.00	1.00	0.95	1.00	0.95	0.95
PL, $l = 102 - h$	1.00	1.00	0.95	1.00	0.95	0.95
PL, $l = 96 - h$	1.00	1.00	0.95	1.00	0.95	0.95
PL, $l = 90 - h$	1.00	1.00	0.95	1.00	0.90	0.95
PL, $l = 84 - h$	1.00	1.00	1.00	1.00	0.95	0.95
PL, $l = 78 - h$	1.00	1.00	1.00	1.00	0.95	0.95
PL, $l = 72 - h$	1.00	1.00	1.00	1.00	0.95	0.95
PL, $l = 66 - h$	1.00	1.00	1.00	1.00	0.95	0.95
PL, $l = 60 - h$	1.00	1.00	1.00	1.00	0.95	0.95
PL, $l = 54 - h$	1.00	1.00	1.00	1.00	0.95	0.95
PL, $l = 48 - h$	1.00	1.00	1.00	1.00	0.95	0.95
PL, $l = 42 - h$	1.00	1.00	1.00	1.00	0.95	0.95
PL, $l = 36 - h$	1.00	1.00	1.00	1.00	0.95	0.95
PL, $l = 30 - h$	1.00	1.00	1.00	1.00	0.95	1.00
PL, $l = 24 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 18 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 12 - h$	1.00	1.00	1.00	1.00	1.00	1.00

Table A.7 Empirical vs. nominal coverage probability, Japan.

	$h=1$		$h=2$		$h=4$	
Nominal coverage	0.90	0.95	0.90	0.95	0.90	0.95
AR	0.90	0.95	0.75	0.95	0.80	0.90
DI-AR	0.85	0.90	0.85	0.95	0.80	0.90
DI	0.90	0.95	0.85	0.95	0.80	0.95
DI-AR, $v=1$	0.95	0.95	0.95	0.95	0.90	0.95
DI-AR, $v=2$	0.95	0.95	0.90	0.95	0.95	0.95
DI-AR, $v=3$	0.90	0.95	0.90	0.95	0.95	0.95
DI-AR, $v=4$	0.85	0.90	0.85	0.95	0.80	0.90
DI, $v=1$	0.95	0.95	0.95	0.95	0.95	0.95
DI, $v=2$	0.95	0.95	0.90	0.95	0.90	0.95
DI, $v=3$	0.90	0.95	0.90	0.95	0.95	0.95
DI, $v=4$	0.90	0.95	0.85	0.95	0.80	0.95
ML	0.90	0.90	0.90	0.90	0.85	0.95
PL, $l=108-h$	0.85	0.90	0.90	0.95	0.90	0.95
PL, $l=102-h$	0.80	0.95	0.95	0.95	0.80	0.95
PL, $l=96-h$	0.80	0.90	0.90	0.95	0.85	0.95
PL, $l=90-h$	0.85	0.90	0.95	0.95	0.85	0.95
PL, $l=84-h$	0.90	0.90	0.95	0.95	0.90	0.95
PL, $l=78-h$	0.90	0.90	0.90	0.95	0.90	0.95
PL, $l=72-h$	0.90	0.95	0.90	0.95	0.85	0.95
PL, $l=66-h$	0.90	0.95	0.90	0.95	0.85	0.95
PL, $l=60-h$	0.90	0.95	0.90	0.95	0.90	0.95
PL, $l=54-h$	0.90	0.95	0.90	0.95	0.90	0.95
PL, $l=48-h$	0.90	0.90	0.90	0.95	0.95	0.95
PL, $l=42-h$	0.90	0.90	0.90	0.95	0.95	0.95
PL, $l=36-h$	0.90	0.95	0.90	0.95	0.95	0.95
PL, $l=30-h$	0.90	0.90	0.90	0.95	0.95	0.95
PL, $l=24-h$	0.90	0.95	0.90	0.90	0.95	0.95
PL, $l=18-h$	0.90	0.95	0.90	0.95	0.90	0.95
PL, $l=12-h$	0.90	0.95	0.95	0.95	0.90	0.95

Table A.8 Empirical vs. nominal coverage probability, UK and US.

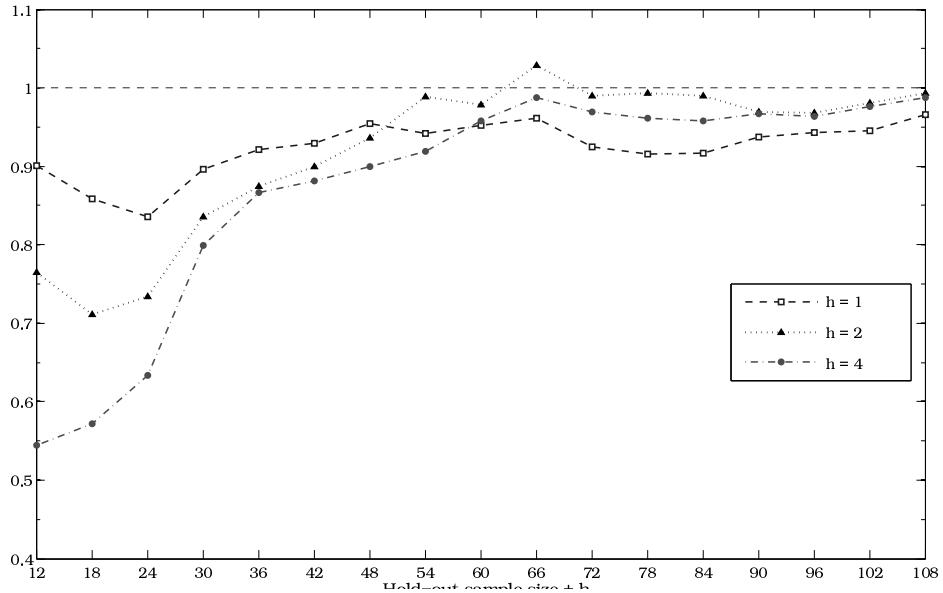
	$h = 1$		$h = 2$		$h = 4$	
Nominal coverage	0.90	0.95	0.90	0.95	0.90	0.95
AR	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR	1.00	1.00	1.00	1.00	1.00	1.00
DI	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR, $v = 1$	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR, $v = 2$	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR, $v = 3$	1.00	1.00	1.00	1.00	1.00	1.00
DI-AR, $v = 4$	1.00	1.00	1.00	1.00	1.00	1.00
DI, $v = 1$	1.00	1.00	1.00	1.00	1.00	1.00
DI, $v = 2$	1.00	1.00	1.00	1.00	1.00	1.00
DI, $v = 3$	1.00	1.00	1.00	1.00	1.00	1.00
DI, $v = 4$	1.00	1.00	1.00	1.00	1.00	1.00
ML	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 108 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 102 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 96 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 90 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 84 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 78 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 72 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 66 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 60 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 54 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 48 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 42 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 36 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 30 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 24 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 18 - h$	1.00	1.00	1.00	1.00	1.00	1.00
PL, $l = 12 - h$	1.00	1.00	1.00	1.00	1.00	1.00

Appendix B

Figures

Figure B.1 Ratio of RMSFE for predictive likelihood and marginal likelihood as a function of the hold-out sample size.

(a) Canada



(b) Germany

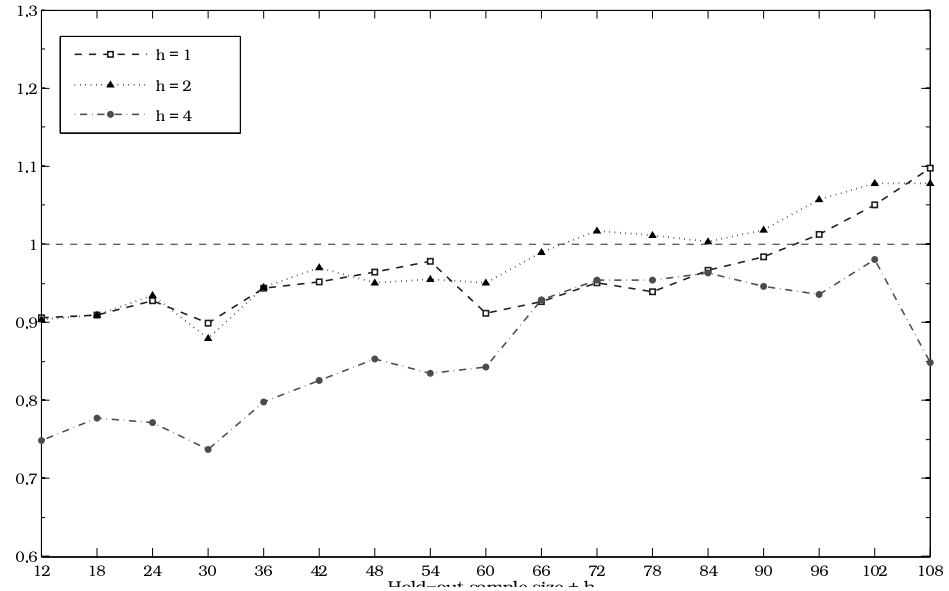
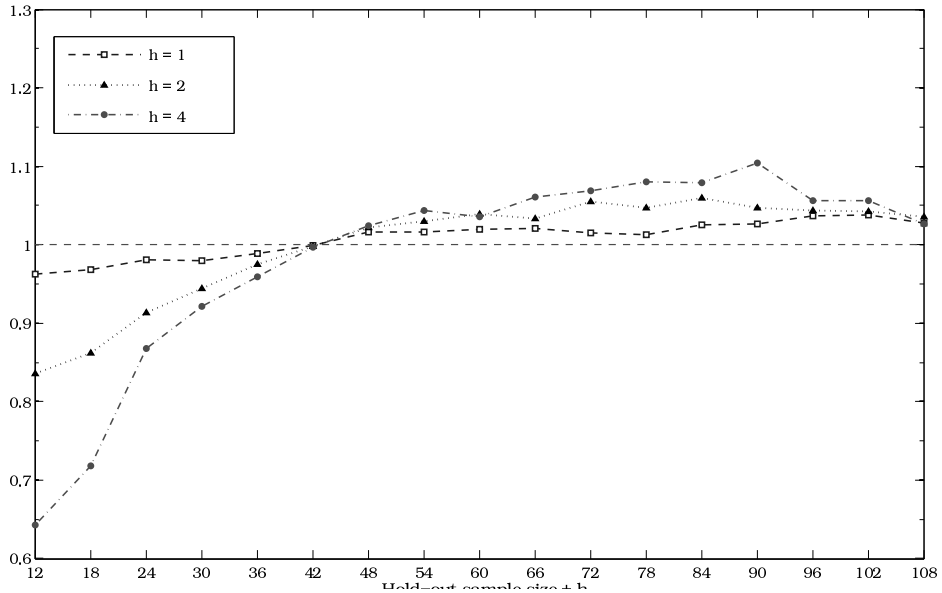


Figure B.1 Ratio of RMSFE for predictive likelihood and marginal likelihood as a function of the hold-out sample size.

(c) Italy



(d) Japan

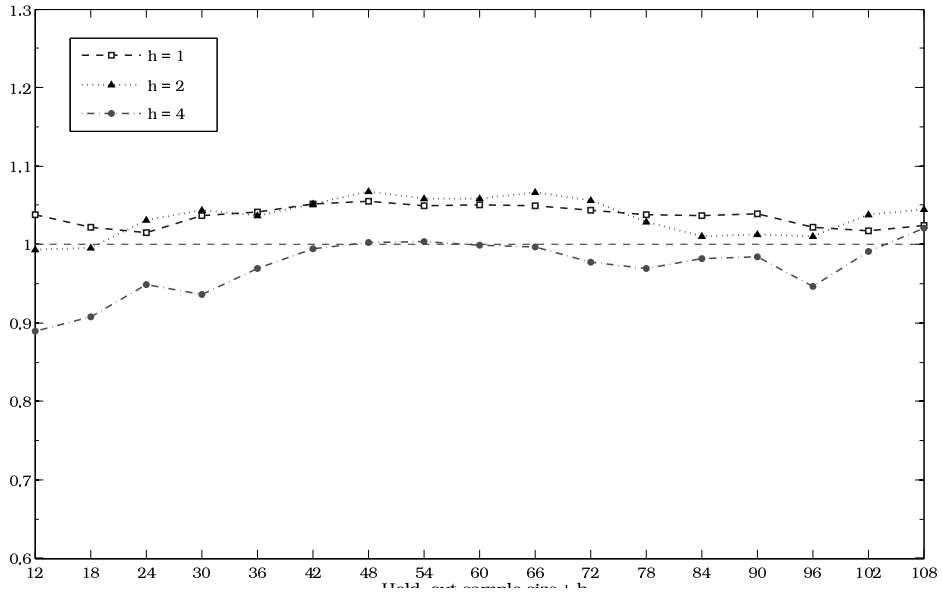
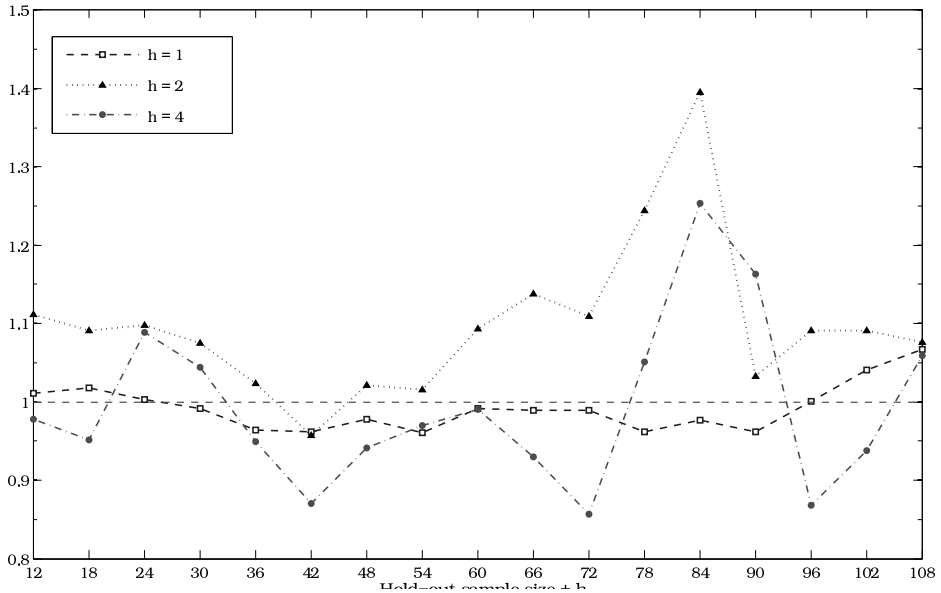
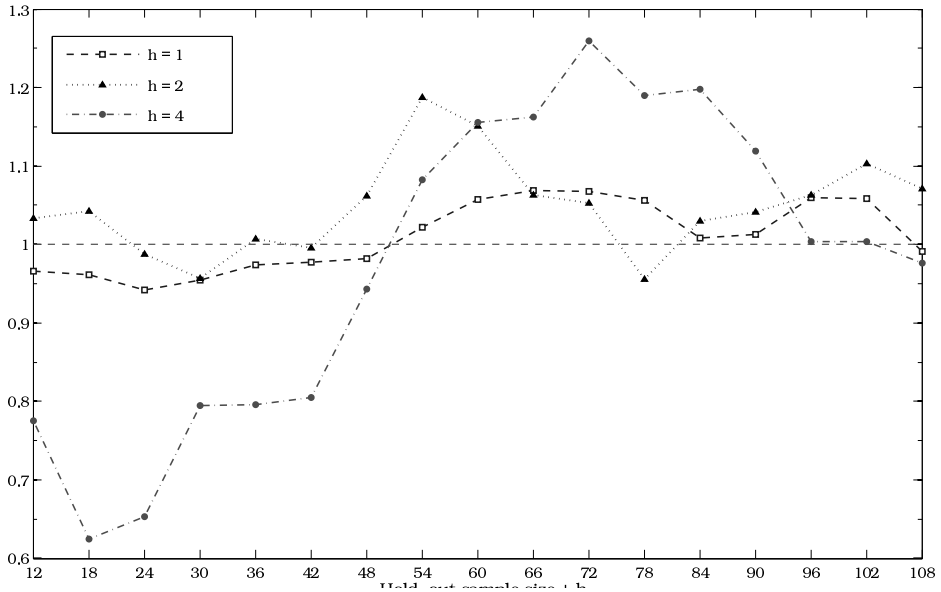


Figure B.1 Ratio of RMSFE for predictive likelihood and marginal likelihood as a function of the hold-out sample size.

(e) United Kingdom



(f) Unites States



Appendix C

Datasets summary

Data set for each country consist of those series (and transfomations) indicated by $\times$. The total number of series for each country is in Table C.4.

Table C.1 Asset Prices

Variable	Description		CA	DE	IT	JP	UK	US
ROvnght	Interest rate, overnight	level				×		
RTBill	Int. rate, short term gov. bill	level	×					×
RBndL	Int. rate, long term gov. bonds	level					×	
ROvnght	Interest rate, overnight	Δ		×				×
RBndM	Int. rate, med. term gov. bonds	Δ						×
RBndL	Int. rate, long term gov. bonds	Δ		×			×	
rROvnght	Real overnight rate	level		×				×
rRBndM	Real medium term bond rate	level			×			×
rRBndL	Real long term bond rate	level	×		×			
rROvnght	Real overnight rate	Δ				×		
rRTBbill	Real short term bill rate	Δ	×					×
rRBndM	Real medium term bond rate	Δ			×			×
rRBndL	Real long term bond rate	Δ	×		×			×
RSpread	Term spread: RBndL -ROvnght	level		×				×
StockP	Stock Price Index	$\Delta \ln$	×	×	×	×	×	×
rStockP	Real Stock Price Index	$\Delta \ln$	×	×	×	×	×	×
Gold	Gold Prices	$\Delta \ln$	×	×	×	×	×	×
Gold	Gold Prices	$\Delta^2 \ln$	×	×	×	×	×	×
rGold	Real Gold Prices	ln	×	×	×	×	×	×
rGold	Real Gold Prices	$\Delta \ln$	×	×	×	×	×	×

Table C.2 Wages, Goods and Commodity Prices

Variable	Description		CA	DE	IT	JP	UK	US
Earn	Wages	$\Delta \ln$	×			×		×
Earn	Wages	$\Delta^2 \ln$	×			×		×
rOil	Real Oil Prices	ln	×	×	×	×	×	×
Commod	Commodity Price Index	$\Delta \ln$	×	×	×	×	×	×
Commod	Commodity Price Index	$\Delta^2 \ln$	×	×	×	×	×	×
rCommod	Real Commodity Price Index	ln	×	×	×	×	×	×
rCommod	Real Commodity Price Index	$\Delta \ln$	×	×	×	×	×	×

Table C.3 Activity

Variable	Description		CA	DE	IT	JP	UK	US
rGDP	Real GDP	$\Delta \ln$	×	×	×	×	×	×
rGDP	Real GDP	HP	×	×	×	×	×	×
IP	Index of Industrial Production	$\Delta \ln$	×	×	×	×	×	×
IP	Index of Industrial Production	HP	×	×	×	×	×	×
Emp	Employment	$\Delta \ln$	×	×		×	×	×
Emp	Employment	HP	×	×		×	×	×
Unemp	Unemployment Rate	level	×		×	×	×	×
Unemp	Unemployment Rate	Δ	×		×	×	×	×
Unemp	Unemployment Rate	HP	×		×	×	×	×
PGDP	GDP Deflator	$\Delta \ln$	×	×	×	×	×	×
PGDP	GDP Deflator	$\Delta^2 \ln$	×	×		×	×	×
CPI	Consumer Price Index	$\Delta \ln$	×	×	×	×	×	×
CPI	Consumer Price Index	$\Delta^2 \ln$	×	×	×	×	×	×
PPI	Producer Price Index	$\Delta \ln$	×	×		×	×	×
PPI	Producer Price Index	$\Delta^2 \ln$	×	×		×	×	×

Table C.4 Money

Variable	Description		CA	DE	IT	JP	UK	US
M1	Money: M1	$\Delta \ln$	×	×				×
M1	Money: M1	$\Delta^2 \ln$	×	×				×
M2	Money: M2	$\Delta \ln$		×				×
M2	Money: M2	$\Delta^2 \ln$		×				×
rM1	Real Money: M1	$\Delta \ln$	×	×				×
rM2	Real Money: M2	$\Delta \ln$		×				×
Total Number of Series			35	33	25	30	28	43

Chapter 4

Computational Efficiency in Bayesian Model and Variable Selection

4.1 Introduction

Bayesian model averaging (BMA) was introduced by Leamer (1978) some 30 years ago. It has grown in popularity in the last decade as new theoretical developments have become available and computer power easily accessible. Hoeting, Madigan, Raftery, and Volinsky (1999) provide a comprehensive historical overview and summarization of the literature on this topic. BMA has been applied successfully to many statistical model classes including linear regression, Raftery, Madigan, and Hoeting (1997), Fernández, Ley, and Steel (2001); discrete graphical models, Madigan and Raftery (1994); survival analysis, Raftery, Madigan, and Volinsky (1995); factor-based models, Koop and Potter (2004); and large macroeconomic panels, Jacobson and Karlsson (2004), in all cases improving predictive performance in the presence of model uncertainty.

At first glance, Bayesian model averaging is straightforward to implement: one needs the marginal distribution of the data, the prior probabilities of the models and the posterior distribution of the quantity of interest conditional on each model. In linear regression, for nicely behaved prior distributions, all these components are available in closed form. Even if one has the closed form, another problem is that in many applications, the model space is too large to allow enumeration of all models, and beyond 20-25 variables, see George and McCulloch (1997), estimation of posterior model probabilities and BMA must be based on a sample of models.

For a variable selection problem in a linear regression setting with 50 potential explanatory variables, implying $2^{50} \approx 10^{15}$ different models, the CPU time for a brute force attack would be close to 5 millennia with our fastest algorithm on a standard desktop PC. While this can obviously be speeded up by throwing more hardware at the problem and the computations are trivial to parallelize, the computational burden is simply too large. Assuming a 100-fold increase in processor speed it would still take over 17 days to evaluate all the models on a 1024 node cluster, which clearly puts a brute force approach beyond everyday use. Markov chain Monte Carlo (MCMC) methods provide a stochastic method of obtaining sample from the model posterior distributions. For the commonly used linear regression models with uninformative priors when there is a wealth of potential predictors, these methods have to be run sufficiently long in order to achieve satisfactory level of convergence. It is thus important to be able to evaluate the models directly by an analytical marginal likelihood, or indirectly by generating pseudo random numbers from the posterior distribution, in an efficient manner. It is of equal importance that the MCMC scheme is well designed and moves quickly through the model

space.

This chapter investigates these issues in the context of linear regression models. Algorithms for solving least squares problems and various MCMC algorithms for exploring the model space are evaluated both in terms of speed and accuracy.

The chapter is organized as follows, the next section 4.2 introduces Bayesian model averaging and sets the stage for the remainder of the chapter. Section 4.3 reviews methods for solving linear normal equations and Section 4.4 evaluates the speed and accuracy of the algorithms. Section 4.5 describes several Markov chain samplers and Section 4.6 evaluates their performance. Finally, Section 4.7 concludes.

4.2 Bayesian model averaging and linear regression

Bayesian model averaging and model selection depends on the ability to calculate posterior model probabilities

$$p\left(\mathcal{M}_i|\,\mathbf{y}\right) = \frac{m\left(\mathbf{y}|\,\mathcal{M}_i\right)p\left(\mathcal{M}_i\right)}{\sum_{j=1}^{M} m\left(\mathbf{y}|\,\mathcal{M}_j\right)p\left(\mathcal{M}_j\right)}, \tag{4.1}$$

and in particular the marginal likelihoods

$$m\left(\mathbf{y}|\,\mathcal{M}_i\right) = \int L\left(\mathbf{y}|\,\boldsymbol{\theta}_i, \mathcal{M}_i\right)p\left(\boldsymbol{\theta}_i|\,\mathcal{M}_i\right)\mathrm{d}\boldsymbol{\theta}_i, \tag{4.2}$$

for the considered models $\mathcal{M}_i$, $i = 1, \ldots, M$. If M is not too large and the models are tractable in the sense that the marginal likelihoods are available in closed form, this can be solved by brute force calculation of all the posterior model probabilities. In the large scale problems motivating the current chapter, brute force calculations are not feasible even when the marginal likelihood is available in closed form.

Instead, Markov chain Monte Carlo methods are used to estimate the posterior model probabilities. These are typically variations on the reversible jump MCMC (RJMCMC) scheme developed by Green (1995). A typical RJMCMC algorithm, and the basis for all the algorithms considered in this chapter, is given as Algorithm 4.1. The Markov chain moves between models and converges to the posterior model probabilities under general conditions, and the output can be used to estimate the posterior model probabilities after a suitable burn-in. With the very large model sets considered here, it is clear that the chain can not visit all, or even a majority, of the models in a reasonable amount of time. It is thus important that the chain moves well through the

Algorithm 4.1 Reversible jump Markov chain Monte Carlo

Suppose that the Markov chain is at model $\mathcal{M}$, having parameters $\boldsymbol{\theta}_{\mathcal{M}}$, where $\boldsymbol{\theta}_{\mathcal{M}}$ has dimension $\dim(\boldsymbol{\theta}_{\mathcal{M}})$.

A: Marginal likelihood is not available in closed form

1. Propose a jump from model $\mathcal{M}$ to a new model $\mathcal{M}'$ with probability $j(\mathcal{M}'|\mathcal{M})$.
2. Generate a vector $\mathbf{u}$ (which can have a different dimension than $\boldsymbol{\theta}_{\mathcal{M}'}$) from a specified proposal density $q(\mathbf{u}|\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}, \mathcal{M}')$.
3. Set $(\boldsymbol{\theta}_{\mathcal{M}'}, \mathbf{u}') = g_{\mathcal{M},\mathcal{M}'}(\boldsymbol{\theta}_{\mathcal{M}}, \mathbf{u})$, where $g_{\mathcal{M},\mathcal{M}'}$ is a specified invertible function. Hence $\dim(\boldsymbol{\theta}_{\mathcal{M}}) + \dim(\mathbf{u}) = \dim(\boldsymbol{\theta}_{\mathcal{M}'}) + \dim(\mathbf{u}')$. Note that $g_{\mathcal{M},\mathcal{M}'} = g^{-1}_{\mathcal{M}',\mathcal{M}}$.
4. Accept the proposed move with probability

$$\alpha = \min\left\{1, \frac{L\left(\mathbf{y}|\boldsymbol{\theta}_{\mathcal{M}',\mathcal{M}'}\right) p(\boldsymbol{\theta}_{\mathcal{M}'}|\mathcal{M}')\, p(\mathcal{M}')\, j(\mathcal{M}|\mathcal{M}')}{L(\mathbf{y}|\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M})\, p(\boldsymbol{\theta}_{\mathcal{M}}|\mathcal{M})\, p(\mathcal{M})\, j(\mathcal{M}'|\mathcal{M})} \times \frac{q(\mathbf{u}'|\boldsymbol{\theta}_{\mathcal{M}'}, \mathcal{M}', \mathcal{M})}{q(\mathbf{u}|\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}, \mathcal{M}')} \left|\frac{\partial g_{\mathcal{M},\mathcal{M}'}(\boldsymbol{\theta}_{\mathcal{M}}, \mathbf{u})}{\partial(\boldsymbol{\theta}_{\mathcal{M}}, \mathbf{u})}\right|\right\}. \quad (4.3)$$

5. Set $\mathcal{M} = \mathcal{M}'$ if the move is accepted and stay at $\mathcal{M}$ otherwise.

B: Marginal likelihood is available in closed form

When the marginal likelihood is available in closed form the acceptance probability (4.3) can be simplified substantially by employing the fiction that the proposal distribution for the parameters is the posterior distribution, $q(\mathbf{u}|\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}, \mathcal{M}')$ is the posterior $p(\boldsymbol{\theta}_{\mathcal{M}'}|\mathbf{y}, \mathcal{M}')$. The Jacobian is then unity and the acceptance probability simplifies to

$$\alpha = \min\left\{1, \frac{m(\mathbf{y}|\mathcal{M}')\, p(\mathcal{M}')\, j(\mathcal{M}|\mathcal{M}')}{m(\mathbf{y}|\mathcal{M})\, p(\mathcal{M})\, j(\mathcal{M}'|\mathcal{M})}\right\}. \quad (4.4)$$

Clearly steps 2 and 3 above are then unnecessary and the algorithm simplifies to

1. Propose a jump from model $\mathcal{M}$ to a new model $\mathcal{M}'$ with probability $j(\mathcal{M}'|\mathcal{M})$.
2. Accept the move with probability (4.4) otherwise stay at the current model.

model space and finds the models with high posterior probabilities.[1] The output of the chain can then be taken as an estimate of the (normalized) posterior probabilities for the visited models. When the marginal likelihoods are available in closed form, the exact posterior probabilities, conditional on the set of visited models, can be calculated. Besides being more accurate than the relative frequencies from the chain, this can also be used as a check of the convergence properties of the chain.

The way the chain moves through the model space is largely determined by the type of moves allowed and their probabilities, or in terms of Algorithm 4.1, the jump distribution $j(\mathcal{M}'|\mathcal{M})$. Let $\boldsymbol{\gamma}$ be a binary vector with $\gamma_j = 1$ indicating inclusion and $\gamma_j = 0$ exclusion of variable j in a variable selection problem. A basic implementation of the RJMCMC algorithm for variable selection would use simple, local, model changing moves:

- **Add/Drop** Draw an integer j from a uniform distribution on $1, \ldots, N$ and flip the value of γ_j. This adds or drops variable j from the model. This gives $j(\mathcal{M}'|\mathcal{M}) = j(\mathcal{M}|\mathcal{M}') = 1/N$.
- **Swap** Select an index i at random from $\{i : \gamma_i = 1\}$, an index j from $\{j : \gamma_j = 0\}$ and set $\gamma_i = 0, \gamma_j = 1$. This replaces variable i with variable j. The jump probabilities are $j(\mathcal{M}'|\mathcal{M}) = j(\mathcal{M}|\mathcal{M}') = 1/k(N-k)$, where k is the number of variables in the model. This replaces variable i with variable j.

Using only the Add/Drop move yields the Markov chain Monte Carlo model composition $(\text{MC})^3$ algorithm of Madigan and York (1995). The Swap move is used in conjunction with the Add/Drop by, among others; Denison, Mallick, and Smith (1998), and Jacobson and Karlsson (2004).

Moves like Add/Drop and Swap are attractive since they are very easy to implement, but their simple nature might cause the chain to mix badly and converge slowly. Consequently a number of alternative schemes has been proposed with the aim of speeding up convergence, in particular for the common situation, where there is a high degree of multicollinearity between potential explanatory variables. However, little is known about the relative merits of these sampling schemes, and one aim of this chapter is to evaluate them against a common set of benchmarks. There are two important aspects which we focus on:

1. How quickly will the algorithm account for all but a negligible portion of the total posterior probability?

[1] The capture-recapture technique proposed by George and McCulloch (1997) can be used to estimate the fraction of the total posterior probability accounted for by the chain.

2. How quickly will the algorithm converge to the posterior distribution over the (visited) models?

In general, the first item will be easier to satisfy, and is needed with models where the marginal likelihood is available in closed form and exact posterior model probabilities can be calculated for the visited models. The second might require many more steps of the Markov chain and is needed in order to estimate the posterior model probabilities from the output of the chain.

Linear regression is an important special case where the marginal likelihood is available in closed form for suitable prior distributions. The calculation of the marginal likelihood essentially reduces to solving a least squares problem. For example, with the default prior proposed by Fernández, Ley, and Steel (2001) the marginal likelihood is given by

$$m\left(\mathbf{y}\middle|\,\mathcal{M}\right) \propto (c+1)^{-(k+1)/2}\left(\frac{c\cdot \mathrm{RSS}+\mathrm{TSS}}{c+1}\right)^{-T/2}, \tag{4.5}$$

where RSS is the residual sum of squares from a least squares fit and TSS the (corrected) total sum of squares and c is a tuning constant. See Appendix A in Chapter 2 for more details.

The simplified version B of Algorithm 4.1 applies in this case, which eases the computational burden considerably, but computational efficiency is still important. In many cases it might be needed to run the chain for 5 million steps or more. Scaling up our benchmark this can take as much as 4 hours when using a standard OLS routine to solve the OLS problem. But this is, as many authors have noted, obviously wasteful since the proposed model is a minor variation on the current model. Algorithms that update the current model rather than redoing all the calculations can reduce the time needed by as much as 85% compared to otherwise optimized algorithms. While promising increased speed, there is a potential cost in the form of loss of numerical accuracy due to accumulated round-off errors. We evaluate the speed and accuracy of a number of algorithms for updating OLS estimates when variables are added to or dropped from a model.

It is obvious that the RJMCMC scheme involves small changes between models, with the move types discussed above. This is also to a large extent the case with the more complex updating schemes evaluated in Section 4.5 and we expect similar savings in computational time. There are clearly similar benefits when calculating the marginal likelihoods by brute force since the models can be enumerated by the binary indicator vector $\boldsymbol{\gamma}$.

4.3 Solving least squares problems

Consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4.6}$$

where $\mathbf{y}$ is a $(T \times 1)$ vector, $\mathbf{X}$ is a $(T \times k)$ matrix of explanatory variables, with $T \geq k$, and $\boldsymbol{\beta}$ is a $(k \times 1)$ vector of unknown parameters. A basic computational problem in regression analysis is to obtain the solution to the normal equations

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}. \tag{4.7}$$

The representation of the solution as $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is mostly a notational or theoretical convenience. In practice this is almost never solved by first computing the inverse and then the product. Instead least squares solvers are based on a factorization $\mathbf{A} = \mathbf{BC}$, where $\mathbf{B}$ and $\mathbf{C}$ are non-singular and easily invertible, for example triangular or orthogonal matrices. The most commonly used factorization is probably the QR factorization, which is considered to have good numerical properties. The Cholesky decomposition is also commonly used, but is in general considered to be more susceptible to round-off errors. Other algorithms that can be used include singular value decomposition and the LU factorization. The latter two are not considered here, because they are difficult to update or are relatively inefficient. We will instead include the Sweep operator in our comparison since it is relatively efficient and eminently suited to calculating successive least squares estimates for models that differ by a few explanatory variables. To our knowledge this is the first time the Sweep operator is considered in this context.

Table 4.1 gives a general impression of the efficiency of these algorithms as least squares solvers. The table gives the leading terms in the number of floating point operations (flop) needed for the different steps in order to solve for $\hat{\boldsymbol{\beta}}$ and calculate the residual sum of squares. It is clear that there is little difference for moderate size problems and that the advantage of the Cholesky decomposition and the Sweep operator increases with T. The increased saving in computations is due to these algorithms operating on $\mathbf{X}'\mathbf{X}$ rather than $\mathbf{X}$. This is also a great advantage when repeatedly solving related least squares problems since $\mathbf{X}'\mathbf{X}$, $\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'\mathbf{y}$ can be precomputed, and the Cholesky decomposition and the Sweep operator become $O\left(k^3\right)$ algorithms, whereas the QR decomposition is $O\left(Tk^2\right)$.[2]

In what follows, we review the algorithms and show how they can be used to update least squares solutions when variables are dropped from or added

[2]The QR decomposition can be applied to $\mathbf{X}'\mathbf{X}$, but this requires about 4 times as much calculations as a Cholesky decomposition.

Table 4.1 Order of floating point operations for OLS

Operation	Cholesky	Sweep	QR (Householder)
$\mathbf{X}'\mathbf{X}, \mathbf{X}'\mathbf{y}, \mathbf{y}'\mathbf{y}$	$Tk^2/2$	$Tk^2/2$	-
Factor	$k^3/3$	-	$2Tk^2 - 2k^3/3$
Calculate $\mathbf{Q}'\mathbf{y}$	-	-	$4Tk - 2k^2$
Sweep	-	$3k^3/2$	-
Solve for $\hat{\boldsymbol{\beta}}$	k^2	-	k^2
Calculate RSS	$4k + k^2$	-	$2(T - k)$
Overall order	$Tk^2/2 + k^3/3$	$Tk^2/2 + 3k^3/2$	$2Tk^2 - 2k^3/3$

to a model.[3] In addition to theoretical predictions about their numerical efficiency based on flop counts, we also evaluate the efficiency and accuracy of the algorithms in scenarios designed to mimic the sequence of computations in MCMC-based variable selection and BMA exercises. It should be noted that other implementations of the algorithms might be more numerically accurate. In particular, we do not use pivoting with the Cholesky and QR decompositions since this makes updating more complicated.

4.3.1 QR decomposition

The QR decomposition operates directly on the $(T \times k)$ matrix

$$\mathbf{X} = \mathbf{QR} = \mathbf{Q}_1\mathbf{R}_1, \tag{4.8}$$

decomposing it into a $(T \times T)$ orthogonal matrix $\mathbf{Q} = \begin{bmatrix}\mathbf{Q}_1 & \mathbf{Q}_2\end{bmatrix}$ and $\mathbf{R} = \begin{bmatrix}\mathbf{R}_1' & \mathbf{O}'\end{bmatrix}'$, where $\mathbf{R}_1$ is a $(k \times k)$ upper triangular matrix and $\mathbf{O}$ a $((T-k) \times k)$ null matrix. The decomposition $\mathbf{X} = \mathbf{Q}_1\mathbf{R}_1$ is a so-called thin factorization. The normal equations reduce to

$$\begin{aligned}\mathbf{R}_1'\mathbf{R}_1\hat{\boldsymbol{\beta}} &= \mathbf{R}_1'\mathbf{Q}_1'\mathbf{y} \\ \mathbf{R}_1\hat{\boldsymbol{\beta}} &= \mathbf{Q}_1'\mathbf{y},\end{aligned} \tag{4.9}$$

which is trivial to solve due to the triangularity of $\mathbf{R}_1$. The residual sum of squares can be obtained by summing the squares of the vector $\mathbf{Q}_2'\mathbf{y}$.

The QR decomposition can alternatively be applied on an augmented matrix

$$\mathbf{A} = \begin{bmatrix}\mathbf{X} & \mathbf{y}\end{bmatrix} = \mathbf{Q}^*\mathbf{R}^*. \tag{4.10}$$

[3] The exposition draws on Golub and van Loan (1996) and the reader is referred to this book for additional background.

$\mathbf{R}_1$ is then the leading $(k \times k)$ submatrix of $\mathbf{R}^*$ and $\mathbf{Q}_1'\mathbf{y}$ can be found in the first k rows of the last $(k+1)$ column of $\mathbf{R}^*$. The square of the last diagonal element of $\mathbf{R}^*$ is the residual sum of squares. In this way the multiplication $\mathbf{Q}'\mathbf{y}$ is not necessary and the calculation time is reduced.

Methods for calculating the QR decomposition include Householder reflections, Givens rotations, and Gram-Schmidt orthogonalization.

QR decomposition by Householder reflections

A Householder reflection is a $(m \times m)$ symmetric, and orthogonal matrix of the form

$$\mathbf{H} = \mathbf{I} - \frac{2}{\mathbf{u}'\mathbf{u}}\mathbf{u}\mathbf{u}', \tag{4.11}$$

where $\mathbf{u}$ is a $(m \times 1)$ Householder vector defined as

$$\mathbf{u} = \mathbf{x} + \operatorname{sgn}(x_1) \|\mathbf{x}\|_2 \mathbf{e}_1, \tag{4.12}$$

where $\mathbf{x}$ is a non-zero $(m \times 1)$ vector, $\mathbf{e}_1$ is the first vector of an identity matrix, $\|\cdot\|_2$ denotes the Euclidean norm, and $\operatorname{sgn}(x_1)$ is the sign of the first element in $\mathbf{x}$.

Using the Householder reflections it is easy to transform a non zero vector

$$\mathbf{x} = \begin{pmatrix} x_1, & x_2, & x_3, & \ldots, & x_m \end{pmatrix}' \tag{4.13}$$

into a vector

$$\tilde{\mathbf{x}}_{\mathbf{H}} = \begin{pmatrix} \tilde{x}_{H1}, & 0, & 0, & \ldots, & 0 \end{pmatrix}', \tag{4.14}$$

thus introducing zeros on a grand scale.

Let $\mathbf{H}_1$ be the $(T \times T)$ Householder matrix that sets the elements below the diagonal of the first column of $\mathbf{X}$ to zero. Denote the resulting matrix as $\mathbf{X}_1 = \mathbf{H}_1\mathbf{X}$. Next consider the second column of $\mathbf{X}_1$. The elements below the diagonal can be set to zero by applying a $((T-1) \times (T-1))$ Householder matrix $\widetilde{\mathbf{H}}_2$ to the last $T-1$ rows or, equivalently, premultiplying $\mathbf{X}_1$ by $\mathbf{H}_2 = \operatorname{diag}\left(1, \widetilde{\mathbf{H}}_2\right)$. Continuing in this fashion and noting that $\mathbf{H}_j$ only operates on the lower right $T-j+1$ submatrix from the previous step, we obtain $\mathbf{R}$ as

$$\mathbf{R} = \mathbf{H}_k\mathbf{H}_{k-1}\ldots\mathbf{H}_2\mathbf{H}_1\mathbf{X}, \tag{4.15}$$

and the orthogonal matrix $\mathbf{Q}$ is given by

$$\mathbf{Q} = \mathbf{H}_1\mathbf{H}_2\ldots\mathbf{H}_{k-1}\mathbf{H}_k. \tag{4.16}$$

Table 4.2 Order of floating point operations for OLS updates

	Add	Drop (average)	Swap
HouseUp	$4Tk - k^2$	$4Tk^2/3 - 2k^3/3$	$4Tk^2/3 - 2k^3/3$
GGSUp	$4Tk$	$3Tk + k^2$	$11Tk + k^2$
Cholesky	$k^3/3 + 3k^2/2$	$k^3/3 - k^2/3$	$k^3/3 + k^2$
Sweep	$3N^2/2$	$3N^2/2$	$3N^2$

In practice $\mathbf{Q}$ is rarely needed, and $\mathbf{Q}'\mathbf{y}$ can be calculated alongside $\mathbf{R}$ as $\mathbf{Q}'\mathbf{y} = \mathbf{H}_k\mathbf{H}_{k-1}\dots\mathbf{H}_2\mathbf{H}_1\mathbf{y}$, which is an $O(Tk)$ operation, compared to accumulating $\mathbf{Q}$ and explicitly calculating $\mathbf{Q}'\mathbf{y}$, which are $O\left(T^2k\right)$ and $O\left(T^2\right)$, respectively. Our algorithm, denoted as House, uses the former approach. It also factors $\mathbf{X}$ rather than $\begin{bmatrix}\mathbf{X} & \mathbf{y}\end{bmatrix}$, since the latter would complicate the updating procedure discussed next. The Householder routines are based on the LINPACK subroutines DQRDC and DQRSL. See Dongarra, Bunch, Moler, and Stewart (1979) for details.

Updating QR decomposition using Householder reflections

The vectors $\mathbf{u}$ needed to form $\mathbf{H}_j$ can be stored efficiently[4] and the Householder reflections are easily recreated. This, together with the fact that a Householder reflections requires much less work than a full matrix product, makes a simple algorithm for adding or deleting columns from $\mathbf{X}$ available.

A column $\mathbf{z}$ is simply added to $\mathbf{X}$ by applying the Householder reflections $\mathbf{H}_1, \dots, \mathbf{H}_k$ in turn, which forms $\mathbf{Q}'\mathbf{z}$. Next, elements $k+2$ to T of $\mathbf{Q}'\mathbf{z}$ are zeroed with $\mathbf{H}_{k+1}$, and the first $k+1$ elements of the result are appended to $\mathbf{R}_1$ as column $k+1$ in the upper triangle of $\mathbf{X}$. The vector $\mathbf{u}$ defining $\mathbf{H}_{k+1}$ is stored below the diagonal and the previously stored $\mathbf{Q}'\mathbf{y}$ is premultiplied with $\mathbf{H}_{k+1}$.

To remove column j, $1 \le j \le k$, from $\mathbf{X}$ the reflections for columns $i = k, k-1, \dots, j$ are first undone by premultiplying the previously saved $\mathbf{Q}'\mathbf{y}$ and columns $i+1$ to k of $\mathbf{R}_1$ with $\mathbf{H}_i$ and restoring the elements on or below the diagonal of column i of $\mathbf{X}$ from the saved $\mathbf{u}$ if $i > j$. Columns $j+1$ to k are then shifted left and new Householder matrices $\mathbf{H}_j, \dots, \mathbf{H}_{k-1}$ are used to zero the elements below the diagonal and update $\mathbf{Q}'\mathbf{y}$.

A swap of variables is implemented by first removing the column in question and then adding the new variable as the last column k.

[4] In practice, $\mathbf{R}_1$ overwrites the upper triangle of $\mathbf{X}$, the last $T-j$ elements of $\mathbf{u}$ are stored below the diagonal in column j and the first element in an auxiliary vector.

The leading terms of the flop count for this algorithm, denoted as HouseUp, is given in Table 4.2. While this algorithm provides an efficient way of adding a variable to the model it is quite expensive to remove a variable. A much more efficient way of dropping variables is available using Givens rotations, which can be used to selectively zero elements in a matrix.

QR decomposition by Givens rotations

The Givens rotation method, also known as Jacobi rotations, is defined by the $(m \times m)$ matrix $\mathbf{G}$

$$\mathbf{G}(i,k,\phi) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \begin{matrix} \\ \\ i \\ \\ k \\ \\ \\ \end{matrix}, \qquad (4.17)$$
$$\begin{matrix} & & i & & k & & \end{matrix}$$

where $c = \cos(\phi)$, and $s = \sin(\phi)$ for some angle ϕ. Premultiplication by $\mathbf{G}(i,k,\phi)'$ amounts to a counterclockwise rotation of ϕ radians in the (i,k) coordinate plane. By choosing ϕ to satisfy

$$\cos(\phi) = \frac{x_i}{\|\mathbf{x}\|_2} = \frac{x_i}{\sqrt{x_i^2 + x_k^2}}, \qquad (4.18)$$

$$\sin(\phi) = \frac{x_k}{\|\mathbf{x}\|_2} = \frac{-x_k}{\sqrt{x_i^2 + x_k^2}}, \qquad (4.19)$$

for a $(m \times 1)$ vector $\mathbf{x}$, the elements of $\tilde{\mathbf{x}} = \mathbf{G}(i,k,\phi)'\mathbf{x}$ can be obtained as

$$\tilde{x}_j = \begin{cases} cx_i - sx_k & j = i \\ 0 & j = k \\ x_j & j \neq i,k. \end{cases} \qquad (4.20)$$

That is, element k of $\mathbf{x}$ is zeroed by applying the Givens rotation matrix. In practice the rotations are typically applied to rows i and $i+1$, and we refer to a Givens rotation matrix that zeroes element $i+1$ of column j in a matrix as $\mathbf{G}_{i,j}$. Note that the product $\mathbf{G}'_{i,j}\mathbf{X}$ only operates on rows i and $i+1$ of $\mathbf{X}$.

The QR factorization of $\mathbf{X}$ can be calculated by applying a series of Givens rotations. First, the $T-1$ elements below the diagonal of column 1 are zeroed with $\mathbf{G}_{T-1,1}$ through $\mathbf{G}_{1,1}$ yielding

$$\mathbf{X}_1 = \mathbf{G}'_{1,1} \dots \mathbf{G}'_{T-1,1}\mathbf{X}. \tag{4.21}$$

The $T-2$ elements below the diagonal in the second column of the matrix $\mathbf{X}_1$ are zeroed by applying $\mathbf{G}_{T-1,2}$ through $\mathbf{G}_{2,2}$ and so on for the remaining columns. This yields

$$\mathbf{R} = \mathbf{G}'_{k,k} \dots \mathbf{G}'_{T-1,k}\mathbf{G}'_{k-1,k-1} \dots \mathbf{G}'_{T-1,k-1}\mathbf{G}'_{k-2,k-2} \dots \mathbf{G}'_{T-1,1}\mathbf{X} \tag{4.22}$$

and

$$\mathbf{Q} = \mathbf{G}_{T-1,1}\mathbf{G}_{T-2,1} \dots \mathbf{G}_{1,1}\mathbf{G}_{T-1,2} \dots \mathbf{G}_{2,2} \dots \mathbf{G}_{k,k}. \tag{4.23}$$

This factorization algorithm is $O\left(k^2\left(3T-k\right)\right)$ and less efficient than using Householder matrices. A variation on Givens rotations known as Fast Givens or Modified Givens is of the same order as using Householder matrices, but the maintenance of a vector of scaling constants as well as occasional rescaling is required. Hanson and Hopkins (2004) evaluate the relative performance of Givens and Modified Givens rotations, and find that it is impossible to predict which will perform better on a given hardware/compiler combination.

Deleting a column using Givens rotations

Assume that we have the thin QR factorization

$$\mathbf{X} = \mathbf{Q}_1\mathbf{R}_1, \tag{4.24}$$

and delete the column i in the matrix $\mathbf{X}$. Denote this new matrix as $\widetilde{\mathbf{X}}$. Deleting the i-th column of $\mathbf{R}_1$ gives

$$\widetilde{\mathbf{R}}_1 = \mathbf{Q}'_1\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{R}_{1,1} & \mathbf{T}_1 \\ \mathbf{0} & \mathbf{T}_2 \end{bmatrix}, \tag{4.25}$$

where $\mathbf{R}_{1,1}$ is a $((i-1)\times(i-1))$ upper triangular matrix, $\mathbf{T}_1$ is a rectangular $((i-1)\times(k-i))$ matrix and $\mathbf{T}_2$ is a $(k-i+1)\times(k-i)$ upper Hessenberg matrix[5]. The unwanted sub-diagonal elements $\tilde{r}_{i+1,i}, \dots, \tilde{r}_{k,k-1}$, present in $\mathbf{T}_2$, can be zeroed by a applying a sequence of Givens rotations to $\widetilde{\mathbf{R}}_1$ and form the new

$$\mathbf{R}_1 = \mathbf{G}'_{k-1,k-1}\mathbf{G}'_{k-2,k-2} \cdots \mathbf{G}'_{i,i}\widetilde{\mathbf{R}}_1.$$

[5]In an upper Hessenberg matrix all elements below the first subdiagonal are equal to zero.

$\mathbf{Q}'\mathbf{y}$ is updated similarly by performing the same sequence of Givens rotations and $\mathbf{Q}$ can, if needed, be accumulated by postmultiplying with the Givens matrices.

These are $O\left(3Tk + k^2\right)$ operations and considerably more efficient than the Householder-based algorithm for deleting a column. This algorithm for deleting columns is, however, difficult to combine with the quite efficient approach of adding columns using Householder reflections.

Adding a column using Givens rotations

The standard textbook treatment of algorithms for adding a column to $\mathbf{X}$ assumes that the full decomposition $\mathbf{X} = \mathbf{QR}$ is available. Inserting a column $\mathbf{z}$ at position j of $\mathbf{X}$ gives

$$\mathbf{X}^* = \left(\mathbf{x}_1, \quad \ldots, \quad \mathbf{x}_{j-1}, \quad \mathbf{z}, \quad \mathbf{x}_j, \quad \ldots, \quad \mathbf{x}_k\right). \tag{4.26}$$

Premultiplying by $\mathbf{Q}'$ produces

$$\mathbf{Q}'\mathbf{X}^* = \left(\mathbf{Q}'\mathbf{x}_1, \quad \ldots, \quad \mathbf{Q}'\mathbf{x}_{j-1}, \quad \mathbf{Q}'\mathbf{z}, \quad \mathbf{Q}'\mathbf{x}_j, \quad \ldots, \quad \mathbf{Q}'\mathbf{x}_N\right) = \mathbf{R}^*, \tag{4.27}$$

which is $\mathbf{R}$ with $\mathbf{Q}'\mathbf{z}$ inserted at position j. $\mathbf{R}^*$ is upper triangular except for the j-th column. Applying Givens rotations to zero the elements below the diagonal in column j reduces $\mathbf{R}^*$ to upper Hessenberg form. The unwanted elements $r^*_{j+2,j+1}, \ldots, r^*_{k+2,k+1}$ are then zeroed with a series of Givens rotations. $\mathbf{Q}'\mathbf{y}$ (and $\mathbf{Q}$) can be updated in parallel using the same sequence of Givens rotations. This algorithm is efficient if it is cheap to form $\mathbf{Q}'\mathbf{z}$.

There is a substantial penalty for using a Householder reflection to zero the elements below the diagonal in column j since this turns *all* the elements below the diagonal in columns $j+1, \ldots, k+1$ into non-zero elements, which in turn will have to be zeroed. There is, of course, no penalty if $\mathbf{z}$ is appended to $\mathbf{X}$ as column $k+1$ rather than inserting it as an interior column $j \leq k$. The Householder updating algorithm takes advantage of this and forms $\mathbf{Q}'\mathbf{z}$ directly by applying a small number of Householder reflections, rather than maintaining the $(T \times T)$ matrix $\mathbf{Q}$ and calculating $\mathbf{Q}'\mathbf{z}$ as a matrix product. A similar level of efficiency can be achieved with Givens rotations if a limited number of rotations are needed. However, the use of these algorithms in the settings outlined in the previous section causes the number of rotations to grow without bound with the number of updates of the least squares solution. We are thus reduced to maintaining a full $\mathbf{Q}$ matrix and calculating the matrix product $\mathbf{Q}'\mathbf{z}$, which turns the step of adding a column into an $O\left(T^2k\right)$ operation.

There is thus a trade off. The Householder updating algorithm limits the number of reflections by undoing them in the delete step, which makes an effective add step possible. The Givens-based delete step is much more efficient than the Householder delete step, but the number of rotations grows without bounds, which leads to an inefficient add step.

Gram-Schmidt orthogonalization

Gram-Schmidt orthogonalization is essentially an algorithm for appending columns, which has several advantages in addition to being possible to combine with the efficient Givens-based delete step. It is sufficient to maintain the $(T \times k)$ matrix $\mathbf{Q}_1$ instead of a full $\mathbf{Q}$ since only $\mathbf{Q}_1'\mathbf{z}$ and not $\mathbf{Q}'\mathbf{z}$ is needed for the update. The use of Gram-Schmidt orthogonalization as a tool for adding columns is due to Daniel, Gragg, Kaufman, and Stewart (1976), and our implementation is based on the Fortran code in Reichel and Gragg (1990).

Suppose we have the thin QR factorization $\mathbf{X} = \mathbf{Q}_1\mathbf{R}_1$ and add a column $\mathbf{z}$ to $\mathbf{X}$ to form (4.26). We then have

$$\mathbf{X}^* = \begin{bmatrix}\mathbf{Q}_1 & \mathbf{z}\end{bmatrix} \begin{bmatrix}\mathbf{R}_{1,1} & \mathbf{0} & \mathbf{R}_{1,2} \\ \mathbf{0}' & 1 & \mathbf{0}'\end{bmatrix}. \tag{4.28}$$

Next, use the Gram-Schmidt orthogonalization procedure to find a $(T \times 1)$ vector $\mathbf{q}$, a $(k \times 1)$ vector $\mathbf{r}$, and a scalar ρ that satisfy

$$\begin{bmatrix}\mathbf{Q}_1 & \mathbf{z}\end{bmatrix} = \begin{bmatrix}\mathbf{Q}_1 & \mathbf{q}\end{bmatrix} \begin{bmatrix}\mathbf{I} & \mathbf{r} \\ \mathbf{0}' & \rho\end{bmatrix}, \tag{4.29}$$

under the conditions that $\mathbf{Q}_1'\mathbf{q} = \mathbf{0}$ and $\mathbf{q}'\mathbf{q} = 1$. The last column on the right hand side of (4.29) equals $\mathbf{z} = \mathbf{Q}_1\mathbf{r} + \rho\mathbf{q}$, which premultiplied by $\mathbf{Q}_1'$, gives the vector

$$\mathbf{r} = \mathbf{Q}_1'\mathbf{z}. \tag{4.30}$$

Setting $\mathbf{z}^* = \mathbf{z} - \mathbf{Q}_1\mathbf{r} = \rho\mathbf{q}$ and using $\mathbf{q}'\mathbf{q} = 1$ we have

$$\rho = \mathbf{z}^{*\prime}\mathbf{z}^*, \tag{4.31}$$

$$\mathbf{q} = \frac{1}{\rho}\mathbf{z}^*. \tag{4.32}$$

Combining (4.28) and (4.29) yields

$$\mathbf{X}^* = \begin{bmatrix}\mathbf{Q}_1 & \mathbf{q}\end{bmatrix} \begin{bmatrix}\mathbf{R}_{1,1} & \mathbf{r} & \mathbf{R}_{1,2} \\ \mathbf{0}' & \rho & \mathbf{0}'\end{bmatrix} = \mathbf{Q}_1^*\mathbf{R}_1^* \tag{4.33}$$

and $\mathbf{R}_1^*$ can be reduced to upper triangular form by a series of Givens rotations. We always append $\mathbf{z}$ as the last, k+1, column of $\mathbf{X}$, thus no rotations are needed and the algorithm is $O(4Tk)$ in this case.

It is possible that the Gram-Schmidt procedure (4.30) yields a solution $\mathbf{q}$ that is not orthogonal to $\mathbf{Q}$. In this case reorthogonalization can be used, by simply setting $\mathbf{z} = \mathbf{z}^*$ and applying (4.30) - (4.32) again. Daniel, Gragg, Kaufman, and Stewart (1976) provide an error analysis indicating that the Gram-Schmidt procedure with reorthogonalization has good numerical properties.

Our algorithm, abbreviated as GGSUp, combines the Givens rotation-based procedure for deleting columns and the Gram-Schmidt procedure for adding columns. A swap of columns is achieved by first removing the column in question and then adding the new column as the last column k. Since only $\mathbf{Q}_1$ is maintained the residuals must be calculated explicitly after solving the least squares problem in order to obtain the residual sum of squares. In addition, $\mathbf{Q}_1'\mathbf{y}$ is calculated directly as a matrix product rather than continuously updating the product.

4.3.2 Cholesky decomposition

The Cholesky factorizations decomposes a $(k \times k)$ symmetric, positive definite matrix $\mathbf{A} = \mathbf{X}'\mathbf{X}$ into a product of a lower triangular matrix $\mathbf{L}$ and its transpose $\mathbf{L}'$,

$$\mathbf{A} = \mathbf{X}'\mathbf{X} = \mathbf{L}\mathbf{L}'. \tag{4.34}$$

The solution to the normal equations is then obtained by solving two triangular equations systems

$$\mathbf{L}\boldsymbol{\xi} = \mathbf{X}'\mathbf{y}, \tag{4.35}$$

$$\mathbf{L}'\boldsymbol{\beta} = \boldsymbol{\xi}. \tag{4.36}$$

The residual sum of squares is calculated using the identity $\text{RSS} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$, where $\mathbf{y}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$ is obtained as a by-product of solving the equations system (4.36).

As already noted, a great benefit of the Cholesky decomposition is that $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$ can be precomputed for the full $(T \times N)$ data matrix and the algorithm is then independent of T, when repeatedly solving least squares problems.

The Cholesky routines are based on the Press, Teukolsky, Vetterling, and Flannery (1992) subroutines CHOLDC and CHOLSL.

Updating the Cholesky decomposition

Assume that the Cholesky decomposition (4.34) has been obtained. When a variable is added to a model, the modified matrix and its decomposition are given as

$$\mathbf{A}^* = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{a}_j & \mathbf{A}_{12} \\ \mathbf{a}_j' & a_{jj} & \boldsymbol{\alpha}_j' \\ \mathbf{A}_{21} & \boldsymbol{\alpha}_j & \mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{L}_{11} & 0 & 0 \\ \mathbf{l}_j' & l_{jj} & 0 \\ \mathbf{L}_{21} & \boldsymbol{\lambda}_j & \mathbf{L}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{L}_{11}' & \mathbf{l}_j & \mathbf{L}_{21}' \\ 0 & l_{jj} & \boldsymbol{\lambda}_j' \\ 0 & 0 & \mathbf{L}_{22}' \end{bmatrix}. \quad (4.37)$$

The updated decomposition is found first by solving

$$\mathbf{l}_j = \mathbf{L}_{11}^{-1}\mathbf{a}_j, \quad (4.38)$$

$$l_{jj} = \sqrt{a_{jj} - \mathbf{l}_j'\mathbf{l}_j}, \quad (4.39)$$

$$\boldsymbol{\lambda}_j = \frac{1}{l_{jj}}\left(\boldsymbol{\alpha}_j - \mathbf{L}_{21}l_{jj}\right), \quad (4.40)$$

and then decomposing the lower right corner $\mathbf{A}_{22}$ as usual, i.e. by using expressions (4.39)-(4.40). This is the inner-product version of the factorization. See Golub and van Loan (1996) for the outer-product alternative. Deleting a column j from a model is equivalent to decomposing the lower right corner $\mathbf{A}_{22}$. When swapping two variables, the variable added to the model simply takes the position of the variable being removed and the update as described above is carried out. When only the Add step is implemented, the corresponding products are always appended as the last row and column, respectively.

4.3.3 Reversible sweep operator

The sweep operator was designed by Beaton (1964) as a tool for inverting symmetric matrices. As Goodnight (1979) points out:

> "The importance of the sweep operator in statistical computing is not so much that it is an inversion technique, but rather that it is a conceptual tool to understanding the least squares process."

Goodnight also provides a review of its use, including: ordinary least squares, two-stage and three-stage least squares, nonlinear least squares, multivariate analysis of variance, regressions by leaps and bounds, stepwise regression, and partial correlation.

The details of the Sweep operator are given as Algorithm 4.2. Applying the sweep operator to the data matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{y} \\ \mathbf{y}'\mathbf{X} & \mathbf{y}'\mathbf{y} \end{bmatrix} \quad (4.41)$$

Algorithm 4.2 The reversible upper triangular sweep operator

Define an auxiliary vector $\mathbf{v} = (v_1, \ldots, v_N) = (1, 1, \ldots, 1)$. If a k-th variable is swept then $v_k = -v_k$. Choose a tolerance $\delta > 0$. If $a_{kk} > \delta$, or if $v_k = -1$, then

1. Let $\eta = a_{kk}$.
2. Set
$$\zeta = \begin{cases} a_{kj}\,\eta^{-1}, & \text{for each } j = 1, 2, \ldots, k-1, \\ v_j\, v_k\, a_{jk}\,\eta^{-1}, & \text{for each } j = k+1, \ldots, N. \end{cases}$$
3. Set
$$\xi = \begin{cases} a_{ik}, & \text{for each } i = j, j+1 \ldots, k-1, \\ v_i\, v_k\, a_{ki}, & \text{for each } i = k+1, \ldots, N. \end{cases}$$
4. Calculate
$$a_{ij} = a_{ij} - \zeta\xi.$$
5. Set
$$\begin{aligned} a_{kj} &= -a_{kj}\,\eta^{-1}, && \text{for each } j = 1, \ldots, k, \\ a_{ik} &= a_{ik}\,\eta^{-1}, && \text{for each } i = k, \ldots, N. \end{aligned}$$
6. Finally, set
$$\begin{aligned} a_{kk} &= \eta^{-1}, \\ v_k &= -v_k. \end{aligned}$$

---

results in a matrix

$$\begin{aligned} \mathbf{A}^S &= \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-1} & (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ -\mathbf{y}'\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-1} & \hat{\boldsymbol{\beta}} \\ -\hat{\boldsymbol{\beta}}' & \text{RSS} \end{bmatrix}, \end{aligned} \tag{4.42}$$

since $\text{RSS} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$.

The estimates of $\boldsymbol{\beta}$, the residual sum of squares and the inverse of $\mathbf{X}'\mathbf{X}$ can therefore be obtained by using simple computational operations. Note that $\mathbf{y}$ and the variables in $\mathbf{X}$ used to form the matrix $\mathbf{A}$ are assumed to be centred. The matrix $\mathbf{X}'\mathbf{X}$ is thus the sums of squares and cross product

matrix (SSCP). The $((N+1)\times(N+1))$ matrix $\mathbf{A}$ is also a SSCP matrix, with element $a_{N+1,N+1}$ being the corrected total sum of squares.

The time taken to perform the sweep operations may be reduced by taking into account the symmetry of the matrix (4.42). The elements below the main diagonal may be constructed as

$$a_{ij} = \begin{cases} -a_{ji} & \text{if variable } i \text{ or variable } j \text{ has been swept,} \\ a_{ji} & \text{otherwise.} \end{cases} \tag{4.43}$$

Two properties of the sweep operator are extremely useful when repeatedly solving closely related least squares problems:

1. It is not necessary to perform the sweep of the pivots in any particular form to obtain the inverse,

2. Sweep operations are reversible, that is, the status of a matrix that existed prior to a sweep on a particular column can be reestablished by repeating a sweep operation on the same column. The reverse sweeps do not need to be performed in the same order in which the forwards steps were originally performed.

In this context the matrix (4.41) is formed using the full $(T\times N)$ data matrix and the algorithm is $O\left(N^2\right)$ and independent of both T and k.

To guarantee the reversibility of the Algorithm 4.2 in the case when dependencies between explanatory variables exist, Goodnight (1979) suggests to sweep the diagonal element a_{kk} only if its value is greater than some tolerance, δ, defined as

$$\delta = \begin{cases} \delta_0 \cdot \text{TSS}, & \text{if TSS} > 0 \\ \delta_0, & \text{otherwise,} \end{cases} \tag{4.44}$$

with $\delta_0 \in \left[10^{-8}, 10^{-12}\right]$. In this way the critical information that defines variable's dependency is preserved.

4.4 Numerical accuracy and computational efficiency

This section examines the numerical accuracy and efficiency of the algorithms in a setting designed to closely resemble the sequence of computations encountered in Bayesian variable selection and model averaging exercises. To

summarize, the algorithms[6] included in the evaluation are:

OLS	the IMSL subroutine DRLSE
House	QR decomposition using Householder reflections
HouseUp	Updating the QR decomposition using Householder reflections
GGSUp	Updating the QR decomposition using Givens rotations and Gram-Schmidt orthogonalization
Chol	Cholesky decomposition
CholUp	Updating the Cholesky decomposition
Sweep	Updating the least squares solution using the Sweep operator

While Tables 4.1 and 4.2 give a good indication of the speed, one can expect from the different algorithms a rough flop count is only part of the story. Other issues such as memory access patterns are at least as important and much more difficult to evaluate using theoretical arguments. An empirical evaluation of the relative performance is thus needed.

Numerical accuracy is, of course, of utmost importance, it is of no use to have a fast algorithm if it produces misleading results. When evaluating the numerical accuracy we use the IMSL (Visual Numerics, Inc. (1994)) subroutine DRLSE for OLS as the benchmark. This routine performs an orthogonal reduction of the matrix $\begin{bmatrix}\mathbf{X} & \mathbf{y}\end{bmatrix}$ to upper triangular form using Fast Givens transformations.

We generate datasets with $N = 25, 50$ and 100 variables for three different sample sizes $T = 100, 250$ and 400.

The design of the experiment is based on Fernández, Ley, and Steel (2001). A matrix of 15 predictors $\mathbf{X}_{(T\times 15)}$ is generated, where the first 10 random variables, $\mathbf{x}_1, \ldots, \mathbf{x}_{10}$, are i.i.d. standard normal and the additional five variables are constructed according to

$$(\mathbf{x}_{11}, \ldots, \mathbf{x}_{15}) = (\mathbf{x}_1, \ldots, \mathbf{x}_5) \begin{pmatrix} 0.3 & 0.5 & 0.7 & 0.9 & 1.1 \end{pmatrix}' \boldsymbol{\iota} + \mathbf{E}, \qquad (4.45)$$

where $\boldsymbol{\iota}$ is a (1×5) vector of ones and $\mathbf{E}$ is a $(T \times 5)$ matrix of i.i.d. standard normals. This produces a correlation between the first five and the last five predictors. The dependent variable is generated as

$$y_t = 4 + 2x_{1,t} - x_{5,t} + 1.5x_{7,t} + x_{11,t} + 0.5x_{13,t} + \sigma\varepsilon_t, \qquad (4.46)$$

where the disturbances ε_t are i.i.d. standard normal and $\sigma = 2.5$. To obtain the desired dataset size the remaining $N - 15$ variables are generated as i.i.d. standard normal.

[6] All algorithms are programmed in Fortran 95 and compiled with Compaq Visual Fortran 6.6C using default compiler settings.

In all cases the explanatory variables are centred and standardized prior to running the experiments. For each dataset we attempt to approximate the behaviour of a RJMCMC scheme like Algorithm 4.1 when the true model size is $k = 5, 10, 15$ or 20 explanatory variables. A specific average model size over the pseudo-MCMC run is achieved by controlling the manner in which the variables are added to or removed from the model. This process starts at a selected model size k, and follows a sequence of steps: swap, add, swap, drop, swap, drop, swap, add, swap, add, swap, drop, etc. For the smallest model size this produces the following sequence of model sizes: 5,6,6,5,5,4,4,5,5,6,6,5, ... In this manner 25% of the MCMC steps add a variable, 25% drop a variable and remaining 50% are swapping a variable. The 'Markov chain' is run for 50 000 steps and all the algorithms considered visit the same set of models in the same order.

4.4.1 Results

Speed

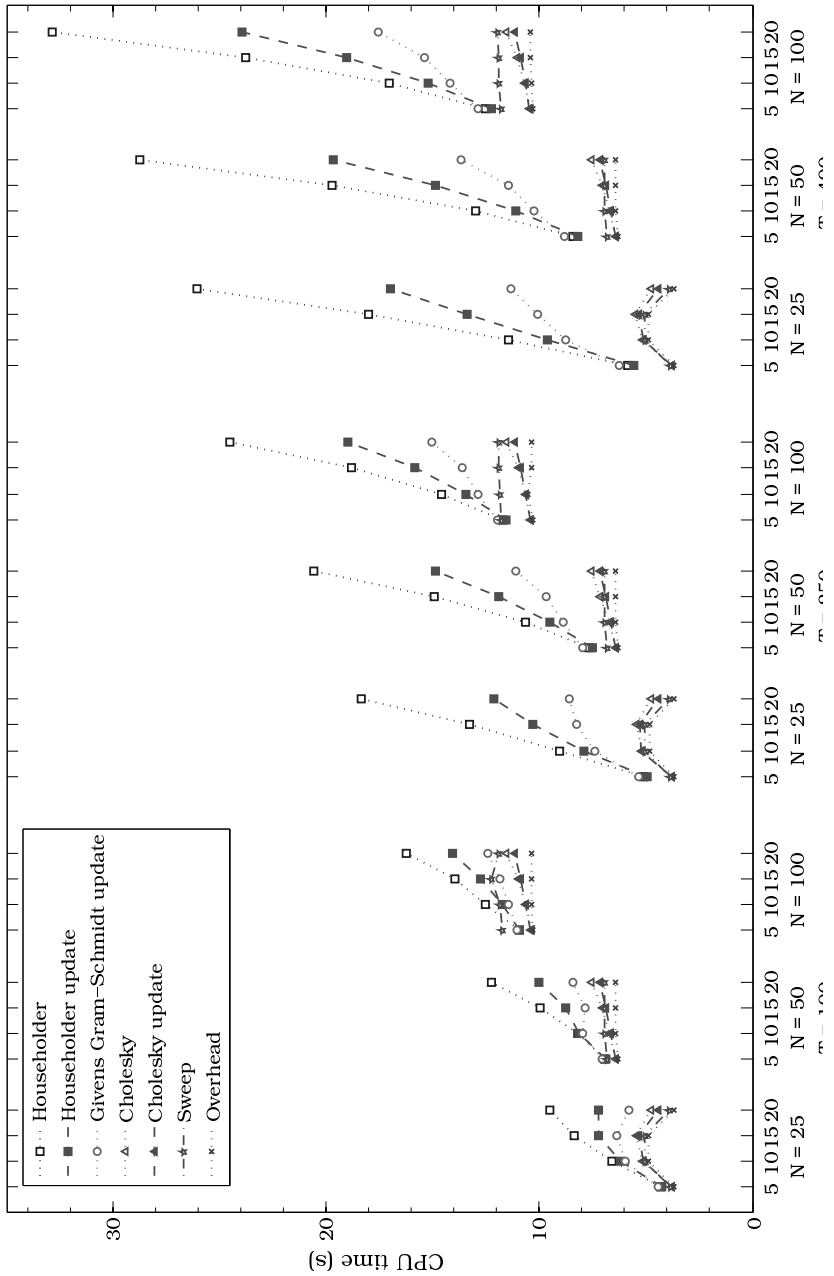
For each combination of the number of potential explanatory variables, N, number of observations, T, and model size, k, the timing experiments were run 5 times and the average CPU time calculated.[7]

The timings include overhead for generating the proposed models and keeping track of the number of visits to a particular model and are thus representative of the actual time requirements for a RJMCMC algorithm. In order to isolate the contribution of the least squares algorithms we also report on the CPU time for the overhead. It should be noted that the time needed to solve the least squares problem is only a small fraction of the total time for the fastest algorithms.

Tables B.1 - B.3 report the times for all the algorithms. Figure 4.1 shows the CPU time in seconds for all algorithms except OLS, which by far is the slowest algorithm for all combinations of k, N and T. It is not surprising that OLS is the slowest algorithm, since it is designed for general use with checks of the data, treatment of missing values etc. that is absent from the other algorithms. In general, increasing the dataset size N has the same effect on all algorithms. Except for the Sweep operator the increase in CPU time can be traced to the increased overhead that is due to the larger number of possible models, and that more distinct models are proposed (and visited) with larger

[7]The experiments were run on an IBM desktop PC running Windows XP SP2 with a Pentium 4 530 processor (3 GHz), 800 MHz memory buss, 1 MB L2 cache and Intel 915G chipset.

Figure 4.1 CPU time (approximation in seconds) for 50 000 steps of a Markov chain for different combinations of k, N and T.



N. For the Sweep operator we see a larger increase in the CPU time than for the other algorithms, reflecting the fact that it always operates on a $(N \times N)$ matrix irrespective of the current model size.

Increasing model size does not influence the Sweep operator and only slightly the Cholesky decomposition and Cholesky update. The concave form for the dataset size $N = 25$, showing dependence on k for these 3 methods, is caused by the fact that there are fewer models in the model set for $k = 5$ and $k = 20$ and thus less overhead than for the remaining two model sizes.

The House, the HouseUp, and GGSUp algorithms all depend on k with House the most sensitive and GGSUp the least affected. This is in rough agreement with the predictions from Tables 4.1 and 4.2.

The Cholesky decomposition, its update, and the Sweep operator are naturally not affected by the sample size T since they operate on the matrix $\mathbf{X}'\mathbf{X}$. Of the remaining algorithms is GGSUp the least affected followed by HouseUp and House. Again this is as predicted by the flop counts in Tables 4.1 and 4.2.

Clearly the Sweep operator, Cholesky, and the Cholesky update algorithms are the most efficient. The Sweep operator is preferred if the number of variables, N, is not too large relative to the average model size, k. The Cholesky update is faster than Cholesky except for the smallest model size, $k = 5$.

Accuracy

For each of the algorithms the values of the estimated beta parameters and the value of the residual sum of squares (RSS) are recorded at every $s = 100$ steps, giving in total 500 control points. Tables 4.3 - 4.4 report the average number of correct significant digits relative to the benchmark algorithm, OLS, for RSS and $\hat{\boldsymbol{\beta}}$, based on the 500 control points. For $\hat{\boldsymbol{\beta}}$ this is an average over both the number of parameters in the model at a particular control point and the control points. That is, the average number of correct digits for $\hat{\boldsymbol{\beta}}$ is given by

$$\zeta_{\hat{\boldsymbol{\beta}}} = \frac{1}{500} \sum_{j=1}^{500} \frac{1}{k_j} \sum_{i=1}^{k_j} \left| \log \left| \frac{\hat{\beta}_i - \hat{\beta}_i^{OLS}}{\hat{\beta}_i^{OLS}} \right| \right|. \tag{4.47}$$

With the increasing sample size T, the number of correct digits is decreasing for all algorithms. The algorithm that is closest to the benchmark is the Cholesky decomposition and its update. The Givens Gram-Schmidt algorithm, however, has the highest number of correct digits for the model parameters. The worst performing algorithm is the Sweep operator with a difference of around one digit.

Table 4.3 Average number of correct significant digits for RSS.

N	k	House	HouseUp	GGSUp	Chol	CholUp	Sweep
T = 100							
	5	15.58	15.59	15.57	15.69	15.69	14.78
25	*10*	15.57	15.57	15.55	15.66	15.66	14.30
	15	15.56	15.58	15.58	15.66	15.66	14.27
	20	15.59	15.55	15.56	15.68	15.67	13.98
	5	15.57	15.56	15.58	15.67	15.67	14.14
50	*10*	15.58	15.57	15.53	15.68	15.68	14.73
	15	15.56	15.55	15.58	15.67	15.68	14.34
	20	15.57	15.54	15.58	15.66	15.66	14.06
	5	15.57	15.56	15.56	15.67	15.67	14.27
100	*10*	15.57	15.56	15.55	15.67	15.67	14.66
	15	15.56	15.56	15.56	15.65	15.65	13.94
	20	15.57	15.56	15.56	15.64	15.63	14.41
T = 250							
	5	15.49	15.49	15.49	15.46	15.46	14.63
25	*10*	15.47	15.47	15.46	15.39	15.39	14.27
	15	15.48	15.44	15.47	15.39	15.39	14.08
	20	15.49	15.46	15.48	15.43	15.43	14.45
	5	15.50	15.45	15.46	15.43	15.42	14.40
50	*10*	15.46	15.45	15.44	15.40	15.40	14.32
	15	15.46	15.51	15.45	15.37	15.37	14.73
	20	15.49	15.49	15.45	15.41	15.41	14.50
	5	15.48	15.51	15.47	15.42	15.42	14.84
100	*10*	15.49	15.49	15.45	15.42	15.42	14.33
	15	15.45	15.47	15.45	15.41	15.41	14.07
	20	15.48	15.47	15.46	15.40	15.41	14.54
T = 400							
	5	15.38	15.35	15.34	15.49	15.49	14.13
25	*10*	15.33	15.35	15.33	15.47	15.47	14.76
	15	15.36	15.36	15.35	15.49	15.49	13.99
	20	15.33	15.33	15.32	15.44	15.44	14.72
	5	15.31	15.33	15.34	15.43	15.43	14.77
50	*10*	15.34	15.34	15.33	15.46	15.46	14.23
	15	15.35	15.33	15.37	15.47	15.47	14.18
	20	15.34	15.37	15.33	15.49	15.50	14.74
	5	15.32	15.32	15.33	15.47	15.47	14.49
100	*10*	15.34	15.32	15.37	15.46	15.46	14.30
	15	15.34	15.31	15.32	15.42	15.43	14.17
	20	15.37	15.35	15.34	15.49	15.48	14.62
Average		15.46	15.46	15.45	15.51	15.51	14.39

Table 4.4 Average number of correct significant digits for estimated model parameters.

N	k	House	HouseUp	GGSUp	Chol	CholUp	Sweep
T = 100							
	5	15.05	14.99	15.13	15.09	15.11	14.12
25	*10*	14.98	14.85	15.05	15.00	15.00	13.99
	15	14.94	14.73	14.99	14.88	14.88	14.01
	20	14.88	14.64	14.94	14.78	14.77	13.95
	5	15.06	15.02	15.12	15.13	15.13	14.02
50	*10*	14.99	14.91	15.07	15.07	15.08	13.96
	15	14.95	14.79	15.01	15.01	15.01	13.90
	20	14.92	14.71	14.97	14.96	14.96	13.94
	5	15.06	15.00	15.12	15.12	15.13	14.08
100	*10*	15.02	14.93	15.08	15.08	15.09	13.94
	15	14.97	14.81	15.01	15.02	15.02	13.89
	20	14.95	14.73	14.97	14.98	14.98	13.87
T = 250							
	5	14.77	14.73	14.83	14.83	14.83	14.06
25	*10*	14.70	14.60	14.79	14.75	14.74	14.00
	15	14.70	14.53	14.78	14.70	14.69	13.98
	20	14.68	14.43	14.76	14.60	14.59	13.92
	5	14.89	14.83	14.94	14.93	14.93	14.16
50	*10*	14.85	14.77	14.92	14.92	14.92	14.13
	15	14.83	14.72	14.89	14.87	14.87	14.02
	20	14.82	14.66	14.89	14.85	14.85	14.06
	5	14.89	14.87	14.97	14.97	14.97	14.19
100	*10*	14.87	14.81	14.93	14.94	14.94	14.11
	15	14.85	14.75	14.91	14.91	14.91	14.05
	20	14.83	14.70	14.90	14.90	14.90	14.03
T = 400							
	5	14.80	14.78	14.88	14.84	14.84	14.06
25	*10*	14.73	14.64	14.83	14.75	14.75	13.98
	15	14.70	14.57	14.80	14.70	14.69	14.07
	20	14.71	14.49	14.81	14.64	14.62	14.03
	5	14.84	14.81	14.90	14.88	14.88	14.10
50	*10*	14.75	14.69	14.85	14.83	14.83	14.08
	15	14.72	14.62	14.81	14.79	14.79	14.00
	20	14.72	14.58	14.80	14.77	14.77	14.02
	5	14.82	14.82	14.88	14.86	14.86	14.16
100	*10*	14.80	14.74	14.87	14.86	14.85	14.10
	15	14.77	14.69	14.86	14.83	14.84	14.08
	20	14.76	14.65	14.83	14.81	14.81	14.01
Average		14.85	14.74	14.92	14.88	14.88	14.03

Figures A.1 - A.3 provide an example of the relative numerical accuracy for RSS at each of the control points for dataset with $N = 50$ variables and average model size $k = 10$. All the algorithms, except the Sweep operator, are oscillating round zero with a constant variance, which is increasing with the sample size. The Sweep operator oscillates round zero for $T = 100$ with a growing variance, but diverges for the other sample sizes. However, for about the first 70 points, i.e. round 7 000 Markov chain steps, it does not perform much worse than the remaining algorithms. A similar level of accuracy could thus be obtained with the Sweep operator by resetting the crossproduct matrix and reinitializing the calculations after some fixed number of updates.

4.5 Methods for model space exploration

The posterior distribution of different possible models can be also viewed as a distribution of a set of binary strings or as a binary spatial field. The posterior model probabilities (4.1) can be then written as

$$p\left(\mathcal{M}_i|\,\mathbf{y}\right) = p\left(\boldsymbol{\gamma}_{\mathcal{M}_i}|\,\mathbf{y}\right) \propto m\left(\mathbf{y}|\,\boldsymbol{\gamma}_{\mathcal{M}_i}\right) p\left(\boldsymbol{\gamma}_{\mathcal{M}_i}\right), \tag{4.48}$$

where $\boldsymbol{\gamma}_{\mathcal{M}_i} = (\gamma_1, \ldots, \gamma_N)'$ is an binary vector, with $\gamma_k = 1$ indicating inclusion and $\gamma_k = 0$ omission of a variable k from model $\mathcal{M}_i$.

When evaluating the different MCMC algorithms we will take the basic RJMCMC algorithm 4.1 with both Add/Drop and Swap moves as our base case. The jump probabilities for the Add/Drop and Swap moves are as described in Section 4.2 and we randomize between the Add/Drop and Swap moves, proposing to swap variables with probability 1/2. We refer to this as the RJ ADS algorithm. We also consider a simplified version with only the Add/Drop move, i.e. the probability of proposing a swap is zero. This is the same as the $(\mathrm{MC})^3$ algorithm of Madigan and York (1995) and we refer to this as the RJ AD algorithm.

4.5.1 Gibbs and Metropolis-Hastings samplers

When the marginal likelihood is available in closed form a Gibbs sampler can be used to simulate the posterior distribution of the binary vector $\boldsymbol{\gamma}$ directly. The Gibbs sampler is obtained as a special case of the basic RJMCMC Algorithm 4.1B by selecting the jump distribution as the conditional posterior

$$p\left(\gamma_i'|\mathbf{y}, \boldsymbol{\gamma}_{\setminus i}\right) = \frac{m\left(\mathbf{y}|\gamma_i', \boldsymbol{\gamma}_{\setminus i}\right) p\left(\gamma_i'|\boldsymbol{\gamma}_{\setminus i}\right)}{\sum_{j=0}^{1} m\left(\mathbf{y}|\gamma_i = j, \boldsymbol{\gamma}_{\setminus i}\right) p\left(\gamma_i = j|\boldsymbol{\gamma}_{\setminus i}\right)} \tag{4.49}$$

Algorithm 4.3 Gibbs sampling scheme for variable selection

Suppose that the Markov chain is at model $\mathcal{M}$ represented by the binary vector $\boldsymbol{\gamma}$.

1. For $i = 1, \ldots, N$, draw γ_i' from the conditional posterior $p\left(\gamma_i' | \mathbf{y}, \boldsymbol{\gamma}_{\setminus i}\right)$
2. Set the new model to $\mathcal{M}'$ as given by $\boldsymbol{\gamma}'$.

when updating component i with current value γ_i to a (possibly) different value γ_i'. This proposes to stay with probability $p\left(\gamma_i | \mathbf{y}, \boldsymbol{\gamma}_{\setminus i}\right) = 1 - p\left(\gamma_i' | \mathbf{y}, \boldsymbol{\gamma}_{\setminus i}\right)$, and move with probability $1 - p\left(\gamma_i | \mathbf{y}, \boldsymbol{\gamma}_{\setminus i}\right) = p\left(1 - \gamma_i | \mathbf{y}, \boldsymbol{\gamma}_{\setminus i}\right)$. The acceptance probability (4.4) clearly simplifies to 1. If in addition the index i to update is drawn with probability $1/N$ it is clear that this is analogous to the basic RJMCMC algorithm with only the Add/Drop move. The probability of moving to a new model is, however, larger for the RJMCMC scheme than for the Gibbs sampler, and it follows from Peskun (1973) that RJMCMC with only Add/Drop moves will mix better and is statistically more efficient.

In practice the Gibbs sampler is usually implemented with a systematic scan updating all the variables, as in Smith and Kohn (1996), rather than drawing a single variable to update at random. In this case the Gibbs sampler corresponds to a thinned RJMCMC sampler where every N^{th} draw is retained.

Kohn, Smith, and Chan (2001) propose several sampling schemes designed to reduce the computational burden by avoiding unnecessary evaluations of the marginal likelihood. The Gibbs sampler and the basic RJMCMC scheme each require that $m\left(\mathbf{y} | \boldsymbol{\gamma}'\right)$ is evaluated in every step but in many, perhaps most, cases we will remain at the same model. In particular, it is likely that an attempt to change γ_i from 0 to 1 will fail since the total number of variables, N, is typically large relative to the number of variables in the model. Here, we consider the sampling scheme 2 (SS2) of Kohn, Smith, and Chan (2001) outlined in Algorithm 4.4. It reduces the number of evaluations of the marginal likelihood by using the conditional prior $p\left(\gamma_i | \boldsymbol{\gamma}_{\setminus i}\right)$ as the proposal distribution when considering flipping γ_i in an Add/Drop move. If the prior probability $p\left(\gamma_i = 1 | \boldsymbol{\gamma}_{\setminus i}\right)$ is small, the proposal will often be to stay at $\gamma_i = 0$ and a likelihood evaluation is avoided. If, on the other hand, $\gamma_i = 1$, it is likely that a flip is proposed and the marginal likelihood must be evaluated, but the situation when $\gamma_i = 1$ should occur relatively less frequently. Kohn, Smith, and Chan (2001) show that the SS2 scheme is statistically less efficient than the Gibbs sampler, but required only about 10% as many likelihood evaluations in

Algorithm 4.4 Kohn, Smith and Chan Metropolis-Hastings sampling scheme with prior for $\boldsymbol{\gamma}$ as a proposal probability

Suppose that the Markov chain is at model $\mathcal{M}$ represented by the binary vector $\boldsymbol{\gamma}$.

1. For $i = 1, \ldots, N$,

   i. Draw γ_i' from $p\left(\gamma_i | \boldsymbol{\gamma}_{\setminus i}\right)$

   ii. Accept the proposal with probability

   $$\alpha = \frac{m\left(\mathbf{y} | \, \gamma_i', \boldsymbol{\gamma}_{\setminus i}\right)}{m\left(\mathbf{y} | \, \gamma_i, \boldsymbol{\gamma}_{\setminus i}\right)}. \tag{4.50}$$

   If $\gamma_i' = \gamma_i$ no likelihood evaluation is needed.

2. Set the new model to $\mathcal{M}'$ as given by $\boldsymbol{\gamma}'$.

their application.

The Kohn, Smith, and Chan (2001) idea of reducing the number of likelihood evaluations can also be applied to the RJMCMC sampling scheme by modifying the Add/Drop move proposal probabilities. Instead of selecting a variable index i with probability $1/N$ and proposing to flip γ_i, we select and index at random and propose to *flip* the value of γ_i with probability $p\left(\gamma_i = 1 | \boldsymbol{\gamma}_{\setminus i}\right)$, and propose to *stay* with probability $p\left(\gamma_i = 0 | \boldsymbol{\gamma}_{\setminus i}\right)$ when $\gamma_i = 0$. The proposal probabilities for $\gamma_i = 1$ are analogous. With just an Add/Drop move this is simply a random scan version of the SS2 scheme. For completeness we give this as Algorithm 4.5.

Kohn, Smith, and Chan (2001) also proposed versions of SS2 that update blocks of indices γ_i rather than a single index. The block schemes turned out to be less statistically efficient than one at a time updating with little additional computational saving. Consequently, we will not consider the block updating schemes here. Instead we consider a recently proposed scheme based on the Swendsen-Wang algorithm where the blocks are formed dynamically.

4.5.2 Swendsen-Wang algorithm for Bayesian variable selection

When there are high posterior correlations between components of $\boldsymbol{\gamma}$, the usual Markov chain Monte Carlo methods for exploring the posterior, which

Algorithm 4.5 RJMCMC with KSC proposal probabilities

Suppose that the Markov chain is at model $\mathcal{M}$ represented by the binary vector $\boldsymbol{\gamma}$ with k variables in the model.

1. With probability δ attempt to add or drop a variable.

   Draw an index $i = 1, \ldots, N$ with probability $1/N$ and γ_i' from $p\left(\gamma_i|\boldsymbol{\gamma}_{\setminus i}\right)$.

   The jump probability is $j(\mathcal{M}'|\mathcal{M}) = \delta p\left(\gamma_i'|\boldsymbol{\gamma}_{\setminus i}\right)$ and the probability of the reverse jump is $j(\mathcal{M}|\mathcal{M}') = \delta p\left(1-\gamma_i'|\boldsymbol{\gamma}_{\setminus i}\right)$

2. Otherwise attempt to swap two variables.

   Select an index i at random from $\{i : \gamma_i = 1\}$, an index j from $\{j : \gamma_j = 0\}$ and set $\gamma_i = 0, \gamma_j = 1$. This replaces variable i with variable j. The jump probabilities are $j(\mathcal{M}'|\mathcal{M}) = j(\mathcal{M}|\mathcal{M}') = (1-\delta)/k(N-k)$.

3. Accept the move with probability

$$\alpha = \min\left\{1, \frac{m\left(\mathbf{y}|\mathcal{M}'\right)p\left(\mathcal{M}'\right)j\left(\mathcal{M}|\mathcal{M}'\right)}{m\left(\mathbf{y}|\mathcal{M}\right)p\left(\mathcal{M}\right)j\left(\mathcal{M}'|\mathcal{M}\right)}\right\} \tag{4.51}$$

   otherwise stay at the current model.

update one component of $\boldsymbol{\gamma}$ at a time, can mix slowly. Nott and Green (2004) propose in such cases to use a MCMC algorithm analogous to Swendsen-Wang algorithm for the Ising model[8]. See Higdon (1998) for a review of the Swendsen-Wang algorithm.

Nott and Leonte (2004) combine the method of Nott and Green (2004) with a RJMCMC to obtain a sampling scheme for Bayesian variable selection in generalized linear models, which is applicable when the regression coefficients can not be integrated out of the posterior distribution analytically.

Let $\mathbf{v} = \{v_{ij} : 1 \leq i < j \leq N\}$ be a collection of auxiliary variables, and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ be the subvector of $\boldsymbol{\beta}$ consisting of nonzero elements. The target distribution of the Markov chain is

$$p\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \mathbf{v}|\mathbf{y}\right) \propto p\left(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}\right)p\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}\right)p\left(\boldsymbol{\gamma}\right)p\left(\mathbf{v}|\boldsymbol{\gamma}\right) \tag{4.52}$$

instead of the usual

$$p\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|\mathbf{y}\right) \propto p\left(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}\right)p\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}\right)p\left(\boldsymbol{\gamma}\right). \tag{4.53}$$

[8]See for example Kindermann R. and J. Laurie Snell (1980): *Markov Random Fields and Their Applications*. The American Mathematical Society, Rhode Island.

Adding auxiliary variables can often simplify calculations and lead to improved mixing. In general, it may be possible to design good sampling algorithms for the distribution $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \mathbf{v})$ more easily then for the distribution $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$. One can simulate from the joint distribution $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}, \mathbf{v})$ and then obtain information about the posterior distribution on the model parameters by ignoring the simulated values for the auxiliary variables $\mathbf{v}$. The auxiliary variables do not need to have an immediate interpretation.

The conditional distribution for the auxiliary variables in (4.52) is

$$p(\mathbf{v}|\boldsymbol{\gamma}) = \frac{\prod_{i<j} \mathit{1}\left(0 \leq v_{ij} \leq \exp\left(\psi_{ij}\mathit{1}\left(\gamma_i = \gamma_j\right)\right)\right)}{\exp\left(\sum_{i<j} \psi_{ij}\mathit{1}\left(\gamma_i = \gamma_j\right)\right)}, \tag{4.54}$$

where $\mathit{1}(\mathcal{K})$ is indicator function, that equals 1 if condition $\mathcal{K}$ is satisfied and 0 otherwise. The parameters ψ_{ij} are interaction parameters and the process how they are determined is explained in the next section. The conditional distribution (4.54) makes v_{ij} given $\boldsymbol{\gamma}$ conditionally independent and uniform on the interval $\left[0, \exp\left(\psi_{ij}\mathit{1}\left(\gamma_i = \gamma_j\right)\right)\right]$.

The auxiliary variables $\mathbf{v}$ impose constraints on the value of $\boldsymbol{\gamma}$. If $\psi_{ij} > 0$ and $1 \leq v_{ij} \leq \exp(\psi_{ij})$, then the condition $v_{ij} \leq \exp\left(\psi_{ij}\mathit{1}\left(\gamma_i = \gamma_j\right)\right)$ is satisfied if $\gamma_i = \gamma_j$. For $\psi_{ij} < 0$ and $\exp(\psi_{ij}) \leq v_{ij} \leq 1$, the condition $v_{ij} \leq \exp\left(\psi_{ij}\mathit{1}\left(\gamma_i = \gamma_j\right)\right)$ is satisfied if $\gamma_i \neq \gamma_j$. There are no constraints for γ_i and γ_j, if $\psi_{ij} > 0$ and $0 \leq v_{ij} < 1$, and if $\psi_{ij} < 0$ and $0 \leq v_{ij} < \exp(\psi_{ij})$. The auxiliary variables define clusters among components of $\boldsymbol{\gamma}$ by the constrains $\gamma_i = \gamma_j$ and $\gamma_i \neq \gamma_j$. The constraints always have at least one feasible solution, because they are based on the current value of $\boldsymbol{\gamma}$. Let $\mathcal{C} = \mathcal{C}(\mathbf{v})$ be a cluster defined by the auxiliary variables and denote by $\boldsymbol{\gamma}(\mathcal{C})$ the subset of $\boldsymbol{\gamma}$ corresponding to the variables in $\mathcal{C}$, and denote by $\boldsymbol{\gamma}\left(\bar{\mathcal{C}}\right)$ the remaining components of $\boldsymbol{\gamma}$. The elements of $\boldsymbol{\gamma}(\mathcal{C})$ can be updated by flipping their values from zero to one and vice versa.

Using clusters allows to update components of $\boldsymbol{\gamma}$ in blocks, rather than one or two at a time. This is very beneficial in situations where predictors are far from orthogonality and thus the sampling schemes for exploring the posterior distribution work very poorly. Algorithm 4.6 describes the RJMCMC algorithm 4.1 extended by the Swendsen-Wang algorithm for exploring the model space.

Algorithm 4.6 Swendsen-Wang reversible jump Markov chain Monte Carlo

Suppose that the Markov chain is at model $\mathcal{M}$ and the interaction parameters ψ_{ij} have been calculated. Denote the binary indicator vector for the parameters of the model $\mathcal{M}$ as $\boldsymbol{\gamma}$.

1. Generate auxiliary variables $\mathbf{v}$ from $p(\mathbf{v}|\boldsymbol{\gamma})$.
2. Uniformly select a variable i from the set of possible variables and find the cluster $\mathcal{C}$ containing the variable i .
3. Propose a jump from model $\mathcal{M}$ to a new model $\mathcal{M}'$ by setting $\boldsymbol{\gamma}'(\mathcal{C}) = 1 - \boldsymbol{\gamma}(\mathcal{C})$.
4. Accept the proposed move with probability
$$\alpha = \min\left\{1, \frac{m(\mathbf{y}|\boldsymbol{\gamma}')\, p(\boldsymbol{\gamma}')}{m(\mathbf{y}|\boldsymbol{\gamma})\, p(\boldsymbol{\gamma})} \exp\left(\sum_{i<j} \psi_{ij}\Big(1(\gamma_i = \gamma_j) - 1\left(\gamma_i' = \gamma_j'\right)\Big)\right)\right\}. \tag{4.55}$$
5. Set $\mathcal{M} = \mathcal{M}'$ if the move is accepted.

Since $\boldsymbol{\gamma}'\left(\bar{\mathcal{C}}\right) = \boldsymbol{\gamma}\left(\bar{\mathcal{C}}\right)$ the acceptance probability is simplified to

$$\alpha = \min\left\{1, \frac{m(\mathbf{y}|\boldsymbol{\gamma}')\, p(\boldsymbol{\gamma}')}{m(\mathbf{y}|\boldsymbol{\gamma})\, p(\boldsymbol{\gamma})} \exp\left(\sum_{(i,j)\partial\mathcal{C}} \psi_{ij}\Big(1(\gamma_i = \gamma_j) - 1\left(\gamma_i' = \gamma_j'\right)\Big)\right)\right\}, \tag{4.56}$$

where $\partial\mathcal{C} = \{(i,j) : i < j \text{ and either } i \in \mathcal{C}, j \notin \mathcal{C} \text{ or } i \notin \mathcal{C}, j \in \mathcal{C}\}$.
Steps 1 and 2 can be combined, since only the components of $\mathbf{v}$, which determine the cluster containing γ_i in step 2, have to be generated.

Choice of interaction parameters

The choice of interaction parameters is very important, as the auxiliary variables conditionally remove interactions among components of $\boldsymbol{\gamma}$. Nott and Green (2004) suggest to compute the interaction parameters ψ_{ij} via the expression

$$\psi_{ij}^U = 0.5 \times \left(\sum_{\substack{\gamma_i=\gamma_j,\\ \gamma_k=\gamma_k^*,\, k\neq i,j}} \log m\left(\mathbf{y}|\boldsymbol{\gamma}\right) - \sum_{\substack{\gamma_i\neq\gamma_j,\\ \gamma_k=\gamma_k^*,\, k\neq i,j}} \log m\left(\mathbf{y}|\boldsymbol{\gamma}\right) \right), \qquad (4.57)$$

where $\boldsymbol{\gamma}^*$ is set to be a vector of ones. The interaction parameters are then formed as

$$\psi_{ij} = c\psi_{ij}^U 1\big(\left|c\psi_{ij}^U\right| \geq t\big), \qquad (4.58)$$

where c is a scaling factor and t is a threshold parameter. The transformation (4.58) scales down the values ψ_{ij}^U by a factor of c, and truncates these values to zero if the scaled values are less than t in magnitude. With $\boldsymbol{\gamma}^*$ to be vector of ones, c is chosen so that all the algorithm interaction parameters lie in the interval $[-1, 1]$, and t is set to 0.1. From the definition of the auxiliary variables and the constraints conditions it is easily seen that if the current states for γ_i and γ_j are such that a constraint is possible between them at the next iteration, then the probability of such a constraint is $1 - \exp\left(-\left|\psi_{ij}\right|\right)$. If $|\psi_{ij}|$ is large, this constraint probability will be close to one. If many of the algorithm interaction parameters are large, the cluster $\mathcal{C}$ tends to contain large numbers of variables, and at least one of the γ in a large cluster will have a value fixed by the likelihood. This means that any proposal to flip the values in a large cluster is unlikely to be accepted, which prevents mixing in this sampling scheme. Scaling down the parameters has therefore a beneficial effect on the performance of the algorithm.

In datasets containing many variables computing all interaction parameters ψ_{ij} can be very time consuming. Nott and Green (2004) suggest to reduce the number of pairs (i, j) for which the interactions parameters have to be calculated by using standard multicollinearity diagnostic, the variance proportions. For linearly independent variables the interaction parameters are set to zero, and only for variables (i, j) involved in severe linear dependence they are allowed to be nonzero.

Write $\mathbf{X}'\mathbf{X} = \mathbf{W}\boldsymbol{\Lambda}\mathbf{W}'$ for the eigenvalue decomposition, where $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues with diagonal entries $\lambda_1, \ldots, \lambda_N$ and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_N]$ is a orthogonal matrix with columns given by the eigenvectors $\mathbf{w}_1, \ldots, \mathbf{w}_N$ of

$\mathbf{X}'\mathbf{X}$. The variance of the estimated parameters, $\hat{\boldsymbol{\beta}}$, can be written as

$$\text{var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} = \sigma^2 \mathbf{W}\boldsymbol{\Lambda}^{-\mathbf{1}}\mathbf{W}', \tag{4.59}$$

and the variance of $\hat{\beta}_i$ is

$$\text{var}\left(\hat{\beta}_i\right) = \sigma^2 \sum_{j=1}^{N} \frac{w_{ij}^2}{\lambda_j}. \tag{4.60}$$

The variance proportion that can be attributed to the eigenvalue λ_k is

$$\xi_{ki} = \frac{w_{ik}^2/\lambda_k}{\sum_{j=1}^{N} w_{ij}^2/\lambda_j}. \tag{4.61}$$

If for a given small value of λ_k both variance proportions ξ_{ki} and ξ_{kj} are large this suggests that variables i and j are involved in a near linear dependence among the regressors. Nott and Green suggest to calculate the interaction parameters ψ_{ij}^U only for such pairs (i, j), where both ξ_{ki} and ξ_{kj} are bigger than some cut-off value for some eigenvalue λ_k. The interaction parameters are then scaled and truncated as in expression (4.58). The cut-off value is set to 0.25 in this chapter.

4.6 Performance of MCMC algorithms

In this section we wish to address the issues outlined in Section 4.2. The two main questions are: how quickly a sampler converges to the posterior distribution and how fast does it account for all but a negligible proportion of the total posterior probabilities. We investigate these in two experiments based on different designs for the matrix of potential explanatory variables. The first experiment is the same as the basic design with 15 variables described in Section 4.4.

The second experiment is more challenging with severe and complicated multicollinearity among potential regressors. It is based on an example originally in George and McCulloch (1997), and used by others, Nott and Green (2004) and Nott and Leonte (2004). We simulate 15 variables $\mathbf{X}_{(T\times 15)}$ as follows. First, 16 i.i.d. standard normal variables, $\mathbf{z}_i$, are generated. Then we construct the regressors as $\mathbf{x}_i = \mathbf{z}_i + 2\mathbf{z}_{16}$ for $i = 1, 3, 5, 8, 9, 10, 12, 13, 14, 15$. Furthermore, $\mathbf{x}_i = \mathbf{x}_{i-1} + 0.15\mathbf{z}_i$, for $i = 2, 4, 6$, $\mathbf{x}_7 = \mathbf{x}_8 + \mathbf{x}_9 - \mathbf{x}_{10} + 0.15\mathbf{z}_7$, and finally $\mathbf{x}_{11} = -\mathbf{x}_{12} - \mathbf{x}_{13} + \mathbf{x}_{14} + \mathbf{x}_{15} + 0.15\mathbf{z}_{11}$. This leads to a correlation

about 0.998 between $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$ for variables $1, 3$ and 5, and strong linear dependencies among $(\mathbf{x}_7, \ldots, \mathbf{x}_{10})$ and $(\mathbf{x}_{11}, \ldots, \mathbf{x}_{15})$. The true model is

$$y_t = \mathbf{x}_t \boldsymbol{\beta} + \sigma \varepsilon_t, \tag{4.62}$$

where

$$\boldsymbol{\beta} = \begin{pmatrix} 1.5 & 0 & 1.5 & 0 & 1.5 & 0 & 1.5 & -1.5 & 0 & 0 & 1.5 & 1.5 & 1.5 & 0 & 0 \end{pmatrix}',$$

the disturbances ε_t are i.i.d. standard normal, and $\sigma = 2.5$. In both experiments we set $T = 250$. We generate 100 different datasets for each experiment. We refer to these two experiments as the FLS and NL experiment, respectively.

In order to assess the performance of the different samplers we test how close the posterior distribution of the chain is to the exact posterior distribution. Calculating the exact posterior distribution is trivial for these experiments since the marginal likelihoods are available in a closed form. With only $N = 15$ predictors in both exercises, and the constant term always included in the model, exactly 32768 models have to be evaluated. Using our fastest least squares routine this takes very little time, about 0.2 seconds.

To assess the convergence we use the Kolmogorov-Smirnov (KS) test as one metric. The KS test is defined as the maximum value of absolute difference between the empirical cumulative distribution function, $S_n(x)$, and the assumed cumulative distribution function, $P(x)$,

$$D = \max_{-\infty < x < \infty} |S_n(x) - P(x)|, \tag{4.63}$$

where n denotes the sample size.

Brooks, Giudici, and Philippe (2003) use the KS test for testing homogeneity of sub-populations of different chains under the assumption that replicated chains that have converged generate similar posterior model probability estimates. The use of the KS test is justified by considering the sample-path of a model indicator, $\mathcal{M}_i$. Brooks, Giudici, and Philippe point out that the KS test is invariant under reparameterisation of x. The test is however not invariant to model labelling, so the value of D may change with a different model labelling. On the other hand, the interpretation of the diagnostic plots over time, in terms of whether or not the simulations are performing well, is generally invariant across different labelling schemes. The KS test requires continuous variables to derive an exact distribution, but as Brooks, Giudici, and Philippe also validate, with sufficiently large number of models, this assumption is valid in an approximate sense. In the test (4.63) the realizations

of x are replaced by $\mathcal{M}_i$, the model indicator, and $n = M$, the total number of models visited by the Markov chain.

We evaluate the following samples:

GIBBS	Gibbs sampler (Algorithm 4.3)
KSC	Kohn, Smith, and Chan's (2001) SS2 algorithm (Algorithm 4.4)
RJ ADS	RJMCMC (Algorithm 4.1) with Add/Drop and Swap moves, as defined on page 126 and $\delta = 1/2$
RJ AD	RJMCMC with Add/Drop step only, $\delta = 0$
RJ KSC ADS	RJMCMC with KSC jump proposal, ADS steps, $\delta = 1/2$, (Algorithm 4.5)
RJ KSC AD	RJMCMC with KSC jump proposal, no swap step, $\delta = 0$
SWANG	Swendsen-Wang algorithm (Algorithm 4.6)

4.6.1 Results

For each sampler and each dataset we run the chain in total for $s = 75\,000$, $250\,000$, $500\,000$, $1\,000\,000$ and $2\,000\,000$ steps. We stop the chain after each $r = s/5$ steps, giving 5 control points. At the control points we record the output of the chain, i.e. the KS test statistics are calculated based on the last r steps. After the control point the chain continues moving form the last visited model $\mathcal{M}_i^{(r)}$, but the count of visited models is discarded. For each s the outcome of the KS tests gives an indication of the burn-in needed, provided that the 'batch-size', r, is sufficient to provide a good estimate of the posterior model probabilities, or at least allow us to visit all but a negligible set of models. The KS test is carried out both for the exact posterior probabilities based on visited models by the Markov chain and the relative frequencies from the chain. We refer to these as exact probabilities and MCMC probabilities, respectively. Tables 4.5 - 4.8 report the number of significant KS tests at significance level $\alpha = 0.05$.

The results for the 'simple' FLS experiment does not show any significant differences among the samplers. All the samplers perform well, and according to the KS test, $15\,000$ steps of the Markov chain is sufficient for estimating the posterior model probabilities.

Turning to the more complicated NL experiment, we find substantial differences. For the exact probabilities the Swendsen-Wang sampler produces posterior distributions very close to the 'true' distribution already for short chains. It is followed by the RJ KSC ADS and RJ ADS samplers. It appears that these three algorithms will visit all but a negligible set of models in $15\,000$

Table 4.5 Number of significant Kolmogorov-Smirnov tests for exact probabilities, FLS experiment.

r		*control step*				
s		1	2	3	4	5
	GIBBS	0	0	0	0	0
	KSC	0	0	0	0	0
15 000	RJ ADS	0	0	0	0	0
75 000	RJ AD	0	0	0	0	0
	RJ KSC ADS	1	0	0	0	0
	RJ KSC AD	0	0	0	0	0
	SWANG	1	0	0	0	0
	GIBBS	0	0	0	0	0
	KSC	0	0	0	0	0
50 000	RJ ADS	0	0	0	0	0
250 000	RJ AD	0	0	0	0	0
	RJ KSC ADS	0	0	0	0	0
	RJ KSC AD	0	0	0	0	0
	SWANG	1	0	0	0	0
	GIBBS	0	0	0	0	0
	KSC	0	0	0	0	0
100 000	RJ ADS	0	0	0	0	0
500 000	RJ AD	0	0	0	0	0
	RJ KSC ADS	0	0	0	0	0
	RJ KSC AD	0	0	0	0	0
	SWANG	1	0	0	0	0
	GIBBS	0	0	0	0	0
	KSC	0	0	0	0	0
200 000	RJ ADS	0	0	0	0	0
1 000 000	RJ AD	0	0	0	0	0
	RJ KSC ADS	0	0	0	0	0
	RJ KSC AD	0	0	0	0	0
	SWANG	1	0	0	0	0
	GIBBS	0	0	0	0	0
	KSC	0	0	0	0	0
400 000	RJ ADS	0	0	0	0	0
2 000 000	RJ AD	0	0	0	0	0
	RJ KSC ADS	0	0	0	0	0
	RJ KSC AD	0	0	0	0	0
	SWANG	0	0	0	0	0

Table 4.6 Number of significant Kolmogorov Smirnov tests for MCMC probabilities, FLS experiment.

r		*control step*				
s		1	2	3	4	5
	GIBBS	0	0	0	0	0
	KSC	0	2	0	0	0
15 000	RJ ADS	1	0	0	0	0
75 000	RJ AD	0	0	0	0	0
	RJ KSC ADS	6	1	3	5	1
	RJ KSC AD	0	0	0	0	0
	SWANG	1	1	0	0	0
	GIBBS	0	0	0	0	0
	KSC	1	0	0	0	0
50 000	RJ ADS	0	0	0	0	0
250 000	RJ AD	0	0	0	0	0
	RJ KSC ADS	2	0	0	0	2
	RJ KSC AD	0	0	0	0	0
	SWANG	1	0	0	0	0
	GIBBS	0	0	0	0	0
	KSC	0	0	0	0	0
100 000	RJ ADS	0	0	0	0	0
500 000	RJ AD	0	0	0	0	0
	RJ KSC ADS	0	0	0	0	0
	RJ KSC AD	0	0	0	0	0
	SWANG	1	0	0	0	0
	GIBBS	0	0	0	0	0
	KSC	0	0	0	0	0
200 000	RJ ADS	0	0	0	0	0
1 000 000	RJ AD	0	0	0	0	0
	RJ KSC ADS	0	0	0	0	0
	RJ KSC AD	0	0	0	0	0
	SWANG	1	0	0	0	0
	GIBBS	0	0	0	0	0
	KSC	0	0	0	0	0
400 000	RJ ADS	0	0	0	0	0
2 000 000	RJ AD	0	0	0	0	0
	RJ KSC ADS	0	0	0	0	0
	RJ KSC AD	0	0	0	0	0
	SWANG	0	0	0	0	0

Table 4.7 Number of significant Kolmogorov Smirnov tests for exact probabilities, NL experiment.

r		*control step*				
s		1	2	3	4	5
	GIBBS	5	13	15	9	6
	KSC	43	24	29	47	46
15 000	RJ ADS	3	0	0	1	0
75 000	RJ AD	14	19	11	2	13
	RJ KSC ADS	2	6	1	4	3
	RJ KSC AD	51	30	49	52	48
	SWANG	0	0	0	0	0
	GIBBS	2	0	1	0	0
	KSC	3	5	13	3	16
50 000	RJ ADS	4	0	0	0	0
250 000	RJ AD	5	0	6	2	4
	RJ KSC ADS	0	0	0	0	0
	RJ KSC AD	12	16	28	33	33
	SWANG	0	0	0	0	0
	GIBBS	1	0	0	0	0
	KSC	0	0	4	1	0
100 000	RJ ADS	1	0	0	0	0
500 000	RJ AD	0	0	0	2	0
	RJ KSC ADS	1	0	0	0	0
	RJ KSC AD	3	16	1	6	0
	SWANG	0	0	0	0	0
	GIBBS	3	0	0	0	0
	KSC	1	0	0	2	0
200 000	RJ ADS	2	0	0	0	0
1 000 000	RJ AD	0	0	0	0	0
	RJ KSC ADS	0	0	0	0	0
	RJ KSC AD	0	0	0	0	0
	SWANG	0	0	0	0	0
	GIBBS	0	0	0	0	0
	KSC	1	0	0	0	0
400 000	RJ ADS	2	0	0	0	0
2 000 000	RJ AD	0	0	0	0	0
	RJ KSC ADS	1	0	0	0	0
	RJ KSC AD	0	0	0	0	0
	SWANG	0	0	0	0	0

Table 4.8 Number of significant Kolmogorov Smirnov tests for MCMC probabilities, NL experiment.

r		*control step*				
s		1	2	3	4	5
	GIBBS	72	55	44	21	56
	KSC	51	51	65	47	46
15 000	RJ ADS	38	45	40	36	31
75 000	RJ AD	29	49	49	72	63
	RJ KSC ADS	38	37	43	49	35
	RJ KSC AD	53	72	74	56	51
	SWANG	0	0	0	1	1
	GIBBS	55	33	40	22	39
	KSC	42	40	22	40	44
50 000	RJ ADS	28	29	20	19	18
250 000	RJ AD	37	59	53	44	52
	RJ KSC ADS	28	40	29	26	22
	RJ KSC AD	46	50	51	45	47
	SWANG	0	0	0	1	0
	GIBBS	24	19	32	26	28
	KSC	22	17	34	37	40
100 000	RJ ADS	24	12	6	11	16
500 000	RJ AD	30	21	31	26	24
	RJ KSC ADS	24	15	22	13	24
	RJ KSC AD	39	32	34	24	50
	SWANG	0	0	0	0	0
	GIBBS	12	20	17	11	12
	KSC	7	19	16	15	22
200 000	RJ ADS	9	4	17	11	7
1 000 000	RJ AD	9	12	16	11	27
	RJ KSC ADS	11	8	19	17	4
	RJ KSC AD	26	21	52	21	24
	SWANG	0	0	0	0	0
	GIBBS	8	4	2	5	3
	KSC	4	6	17	11	16
400 000	RJ ADS	5	9	5	2	2
2 000 000	RJ AD	3	8	2	7	19
	RJ KSC ADS	6	4	0	3	5
	RJ KSC AD	12	35	19	16	12
	SWANG	0	0	0	0	0

steps. The next group of algorithms are the Gibbs sampler and RJ AD where 50 000 draws is sufficient for visiting all the important models. The chains produced by the KSC and RJ KSC AD algorithms perform worst and 100 000 steps appear to be needed. It is interesting to note the role of the Swap move. Algorithms with only the Add/Drop move perform badly, whereas the corresponding algorithm with a Swap move performs quite well. The inferior performance of the algorithms with KSC type proposals is to some extent expected. It remains to be seen if this is made up for by savings in computational time.

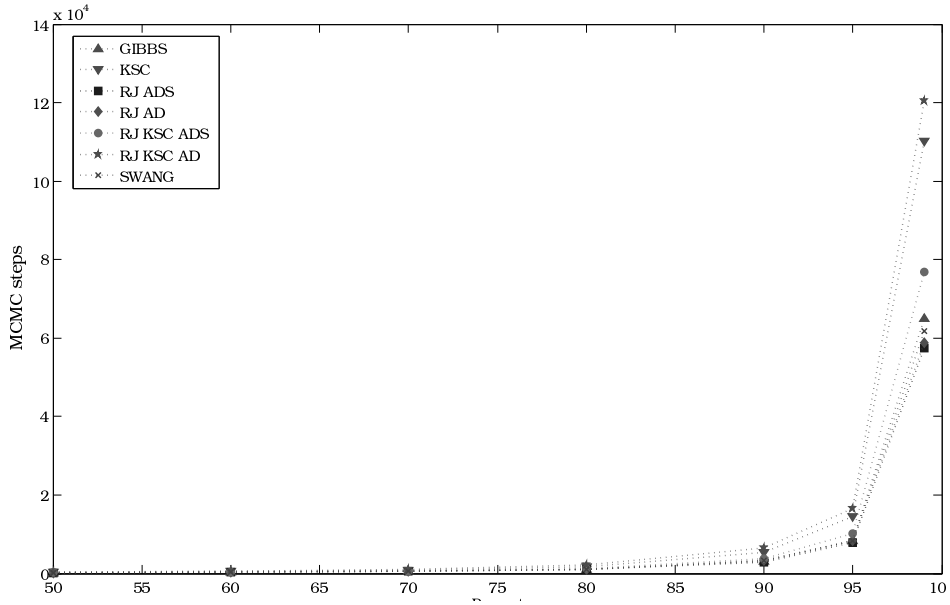
Turning to Table 4.8 and the issue of how well the relative frequencies of the visited models estimate the posterior probabilities, we again find that the Swendsen-Wang sampler performs best. The KS test is unable to reject the null hypothesis that the sampler has converged after only 15 000 steps. The performance of the Gibbs sampler, RJ ADS, RJ AD and RJ KSC ADS is similar to each other, these samplers appear to require about 400 000 steps for acceptable results. Again KSC and RJ KSC AD perform worst with KSC doing somewhat better than RJ KSC AD. For the latter we clearly need more than 400 000 steps of the Markov chain for good estimates of the posterior distribution.

Figures 4.2(a) and 4.2(b) plot the average number of steps that each sampler needed to reach a desired coverage. This is complementary to the results in Tables 4.5 and 4.7 and gives a somewhat more nuanced picture of the FLS experiment. The inferior performance of the KSC and RJ KSC AD algorithms is now evident for this dataset as well with these algorithms requiring about twice the number of steps for 99% coverage. For the NL experiment we again find that Swendsen-Wang performs best followed by RJ ADS with the Gibbs sampler, RJ AD and RJ KSC ADS in a third group.

This far we have not taken the computational requirements into account. Figures 4.3(a) and 4.3(b) show the average CPU time required for a desired level of coverage. In the FLS experiment we find that the Swendsen-Wang sampler requires most CPU time for 99% coverage, and that KSC along with the Gibbs sampler and RJ AD require the least. The RJ ADS, RJ KSC ADS and RJ KSC AD occupy the middle ground with RJ ADS doing best of these three algorithms. The speed of the KSC algorithm clearly compensates for the lack of statistical efficiency in this case. Turning to the NL experiment, we find that the RJ ADS algorithm requires least CPU time followed by KSC and RJ KSC ADS. The Gibbs sampler and the Swendsen-Wang algorithm have similar CPU time requirements, while the RJ AD and RJ KSC AD require most CPU time with RJ KSC AD performing quite poorly.

Figure 4.2 Average number of MCMC steps to reach model space coverage.

(a) FLS experiment



(b) NL experiment

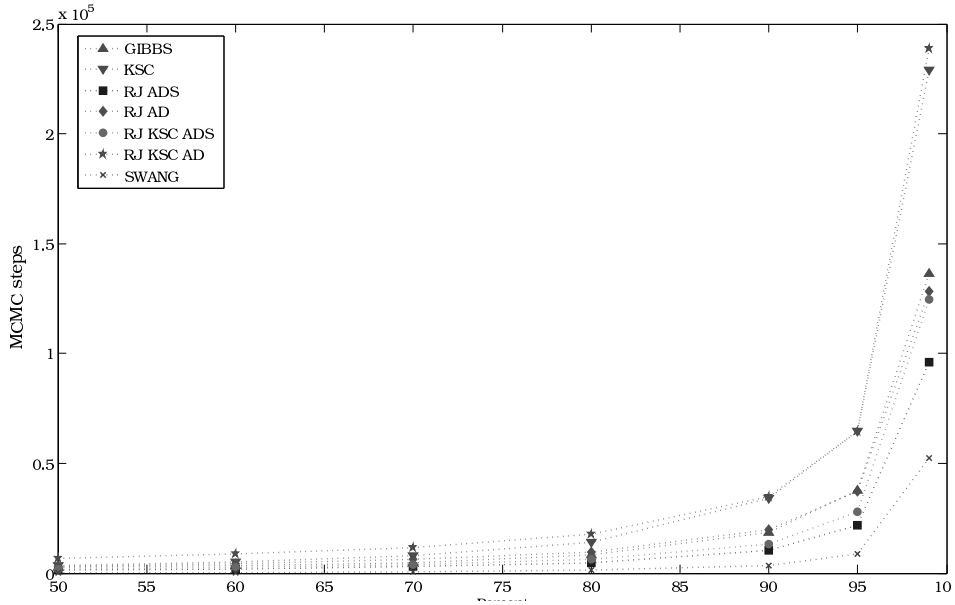
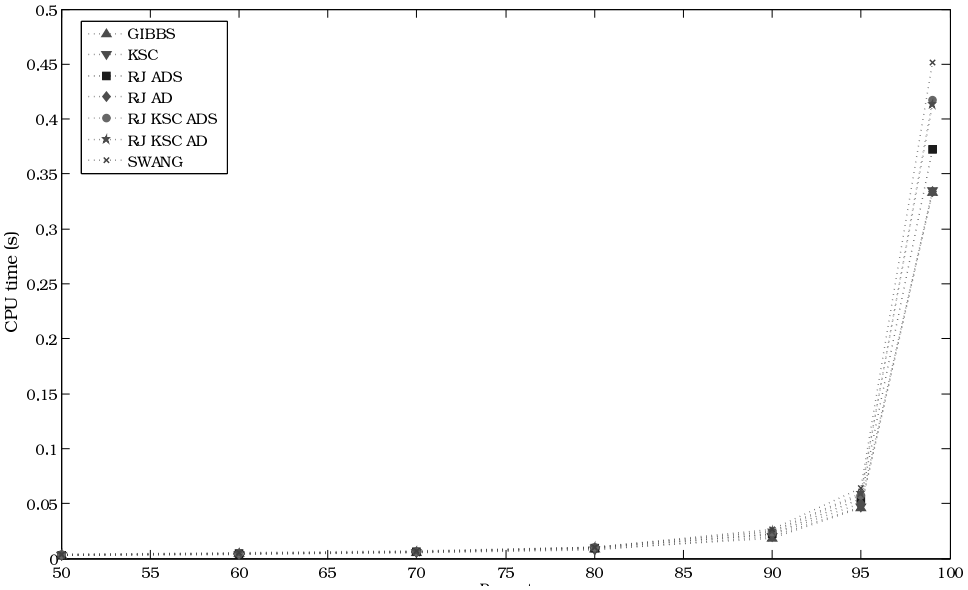
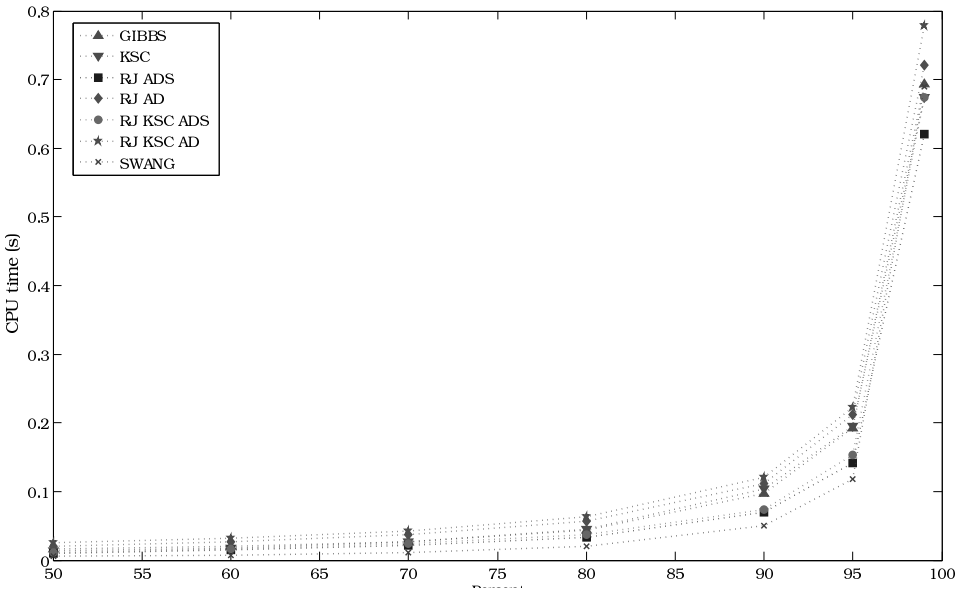


Figure 4.3 Average CPU time to reach model space coverage.

(a) FLS experiment



(b) NL experiment



The Swendsen-Wang algorithm is clearly a superior sampler among the ones considered here and it has much to offer. It is, however, relatively complicated to implement and requires more CPU time. Taking CPU time into account, other algorithms such as RJ ADS strike a balance between statistical efficiency and computational efficiency and might be preferable. It should, however, be noted that this is in a setting where the marginal likelihood is available in closed form and can be evaluated efficiently. For more complicated models the computational cost of generating a proposal is less important and the Swendsen-Wang algorithm might be more efficient in terms of computational time. At the same time this is a case where strategies like KSC, that reduce the number of likelihood evaluation, can be expected to do well.

4.7 Conclusions

In this chapter, the focus had been on a number of MCMC approaches and various algorithms for solving least squares problems. With large numbers of variables and models, situations often encountered in Bayesian selection or averaging exercises, fast calculations of model parameters and efficient search through the model and variable spaces are desirable.

The efficiency and accuracy of the least squares solvers is analysed from both theoretical and empirical perspectives. The results show that using Cholesky and Cholesky updating decompositions as well as the sweep operator, reduces the computational time substantially compared to the QR decomposition. The Cholesky decomposition and its update are also the most accurate algorithms, given the benchmark OLS.

Most MCMC samplers included in the analysis, are local approaches restricting transitions to adjacent subsets of the model space. The most successful is the reversible jump Markov chain Monte Carlo algorithm with Add/Drop and Swap moves implemented. This sampler produces posterior distributions very close to the true distribution and is also computationally efficient. When there is high dependence among the variables and the chain is mixing slowly, the Swendsen-Wang algorithm that allows for more global moves, provides substantial accuracy improvements compared to the local transition samplers.

Bibliography

Beaton, A. E. (1964): "The use of Special Matrix Operators in Statistical Calculus," Research Bulletin 64-51, Educational Testing Service, Princeton, New Jersey.

Brooks, S. P., P. Giudici, and A. Philippe (2003): "Nonparametric Convergence Assessment for MCMC Model Selection," *Journal of Computational & Graphical Statistics*, 12(1), 1–22.

Daniel, J. W., W. B. Gragg, L. Kaufman, and G. W. Stewart (1976): "Reorthogonalization and Stable Algorithms for Updating the Gram-Schmidt QR Factorization," *Mathematics of Computation*, 30(136), 772–795.

Denison, D. G. T., B. K. Mallick, and A. F. M. Smith (1998): "Automatic Bayesian Curve Fitting," *Journal of the Royal Statistical Society B*, 60(2), 333–350.

Dongarra, J. J., J. R. Bunch, C. B. Moler, and G. W. Stewart (1979): *Linpack Users' Guide*. SIAM, Philadelphia.

Fernández, C., E. Ley, and M. F. Steel (2001): "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100(2), 381–427.

George, E. I., and R. E. McCulloch (1997): "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Golub, G. H., and C. F. van Loan (1996): *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 3 edn.

Goodnight, J. H. (1979): "A Tutorial on the Sweep Operator," *The American Statistician*, 33(3), 149–158.

Green, P. J. (1995): "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82(4), 711–732.

Hanson, R. J., and T. Hopkins (2004): "Algorithm 830: Another Visit with Standard and Modified Givens Transformations and a Remark on Algorithm 539," *ACM Transactions on Mathematical Software*, 30(1), 86–94.

Higdon, D. M. (1998): "Auxiliary Variable Methods for Markov Chain Monte Carlo with Applications," *Journal of the American Statistical Association*, 93(442), 585–595.

Hoeting, J., D. Madigan, A. E. Raftery, and C. Volinsky (1999): "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14(4), 382–417.

Jacobson, T., and S. Karlsson (2004): "Finding Good Predictors for Inflation: A Bayesian Model Averaging Approach," *Journal of Forecasting*, 23(7), 479–496.

Kohn, R., M. Smith, and D. Chan (2001): "Nonparametric Regression Using Linear Combinations of Basis Functions," *Statistics and Computing*, 11(4), 313–322.

Koop, G., and S. Potter (2004): "Forecasting in Dynamic Factor Models Using Bayesian Model Averaging," *Econometrics Journal*, 7(2), 550–565.

Leamer, E. E. (1978): *Specification Searches, Ad hoc Inference with Nonexperimental Data*. John Wiley, New York.

Madigan, D., and A. E. Raftery (1994): "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89(428), 1535–1546.

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

Nott, D. J., and P. J. Green (2004): "Bayesian Variable Selection and the Swendsen-Wang Algorithm," *Journal of Computational & Graphical Statistics*, 13(1), 141–157.

Nott, D. J., and D. Leonte (2004): "Sampling Schemes for Bayesian Variable Selection in Generalized Linear Models," *Journal of Computational & Graphical Statistics*, 13(2), 362–382.

Peskun, P. H. (1973): "Optimum Monte-Carlo Sampling Using Markov Chains," *Biometrika*, 60(3), 607–612.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992): *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 2 edn.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997): "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92(437), 179–191.

Raftery, A. E., D. Madigan, and C. Volinsky (1995): "Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance (with Discussion)," in *Bayesian Statistics 5*, ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, pp. 323–349. Oxford University Press, Oxford.

Reichel, L., and W. B. Gragg (1990): "Algorithm 686: FORTRAN Subroutines for Updating the QR Decomposition," *ACM Transactions on Mathematical Software*, 16(4), 369–377.

Smith, M., and R. Kohn (1996): "Nonparametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, 75(2), 317–343.

Visual Numerics, Inc. (1994): *IMSL Fortran Statistical Library, vol. 1.*Visual Numerics, Inc., Houston, Texas, 3 edn.

Appendix A

Figures

Figure A.1 Relative accuracy for RSS, $T = 100, N = 50, k = 10$.

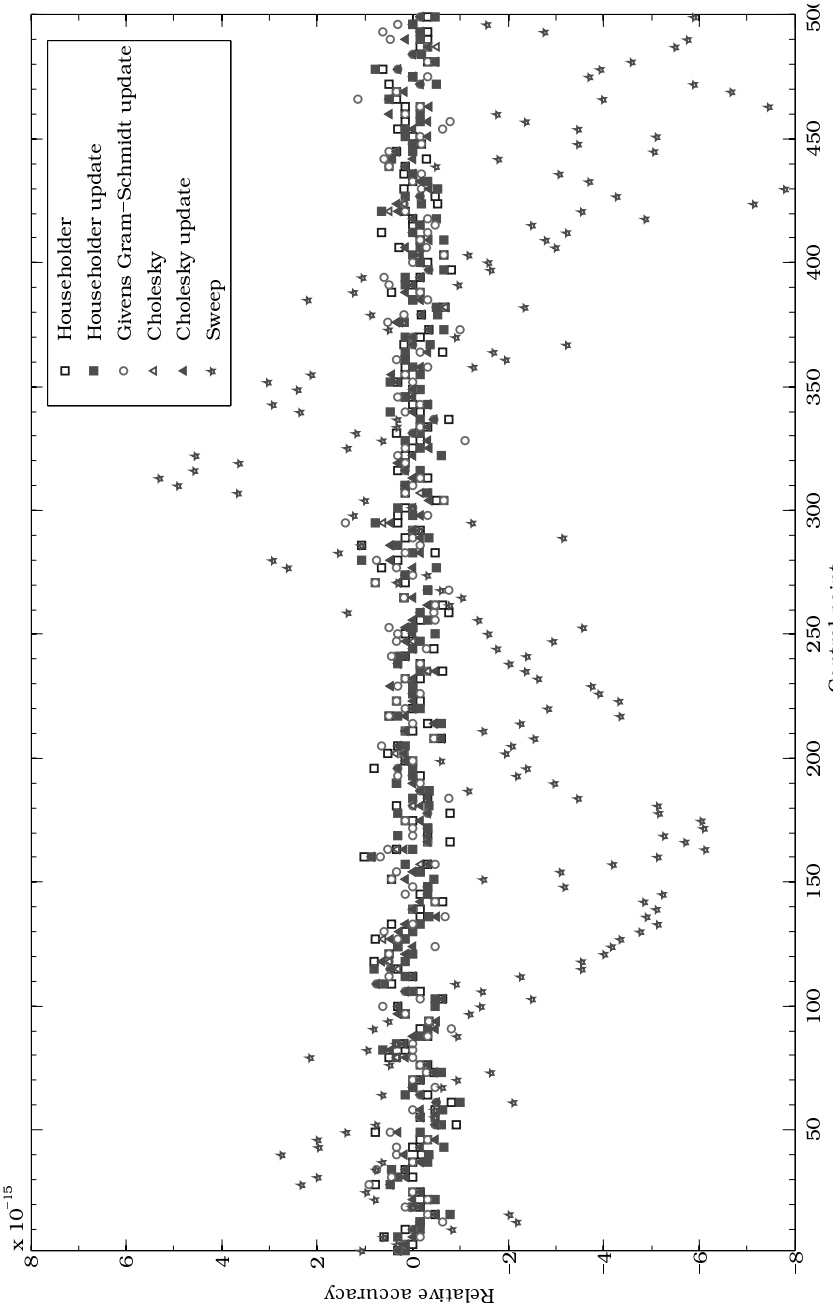


Figure A.2 Relative accuracy for RSS, $T = 250, N = 50, k = 10$.

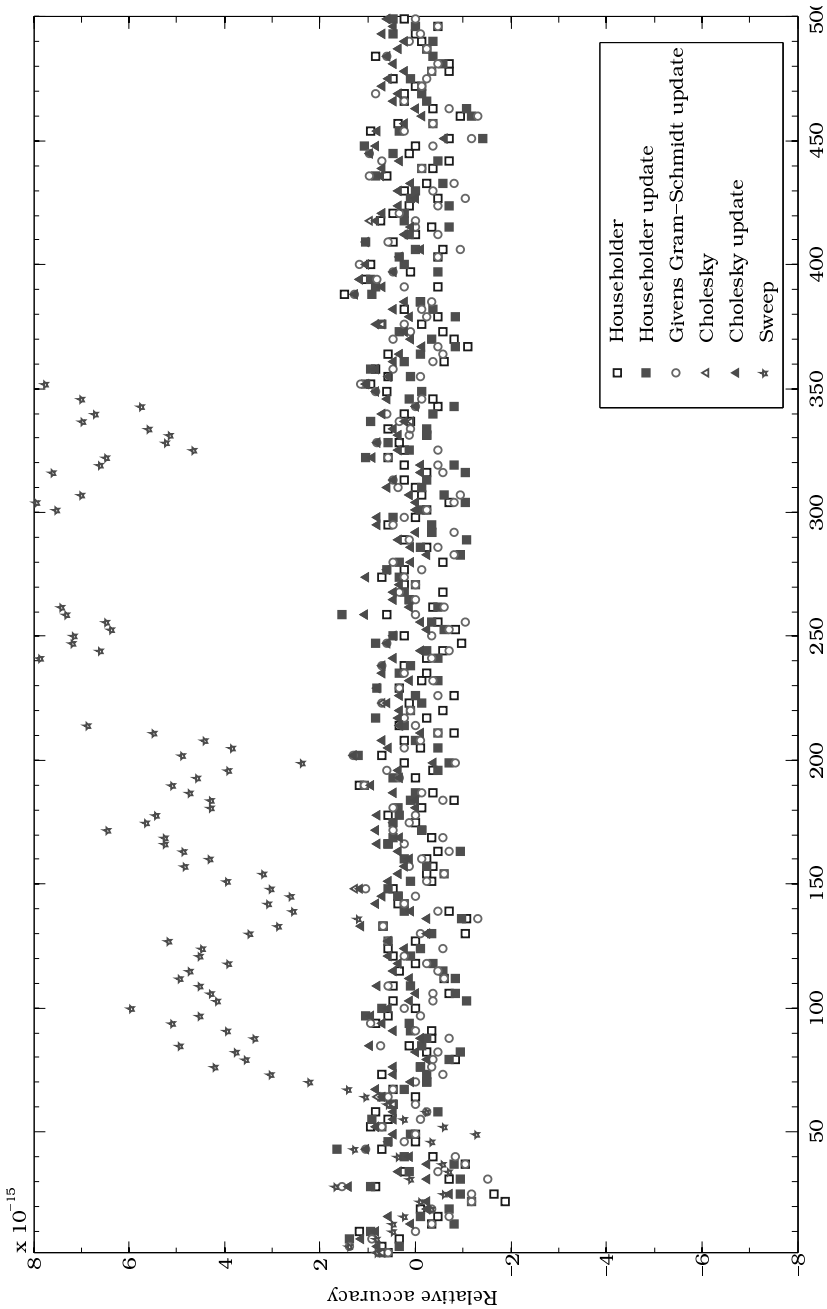
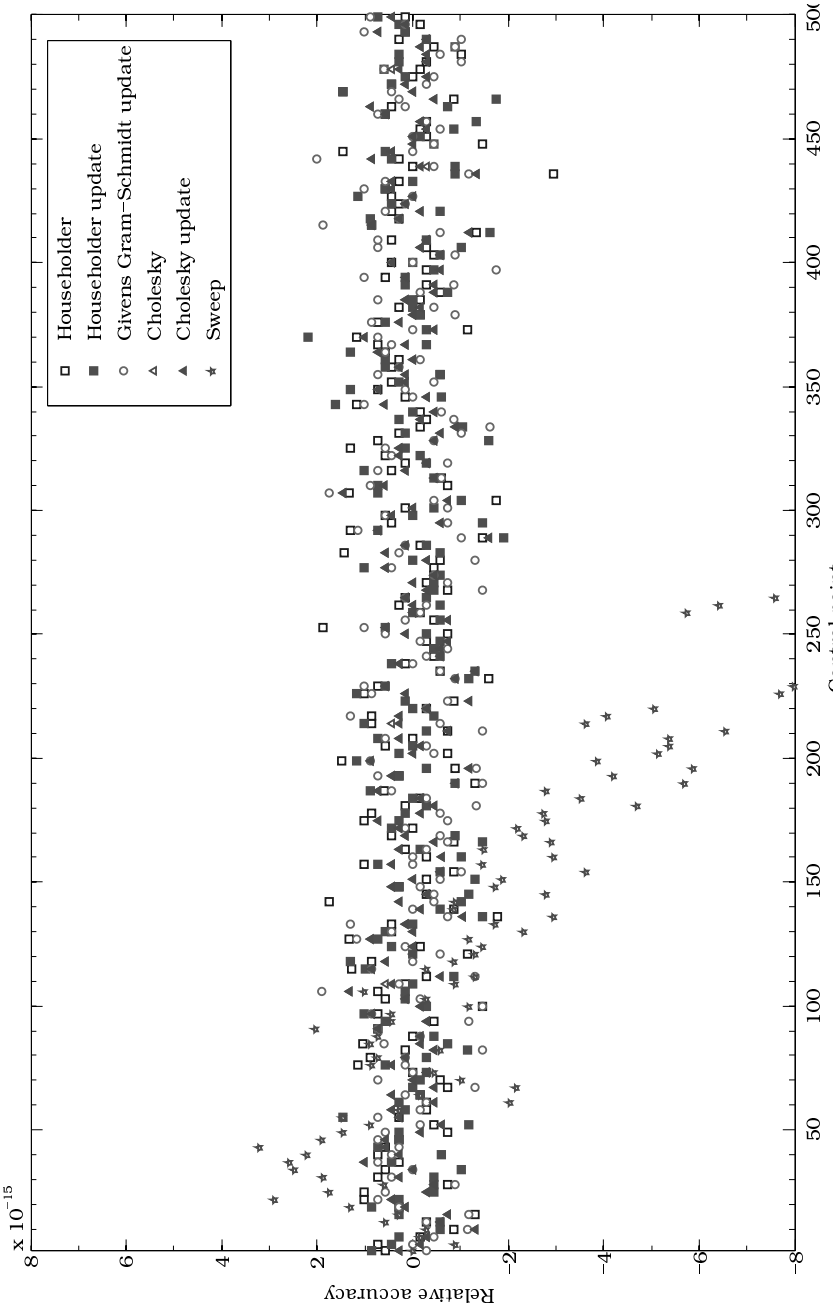


Figure A.3 Relative accuracy for RSS, $T = 400, N = 50, k = 10$.



Appendix B

Tables

Table B.1 CPU time (approximation in seconds) for $T = 100$.

		k			
		5	10	15	20
$N = 25$	OLS	17.56	25.38	32.33	38.91
	Householder	4.31	6.62	8.39	9.52
	Householder update	4.26	6.17	7.20	7.24
	Givens Gram-Schmidt update	4.44	5.98	6.35	5.82
	Cholesky	3.78	5.16	5.47	4.78
	Cholesky update	3.79	5.11	5.31	4.41
	Sweep	3.88	5.06	5.06	3.89
	Overhead	3.68	4.86	4.86	3.69
$N = 50$	OLS	20.21	26.95	33.94	41.74
	Householder	6.91	8.18	9.94	12.24
	Householder update	6.87	8.13	8.76	10.00
	Givens Gram-Schmidt update	7.05	7.94	7.83	8.45
	Cholesky	6.40	6.70	7.03	7.54
	Cholesky update	6.42	6.67	6.87	7.15
	Sweep	6.80	6.93	6.94	6.95
	Overhead	6.29	6.42	6.43	6.44
$N = 100$	OLS	24.26	30.99	38.07	45.61
	Householder	10.91	12.55	13.96	16.27
	Householder update	10.92	11.72	12.78	14.07
	Givens Gram-Schmidt update	11.07	11.46	11.82	12.41
	Cholesky	10.38	10.67	11.01	11.57
	Cholesky update	10.41	10.63	10.86	11.15
	Sweep	11.76	11.85	12.23	11.91
	Overhead	10.29	10.35	10.38	10.39

Table B.2 CPU time (approximation in seconds) for $T = 250$.

		k			
		5	10	15	20
$N = 25$	OLS	37.66	55.38	73.57	91.87
	Householder	5.09	9.05	13.30	18.37
	Householder update	4.93	7.90	10.32	12.14
	Givens Gram-Schmidt update	5.34	7.40	8.25	8.61
	Cholesky	3.78	5.14	5.47	4.79
	Cholesky update	3.80	5.25	5.30	4.41
	Sweep	3.87	5.05	5.06	3.89
	Overhead	3.69	4.84	4.85	3.69
$N = 50$	OLS	40.31	57.52	75.11	95.01
	Householder	7.69	10.67	14.93	20.58
	Householder update	7.53	9.51	11.92	14.90
	Givens Gram-Schmidt update	7.95	8.88	9.66	11.10
	Cholesky	6.42	6.71	7.14	7.54
	Cholesky update	6.43	6.67	6.88	7.15
	Sweep	6.80	6.91	6.94	6.95
	Overhead	6.29	6.42	6.42	6.45
$N = 100$	OLS	44.34	61.18	79.25	98.51
	Householder	11.71	14.58	18.82	24.52
	Householder update	11.56	13.44	15.87	18.99
	Givens Gram-Schmidt update	11.96	12.85	13.63	15.02
	Cholesky	10.40	10.66	11.01	11.56
	Cholesky update	10.41	10.64	10.85	11.15
	Sweep	11.77	11.84	11.88	11.90
	Overhead	10.30	10.51	10.38	10.39

Table B.3 CPU time (approximation in seconds) for $T = 400$.

		k			
		5	10	15	20
$N = 25$	OLS	58.14	86.17	114.64	145.35
	Householder	5.86	11.43	18.02	26.07
	Householder update	5.59	9.61	13.38	16.96
	Givens Gram-Schmidt update	6.23	8.79	10.10	11.32
	Cholesky	3.78	5.17	5.49	4.78
	Cholesky update	3.80	5.12	5.32	4.43
	Sweep	3.88	5.08	5.08	3.90
	Overhead	3.69	4.88	4.87	3.71
$N = 50$	OLS	60.29	87.33	115.85	148.70
	Householder	8.45	12.97	19.70	28.77
	Householder update	8.20	11.09	14.88	19.67
	Givens Gram-Schmidt update	8.83	10.24	11.43	13.65
	Cholesky	6.41	6.72	7.06	7.56
	Cholesky update	6.43	6.68	6.89	7.17
	Sweep	6.81	6.93	6.95	6.96
	Overhead	6.31	6.42	6.45	6.44
$N = 100$	OLS	64.57	91.72	120.01	151.44
	Householder	12.51	17.02	23.77	32.88
	Householder update	12.26	15.20	19.03	23.96
	Givens Gram-Schmidt update	12.85	14.17	15.38	17.57
	Cholesky	10.41	10.68	11.06	11.57
	Cholesky update	10.46	10.66	10.89	11.18
	Sweep	11.80	11.89	11.92	11.94
	Overhead	10.32	10.37	10.40	10.41

Table B.4 Average number of MCMC steps, for $r = 400\,000$ to reach the model space coverage of

	50%	60%	70%	80%	90%	95%	99%
FLS experiment							
GIBBS	152	282	503	1 004	3 023	8 436	64 872
KSC	234	468	894	1 752	5 452	14 568	110 120
RJ ADS	182	298	504	979	2 950	7 926	57 484
RJ AD	164	310	544	1 074	3 237	8 307	58 916
RJ KSC ADS	247	408	680	1 307	3 832	10 097	76 931
RJ KSC AD	308	533	980	2 112	6 510	16 551	120 486
SWANG	151	275	515	1 053	3 208	8 321	61 791
NL experiment							
GIBBS	2 642	3 473	4 858	8 266	18 689	37 609	136 585
KSC	3 629	5 177	7 998	14 236	33 985	64 726	229 126
RJ ADS	1 235	1 885	2 888	4 831	10 561	21 725	96 238
RJ AD	3 141	4 259	6 176	9 603	19 601	37 378	128 123
RJ KSC ADS	1 690	2 497	3 841	6 348	13 112	27 960	124 725
RJ KSC AD	6 783	8 684	11 670	17 845	34 788	64 678	238 735
SWANG	211	358	639	1 289	3 588	8 754	52 321

Table B.5 CPU time (approximation in seconds) to reach to reach the model space coverage of

	50%	60%	70%	80%	90%	95%	99%
FLS experiment							
GIBBS	0.0030	0.0042	0.0054	0.0082	0.0183	0.0459	0.3330
KSC	0.0030	0.0037	0.0053	0.0079	0.0197	0.0469	0.3348
RJ ADS	0.0032	0.0043	0.0059	0.0094	0.0221	0.0545	0.3721
RJ AD	0.0030	0.0038	0.0061	0.0090	0.0217	0.0500	0.3343
RJ KSC ADS	0.0036	0.0047	0.0065	0.0101	0.0239	0.0577	0.4173
RJ KSC AD	0.0033	0.0042	0.0059	0.0100	0.0251	0.0593	0.4133
SWANG	0.0037	0.0051	0.0066	0.0105	0.0266	0.0643	0.4517
NL experiment							
GIBBS	0.0164	0.0206	0.0273	0.0442	0.0974	0.1923	0.6932
KSC	0.0131	0.0178	0.0264	0.0453	0.1038	0.1956	0.6736
RJ ADS	0.0099	0.0146	0.0211	0.0335	0.0704	0.1418	0.6209
RJ AD	0.0204	0.0269	0.0371	0.0567	0.1121	0.2118	0.7215
RJ KSC ADS	0.0117	0.0164	0.0237	0.0366	0.0736	0.1531	0.6746
RJ KSC AD	0.0257	0.0320	0.0426	0.0631	0.1207	0.2220	0.7782
SWANG	0.0055	0.0073	0.0111	0.0201	0.0505	0.1178	0.6902