DISCUSSION PAPER SERIES

No. 5268
CEPR/EABCN No. 23/2005

FORECAST COMBINATION AND MODEL AVERAGING USING PREDICTIVE MEASURES

Jana Eklund and Sune Karlsson

INTERNATIONAL MACROECONOMICS

€ABCN
Euro Area Business Cycle Network
www.eabcn.org

Centre for Economic Policy Research

www.cepr.org

Available online at: www.cepr.org/pubs/dps/DP5268.asp

ISSN 0265-8003

FORECAST COMBINATION AND MODEL AVERAGING USING PREDICTIVE MEASURES

Jana Eklund, Stockholm School of Economics
Sune Karlsson, University of Örebro

Discussion Paper No. 5268
CEPR/EABCN No. 23/2005
October 2005

Centre for Economic Policy Research
90–98 Goswell Rd, London EC1V 7RR, UK
Tel: (44 20) 7878 2900, Fax: (44 20) 7878 2999
Email: cepr@cepr.org, Website: www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programme in **INTERNATIONAL MACROECONOMICS**. Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as a private educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions. Institutional (core) finance for the Centre has been provided through major grants from the Economic and Social Research Council, under which an ESRC Resource Centre operates within CEPR; the Esmée Fairbairn Charitable Trust; and the Bank of England. These organizations do not give prior review to the Centre's publications, nor do they necessarily endorse the views expressed therein.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

CEPR Discussion Paper No. 5268
CEPR/EABCN No. 23/2005

October 2005

ABSTRACT

Forecast Combination and Model Averaging Using Predictive Measures*

We extend the standard approach to Bayesian forecast combination by forming the weights for the model averaged forecast from the predictive likelihood rather than the standard marginal likelihood. The use of predictive measures of fit offers greater protection against in-sample overfitting and improves forecast performance. For the predictive likelihood we show analytically that the forecast weights have good large and small sample properties. This is confirmed in a simulation study and an application to forecasts of the Swedish inflation rate where forecast combination using the predictive likelihood outperforms standard Bayesian model averaging using the marginal likelihood.

JEL Classification: C11, C51, C52 and C53
Keywords: Bayesian model averaging, inflation rate, partial Bayes factor, predictive likelihood and training sample

Jana Eklund
Stockholm School of Economics
Box 6501
SE 11383 Stockholm
SWEDEN
Tel: (46 8) 736 9225
Fax: (46 8) 348 161
Email: jana.eklund@hhs.se

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=161418

Sune Karlsson
Professor of Statistics
Department of Economics, Statistics and Informatics
Örebro University
SE-701 82 Örebro
SWEDEN
Tel: (46 19) 301 257
Email: sune.karlsson@esi.oru.se

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=161419

* The publication of this Paper is funded by the Euro Area Business Cycle Network (www.eabcn.org). This Network provides a forum for the better understanding of the euro area business cycle, linking academic researchers and researchers in central banks and other policy institutions involved in the empirical analysis of the euro area business cycle. Earlier versions of this paper have been presented at the 2005 International Symposium on Forecasting and the EABCN workshop on Recent advances in forecast combination. We are grateful for useful comments from the participants in these conferences. Financial support from Sveriges Riksbank is gratefully acknowledged.

Submitted 13 September 2005

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) for a somewhat dated review, 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 paper 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 (2002), i.e. where there is a wealth of potential predictor variables. For computational simplicity we use simple 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, and Eklund and Karlsson (2005) compare the method of Jacobson and Karlsson with the approach of Stock and Watson based on using the first few principal components of the predictor variables. 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.

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

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

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 paper is organized as follows. The next section introduces the Bayesian model averaging technique and predictive densities, section 3 presents several Bayes factors and their asymptotics. Section 4 studies the small sample properties of the predictive likelihood. Section 5 contains a simulation study, section 6 an application to forecasts of the Swedish inflation and section 7 concludes.

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, $p(\theta_i|\mathcal{M}_i)$, and likelihood function $L(\mathbf{y}|\theta_i,\mathcal{M}_i)$. The posterior probabilities of the models after observing the data $\mathbf{y}$ follow from Bayes rule

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

where

$$m(\mathbf{y}|\mathcal{M}_i)=\int L(\mathbf{y}|\theta_i,\mathcal{M}_i)\,p(\theta_i|\mathcal{M}_i)\,d\theta_i \tag{2}$$

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

$$p(\phi|\mathbf{y})=\sum_{j=1}^{M}p(\phi|\mathbf{y},\mathcal{M}_j)\,p(\mathcal{M}_j|\mathbf{y}). \tag{3}$$

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

[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.

rules of conditional expectation

$$E\left[g\left(\phi\right)\middle|\,\mathbf{y}\right]=\sum_{j=1}^{M}E\left[g\left(\phi\right)\middle|\,\mathbf{y},\mathcal{M}_{j}\right]p\left(\mathcal{M}_{j}\middle|\,\mathbf{y}\right). \tag{4}$$

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

$$\hat{y}_{T+h}=E\left(y_{T+h}\middle|\,\mathbf{y}\right)=\sum_{j=1}^{M}E\left(y_{T+h}\middle|\,\mathbf{y},\mathcal{M}_{j}\right)p\left(\mathcal{M}_{j}\middle|\,\mathbf{y}\right), \tag{5}$$

where $E\left(y_{T+h}\middle|\,\mathbf{y},\mathcal{M}_{j}\right)$ 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 (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

$$\mathbf{y}_{T\times1}=\left[\begin{array}{c}\mathbf{y}^{*}_{m\times1}\\ \widetilde{\mathbf{y}}_{l\times1}\end{array}\right], \tag{6}$$

where the first part of the data is used to convert the parameter priors $p\left(\boldsymbol{\theta}_i\middle|\,\mathcal{M}_i\right)$ 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 $\tilde{\mathbf{y}}=(y_{m+1},y_{m+2},\ldots,y_T)'$, conditional on $\mathbf{y}^*=(y_1,y_2,\ldots,y_m)'$ and model $\mathcal{M}_i$, is

$$p\left(\tilde{\mathbf{y}}\middle|\,\mathbf{y}^{*},\mathcal{M}_{i}\right)=\int_{\boldsymbol{\theta}_i}L\left(\tilde{\mathbf{y}}\middle|\,\boldsymbol{\theta}_i,\mathbf{y}^{*},\mathcal{M}_{i}\right)p\left(\boldsymbol{\theta}_i\middle|\,\mathbf{y}^{*},\mathcal{M}_{i}\right)d\boldsymbol{\theta}_i, \tag{7}$$

where $p\left(\boldsymbol{\theta}_i\middle|\,\mathbf{y}^{*},\mathcal{M}_{i}\right)$ is the posterior distribution of the parameters and $L\left(\tilde{\mathbf{y}}\middle|\,\boldsymbol{\theta}_i,\mathbf{y}^{*},\mathcal{M}_{i}\right)$ 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 $\tilde{\mathbf{y}}$, the expression (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 $p(\tilde{\mathbf{y}}|\, \mathbf{y}^*,\mathcal{M}_i)$.

By replacing the marginal likelihood in (1) with the posterior predictive density (7), the posterior model probabilities, or weights, can be expressed as

$$p(\mathcal{M}_i|\, \tilde{\mathbf{y}}, \mathbf{y}^*) = \frac{p(\tilde{\mathbf{y}}|\, \mathbf{y}^*,\mathcal{M}_i)\, p(\mathcal{M}_i)}{\sum_{j=1}^{M} p(\tilde{\mathbf{y}}|\, \mathbf{y}^*,\mathcal{M}_j)\, p(\mathcal{M}_j)}. \tag{8}$$

The forecast combination based on predictive likelihood weights is then obtained by substituting $p(\mathcal{M}_i|\, \tilde{\mathbf{y}}, \mathbf{y}^*)$ in (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 $\tilde{\mathbf{y}}$ in (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. $\tilde{\mathbf{y}} = \mathbf{y}$, $\mathbf{y}^* = \varnothing$, which yields the marginal density, $m(\mathbf{y})$, of the data.

2. $\tilde{\mathbf{y}} = \{y_r\}$, $\mathbf{y}^* = \mathbf{y}_{-r} = (y_1, y_2, \ldots, y_{r-1}, y_{r+1}, \ldots, y_T)$, which yields the cross-validation density $p(y_r|\, \mathbf{y}_{-r}, \mathcal{M}_i)$, as in Stone (1974) or Geisser (1975).

3. $\tilde{\mathbf{y}}$ contains usually two or three observations, $\mathbf{y}^* = \mathbf{y} - \tilde{\mathbf{y}}$, extending the point 2, as in Peña and Tiao (1992).

4. $\tilde{\mathbf{y}} = \mathbf{y}$, $\mathbf{y}^* = \mathbf{y}$, which yields the posterior predictive density defined in Aitkin (1991).

5. $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{y}^*$, $\mathbf{y}^* = \left(y_1, y_2, \ldots, y_{[\rho T]}\right)$, where [ ] 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. $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{y}^*$, $\mathbf{y}^*$ is a minimal subset, i.e. the least number of data points such that $p(\boldsymbol{\theta}_i|\, \mathbf{y}^*, \mathcal{M}_i)$ is a proper density, 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 $p(\boldsymbol{\theta}_i|\, \mathbf{y}^*, \mathcal{M}_i)$ and therefore the posterior predictive density $p(\tilde{\mathbf{y}}|\, \mathbf{y}^*, \mathcal{M}_i)$ does not diverge and can be calculated.

We adopt a combination of the approaches suggested in O'Hagan (1995) and in Berger and Pericchi (1996) using the sample split in (6). 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 $\tilde{\mathbf{y}}$.

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.

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 $\mathcal{M}_j$

$$BF_{ij} = \frac{P(\mathcal{M}_i|\mathbf{y})}{P(\mathcal{M}_j|\mathbf{y})} \Big/ \frac{P(\mathcal{M}_i)}{P(\mathcal{M}_j)} = \frac{m(\mathbf{y}|\,\mathcal{M}_i)}{m(\mathbf{y}|\,\mathcal{M}_j)}. \tag{9}$$

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

$$\frac{P(\mathcal{M}_i)}{P(\mathcal{M}_j)} PBF_{ij}(\tilde{\mathbf{y}}|\,\mathbf{y}^*), \tag{10}$$

where

$$PBF_{ij}(\tilde{\mathbf{y}}|\,\mathbf{y}^*) = \frac{p(\tilde{\mathbf{y}}|\,\mathbf{y}^*, \mathcal{M}_i)}{p(\tilde{\mathbf{y}}|\,\mathbf{y}^*, \mathcal{M}_j)}. \tag{11}$$

O'Hagan (1995) defines (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 (9) and that the PBF does not depend on arbitrary constants when improper priors are used.

3.1 Asymptotic Bayes factors

Using a Laplace approximation we can write the Bayes factor as

$$BF_{ij}(\mathbf{y}) \approx \frac{L\left(\mathbf{y}|\,\hat{\boldsymbol{\theta}}_i, \mathcal{M}_i\right) p\left(\hat{\boldsymbol{\theta}}_i\right) \left|-\mathcal{H}\left(\hat{\boldsymbol{\theta}}_i\right)^{-1}\right|^{\frac{1}{2}}}{L\left(\mathbf{y}|\,\hat{\boldsymbol{\theta}}_j, \mathcal{M}_j\right) p\left(\hat{\boldsymbol{\theta}}_j\right) \left|-\mathcal{H}\left(\hat{\boldsymbol{\theta}}_j\right)^{-1}\right|^{\frac{1}{2}}} \left(\frac{T}{2\pi}\right)^{\frac{k_j-k_i}{2}}, \tag{12}$$

for large T. In expression (12), $\hat{\boldsymbol{\theta}}_i$ is the maximum likelihood estimator under model $\mathcal{M}_i$, and $\mathcal{H}\left(\hat{\boldsymbol{\theta}}_i\right) = \left(-\frac{\partial \log L(\mathbf{y}|\boldsymbol{\theta}_i, \mathcal{M}_i)}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_i'}\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_i}$ is the observed Hessian matrix and k_i is the number of parameters in model $\mathcal{M}_i$. The asymptotic Bayes factor (12) can be rewritten as

$$-2\log BF_{ij}(\mathbf{y}) \approx -2\log \frac{L\left(\mathbf{y}|\hat{\boldsymbol{\theta}}_i, \mathcal{M}_i\right)}{L\left(\mathbf{y}|\hat{\boldsymbol{\theta}}_j, \mathcal{M}_j\right)} + (k_i - k_j)\log T + A\left(\hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\theta}}_j\right), \tag{13}$$

where the last term

$$A\left(\hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\theta}}_j\right) = -\log \frac{\left|-\mathcal{H}\left(\hat{\boldsymbol{\theta}}_i\right)^{-1}\right|}{\left|-\mathcal{H}\left(\hat{\boldsymbol{\theta}}_j\right)^{-1}\right|} - 2\log \frac{p\left(\hat{\boldsymbol{\theta}}_i\right)}{p\left(\hat{\boldsymbol{\theta}}_j\right)} + (k_j - k_i)\log 2\pi \tag{14}$$

is $O(1)$. For non-nested models, $\log \frac{L\left(\mathbf{y}|\hat{\boldsymbol{\theta}}_i, \mathcal{M}_i\right)}{L\left(\mathbf{y}|\hat{\boldsymbol{\theta}}_j, \mathcal{M}_j\right)}$ is $O_p(T)$ when $\mathcal{M}_i$ or $\mathcal{M}_j$ is the true model, and it follows that the Bayes factor is consistent. For nested models standard results for Likelihood ratio tests apply and consistency follows. See Gelfand and Dey (1994) for additional details. Note that dropping the term $A(\cdot)$ in (13) yields Schwarz's Bayesian information criterion

$$-2\log BF_{ij}(\mathbf{y}) \approx -2\log \frac{L\left(\mathbf{y}|\,\hat{\boldsymbol{\theta}}_i, \mathcal{M}_i\right)}{L\left(\mathbf{y}|\,\hat{\boldsymbol{\theta}}_j, \mathcal{M}_j\right)} + (k_i - k_j)\log T. \tag{15}$$

The previous results relies on the assumption that the true model is included in the model set under consideration. In a recent paper Fernández-Villaverde and Rubio-Ramírez (2004) extended this to the case when all the models are misspecified and showed that, asymptotically, the Bayes factor will select the model which minimizes the Kullback-Leibler distance to the true data density. That is, the best approximation to the true model is selected.

3.2 Asymptotic partial Bayes factors

The partial Bayes factor (11) can be expressed as

$$PBF_{ij}(\tilde{\mathbf{y}}|\,\mathbf{y}^*) = \frac{m(\mathbf{y}|\,\mathcal{M}_i)}{m(\mathbf{y}|\,\mathcal{M}_j)}\Bigg/\frac{m(\mathbf{y}^*|\,\mathcal{M}_i)}{m(\mathbf{y}^*|\,\mathcal{M}_j)}, \tag{16}$$

where $m(\cdot)$ is the marginal likelihood of the full sample, $\mathbf{y}$, and the training sample, $\mathbf{y}^*$, respectively. This gives the approximate partial Bayes factor in the following form

$$\begin{aligned} PBF_{ij}(\tilde{\mathbf{y}}|\,\mathbf{y}^*) \approx{} & \frac{L\left(\mathbf{y}|\,\hat{\boldsymbol{\theta}}_{i,\mathbf{y}}, \mathcal{M}_i\right) p\left(\hat{\boldsymbol{\theta}}_{i,\mathbf{y}}\right) \left|-\mathcal{H}\left(\hat{\boldsymbol{\theta}}_{i,\mathbf{y}}\right)^{-1}\right|^{1/2}}{L\left(\mathbf{y}|\,\hat{\boldsymbol{\theta}}_{j,\mathbf{y}}, \mathcal{M}_j\right) p\left(\hat{\boldsymbol{\theta}}_{j,\mathbf{y}}\right) \left|-\mathcal{H}\left(\hat{\boldsymbol{\theta}}_{j,\mathbf{y}}\right)^{-1}\right|^{1/2}} \left(\frac{T}{2\pi}\right)^{\frac{k_j-k_i}{2}} \\ & \times \left[\frac{L\left(\mathbf{y}^*|\,\hat{\boldsymbol{\theta}}_{i,\mathbf{y}^*}, \mathcal{M}_i\right) p\left(\hat{\boldsymbol{\theta}}_{i,\mathbf{y}^*}\right) \left|-\mathcal{H}\left(\hat{\boldsymbol{\theta}}_{i,\mathbf{y}^*}\right)^{-1}\right|^{1/2}}{L\left(\mathbf{y}^*|\,\hat{\boldsymbol{\theta}}_{j,\mathbf{y}^*}, \mathcal{M}_j\right) p\left(\hat{\boldsymbol{\theta}}_{j,\mathbf{y}^*}\right) \left|-\mathcal{H}\left(\hat{\boldsymbol{\theta}}_{j,\mathbf{y}^*}\right)^{-1}\right|^{1/2}} \left(\frac{m}{2\pi}\right)^{\frac{k_j-k_i}{2}}\right]^{-1}, \end{aligned} \tag{17}$$

where $\hat{\boldsymbol{\theta}}_{i,\mathbf{y}^*}$ is the maximum likelihood estimate of $\boldsymbol{\theta}_i$ for the training sample and $\hat{\boldsymbol{\theta}}_{i,\mathbf{y}}$ for the full sample.

We can thus write the partial Bayes factor as

$$\begin{aligned} -2\log PBF_{ij}(\tilde{\mathbf{y}}|\,\mathbf{y}^*) \approx{} & -2\log \frac{L\left(\mathbf{y}|\hat{\boldsymbol{\theta}}_{i,\mathbf{y}}, \mathcal{M}_i\right)}{L\left(\mathbf{y}|\hat{\boldsymbol{\theta}}_{j,\mathbf{y}}, \mathcal{M}_j\right)} + 2\log \frac{L\left(\mathbf{y}|\hat{\boldsymbol{\theta}}_{i,\mathbf{y}^*}, \mathcal{M}_i\right)}{L\left(\mathbf{y}|\hat{\boldsymbol{\theta}}_{j,\mathbf{y}^*}, \mathcal{M}_j\right)} \\ & + (k_i - k_j)(\log T - \log m) + A\left(\hat{\boldsymbol{\theta}}_{i,(\cdot)}, \hat{\boldsymbol{\theta}}_{j,(\cdot)}\right), \end{aligned} \tag{18}$$

with

$$\begin{aligned} A\left(\hat{\boldsymbol{\theta}}_{i,(\cdot)}, \hat{\boldsymbol{\theta}}_{j,(\cdot)}\right) ={} & -\log \frac{\left|-\mathcal{H}\left(\hat{\boldsymbol{\theta}}_{i,\mathbf{y}}\right)^{-1}\right|}{\left|-\mathcal{H}\left(\hat{\boldsymbol{\theta}}_{j,\mathbf{y}}\right)^{-1}\right|} + \log \frac{\left|-\mathcal{H}\left(\hat{\boldsymbol{\theta}}_{i,\mathbf{y}^*}\right)^{-1}\right|}{\left|-\mathcal{H}\left(\hat{\boldsymbol{\theta}}_{j,\mathbf{y}^*}\right)^{-1}\right|} \\ & - 2\log \frac{p\left(\hat{\boldsymbol{\theta}}_{i,\mathbf{y}}\right)}{p\left(\hat{\boldsymbol{\theta}}_{j,\mathbf{y}}\right)} + 2\log \frac{p\left(\hat{\boldsymbol{\theta}}_{i,\mathbf{y}^*}\right)}{p\left(\hat{\boldsymbol{\theta}}_{j,\mathbf{y}^*}\right)} = O(1). \end{aligned} \tag{19}$$

It follows that model choice based on the predictive likelihood 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) applies 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.

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

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \varepsilon, \tag{20}$$

where $\boldsymbol{\gamma} = (\alpha, \beta')'$, $\mathbf{Z} = (\iota, \mathbf{X})$ and $\boldsymbol{\varepsilon}$ is a vector of $N\left(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I}\right)$ disturbances. Partitioning $\mathbf{Z}$ conformably with (6) into the training and hold-out samples we use a g-prior for the regression parameters

$$\boldsymbol{\gamma}|\,\sigma^2 \sim N\left(0, c\sigma^2\left(\mathbf{Z}^{*\prime}\mathbf{Z}^*\right)^{-1}\right), \tag{21}$$

that is, the prior mean is set to zero indicating 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

$$p\left(\sigma^2\right) \propto 1/\sigma^2. \tag{22}$$

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

$$\begin{aligned} p\left(\tilde{\mathbf{y}}|\,\widetilde{\mathbf{Z}}\right) \propto \left(S^*\right)^{\frac{m}{2}} \left|\mathbf{I}_l + \widetilde{\mathbf{Z}}\left(\mathbf{M}^*\right)^{-1}\widetilde{\mathbf{Z}}\right|^{-\frac{1}{2}} \\ \times\left[S^* + \left(\tilde{\mathbf{y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}_1\right)'\left(\mathbf{I}_l + \widetilde{\mathbf{Z}}\left(\mathbf{M}^*\right)^{-1}\widetilde{\mathbf{Z}}\right)^{-1}\left(\tilde{\mathbf{y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}_1\right)\right]^{-(m+l)/2}. \end{aligned} \tag{23}$$

See Appendix A for further details.

A slight reformulation of the predictive density (23) is quite revealing,

$$\begin{aligned} p\left(\tilde{\mathbf{y}}\right) \propto \left(\frac{S^*}{m}\right)^{-l/2} \frac{\left|\mathbf{M}^*\right|^{\frac{1}{2}}}{\left|\mathbf{M}^* + \widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}\right|^{\frac{1}{2}}} \qquad (24) \\ \times\left[m + \frac{1}{\left(S^*/m\right)}\left(\tilde{\mathbf{y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}_1\right)'\left(\mathbf{I} + \widetilde{\mathbf{Z}}\left(\mathbf{M}^*\right)^{-1}\widetilde{\mathbf{Z}}'\right)^{-1}\left(\tilde{\mathbf{y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}_1\right)\right]^{-(m+l)/2}, \end{aligned}$$

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 $|\mathbf{M}^*|^{1/2} \Big/ \left|\mathbf{M}^* + \widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}\right|^{1/2}$. 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

$$|\mathbf{M}^*| = \left|\frac{c+1}{c}\mathbf{Z}^{*\prime}\mathbf{Z}^*\right| = \left(\frac{c+1}{c}\right)^{k+1} |\mathbf{Z}^{*\prime}\mathbf{Z}^*| \tag{25}$$

and

$$\left|\mathbf{M}^* + \widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}\right| \approx \left|\frac{c+1}{c}\mathbf{Z}^{*\prime}\mathbf{Z}^* + \frac{l}{m}\mathbf{Z}^{*\prime}\mathbf{Z}^*\right| = \left(\frac{1+c\left(1+\frac{l}{m}\right)}{c}\right)^{k+1} |\mathbf{Z}^{*\prime}\mathbf{Z}^*|. \tag{26}$$

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

$$\frac{|\mathbf{M}^*|^{1/2}}{\left|\mathbf{M}^* + \widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}}\right|^{1/2}} \approx \left(\frac{c+1}{1+c\left(1+\frac{l}{m}\right)}\right)^{\frac{k+1}{2}} \approx \left(1+\frac{l}{m}\right)^{-\frac{k+1}{2}} = \left(\frac{m+l}{m}\right)^{-\frac{k+1}{2}}. \tag{27}$$

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

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

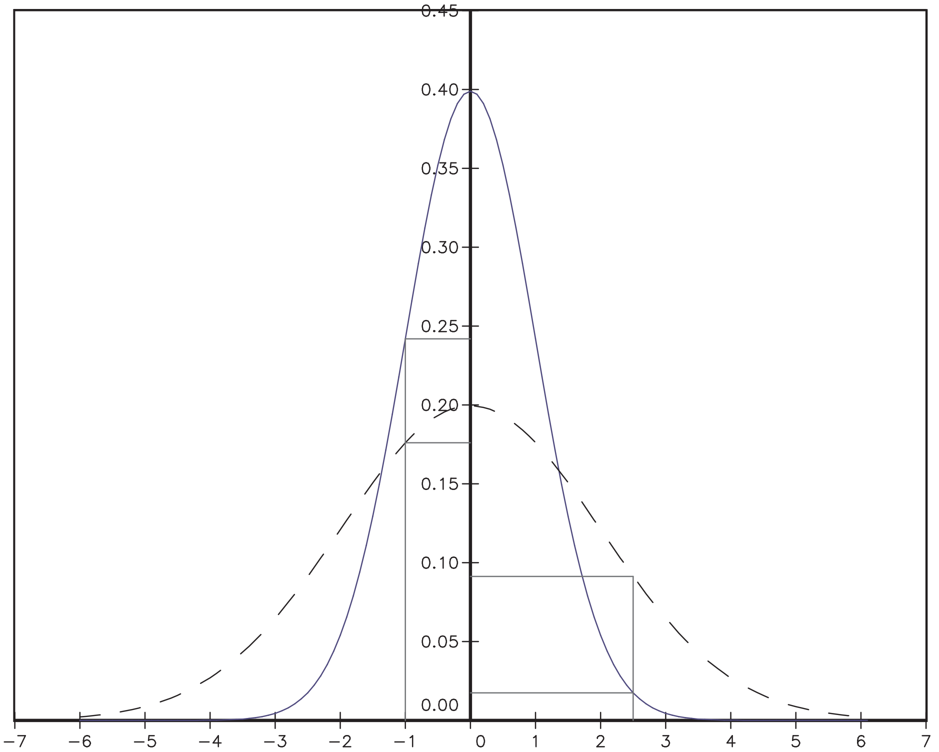
$$\left[m + \frac{1}{(S^*/m)}\left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}_1\right)'\left(\mathbf{I} + \widetilde{\mathbf{Z}}\left(\mathbf{M}^*\right)^{-1}\widetilde{\mathbf{Z}}'\right)^{-1}\left(\widetilde{\mathbf{y}} - \widetilde{\mathbf{Z}}\boldsymbol{\gamma}_1\right)\right]^{-(m+l)/2}. \tag{28}$$

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 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 $T = m + l$. 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 and then the contribution of the in-sample fit is relatively large.

Figure 1 Predictive likelihood for models with small and large prediction error variance.



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 iid standard normal and then construct the additional five variables according to

$$(\mathbf{x}_{11}, \ldots, \mathbf{x}_{15}) = (\mathbf{x}_1, \ldots, \mathbf{x}_5) \begin{pmatrix} 0.3 & 0.5 & 0.7 & 0.9 & 1.1 \end{pmatrix}' \iota + \mathbf{E}, \tag{29}$$

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

$$y_t = 4 + 2x_{1,t} - x_{5,t} + 1.5x_{7,t} + x_{11,t} + 0.5x_{13,t} + \sigma\varepsilon_t, \tag{30}$$

where the disturbances ε_t are iid 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 data sets. 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 data set 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 posterior model probabilities are calculated as in (8) or in (1) for the marginal likelihood based on the full sample. The model averaged forecasts are then formed using (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. 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 data sets we conduct additional experiments where two of the variables, x_1 and x_7, in the true model (30) 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.[3] This corresponds to the $\mathfrak{M}-$open view of Bernardo and Smith.

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 data set 20 forecasts are calculated using the individual models and the forecast combinations. The estimated Root Mean Square Forecast Error (RMSFE) of the different procedures is the average of the RMSFE from the 100 data sets.

The prior specification is the same for all experiments and similar to the one used by Eklund and Karlsson (2005). The prior on the models is given by

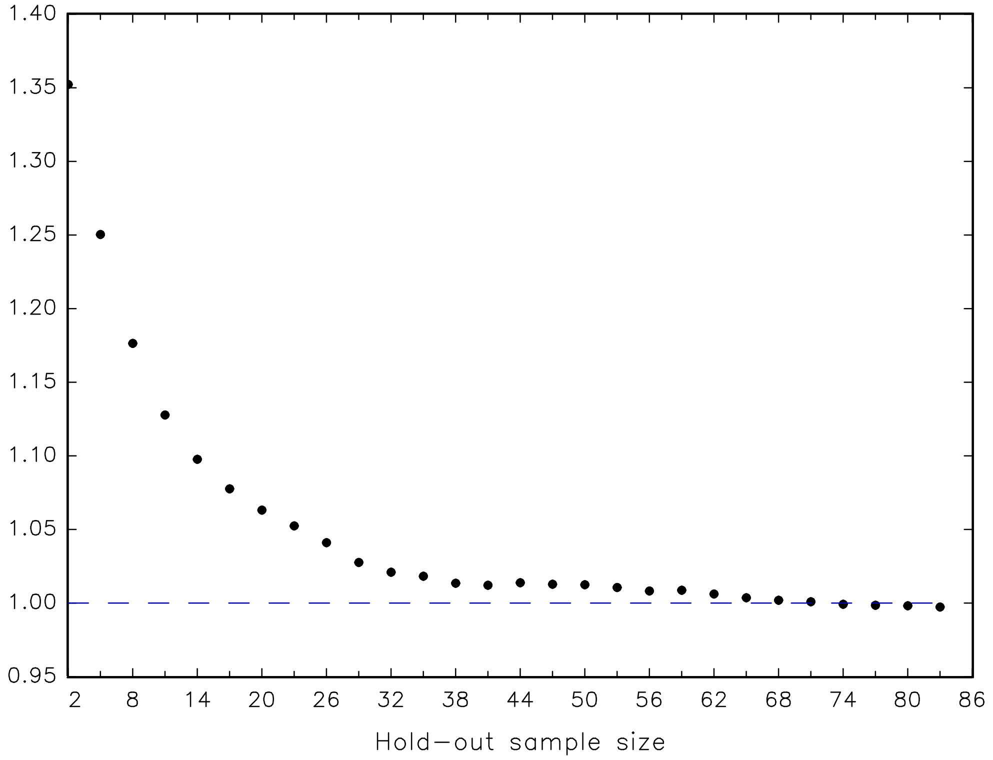
$$p(\mathcal{M}_i) \propto \delta^{k_i} (1-\delta)^{k'-k_i}, \tag{31}$$

where k_i is the number of variables included in model $\mathcal{M}_i$, the maximum number of variables 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 (21) is set to $(k')^3$.

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 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 we use it primarily as a device for identifying the set of practically

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

Figure 2 Ratio of RMSFE for predictive likelihood and marginal likelihood as a function of l for the simulated small data set.



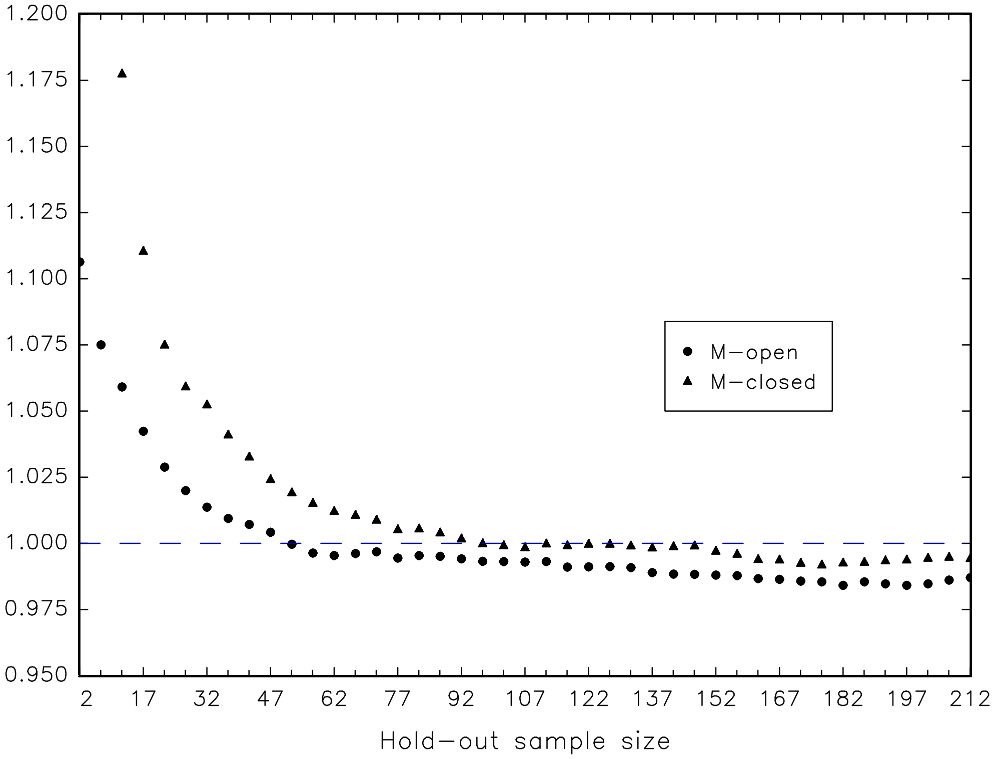
relevant models. That is, models with sufficiently large posterior probability to enter into the forecast combination in a meaningful way. To this end we take the set of relevant models to be the set of models visited by the chain and exact posterior model probabilities are calculated conditional on this set of models. 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 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.

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

For the simulated small data set the simulations include different sizes of the hold-out sample, from $l = 2$, to $l = 83$ with increments of 3. For the medium data set the hold-out sample size varies from $l = 2$, to $l = 212$ with an increment of 5. In the large data set 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 - 4 plot the ratio of the RMSFE for the predictive likelihood to the RMSFE for the marginal likelihood for the three sample sizes. All the results show that the predictive likelihood RMSFE decreases as the size of the hold-out sample increases. For the small data set the predictive likelihood provides a small but insignificant improvement on the marginal likelihood for $l \geq 74$, indicating that at least 70% of the data should be left for model

Figure 3 Ratio of RMSFE for predictive likelihood and marginal likelihood as a function of l for the simulated medium data set.



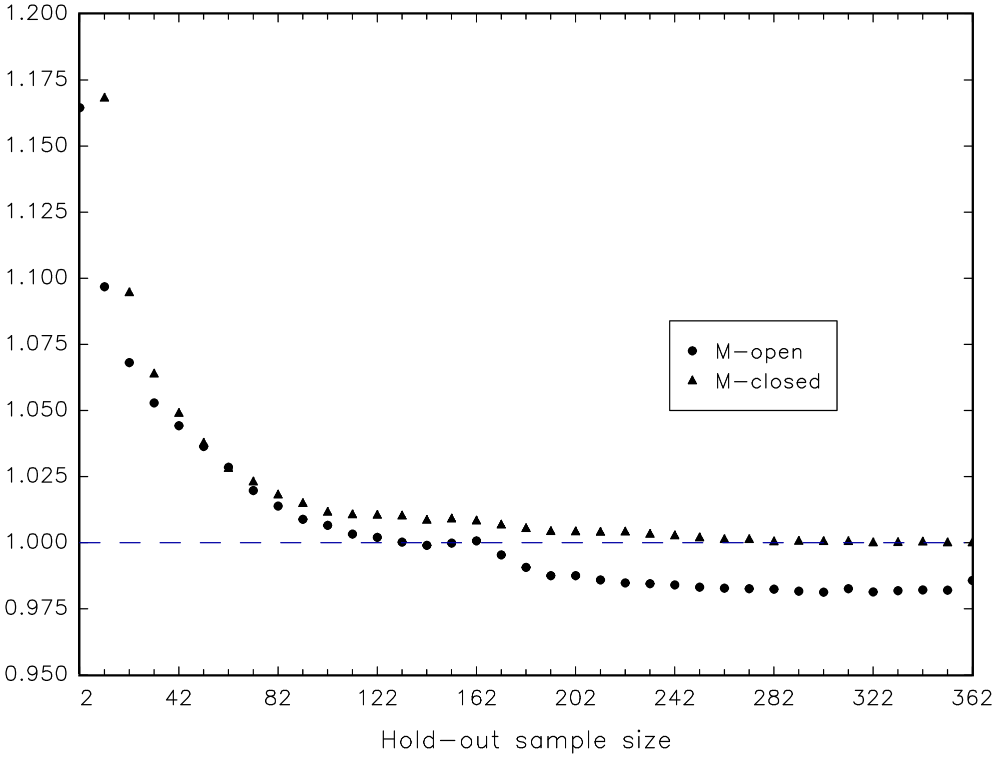
comparison.

In the case of the medium data set the predictive likelihood outperforms the marginal likelihood for $l \geq 97$ (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 data set 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 2.641, 2.527 and 2.531 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 $l = 362$, the average probability of selecting the true model is 50.24%. With this large data set, 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

Figure 4 Ratio of RMSFE for predictive likelihood and marginal likelihood as a function of l for the simulated large data set.



for $l \geq 52$ (the reduction in RMSFE is significant for $l = 107$ and $l \geq 117$) with the medium data set and $l \geq 172$ (significantly for $l \geq 182$) with the large data set, 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 5 and 6. The posterior probability of a variable i being in the model is given by

$$p\left(x_i|\,\mathbf{y}\right) = \sum_{j=1}^{M} I\left(x_i \in \mathcal{M}_j\right) p\left(\mathcal{M}_j|\,\mathbf{y}\right), \tag{32}$$

where $I\left(x_i \in \mathcal{M}_j\right)$ equals one if x_i is included in model j and zero otherwise.

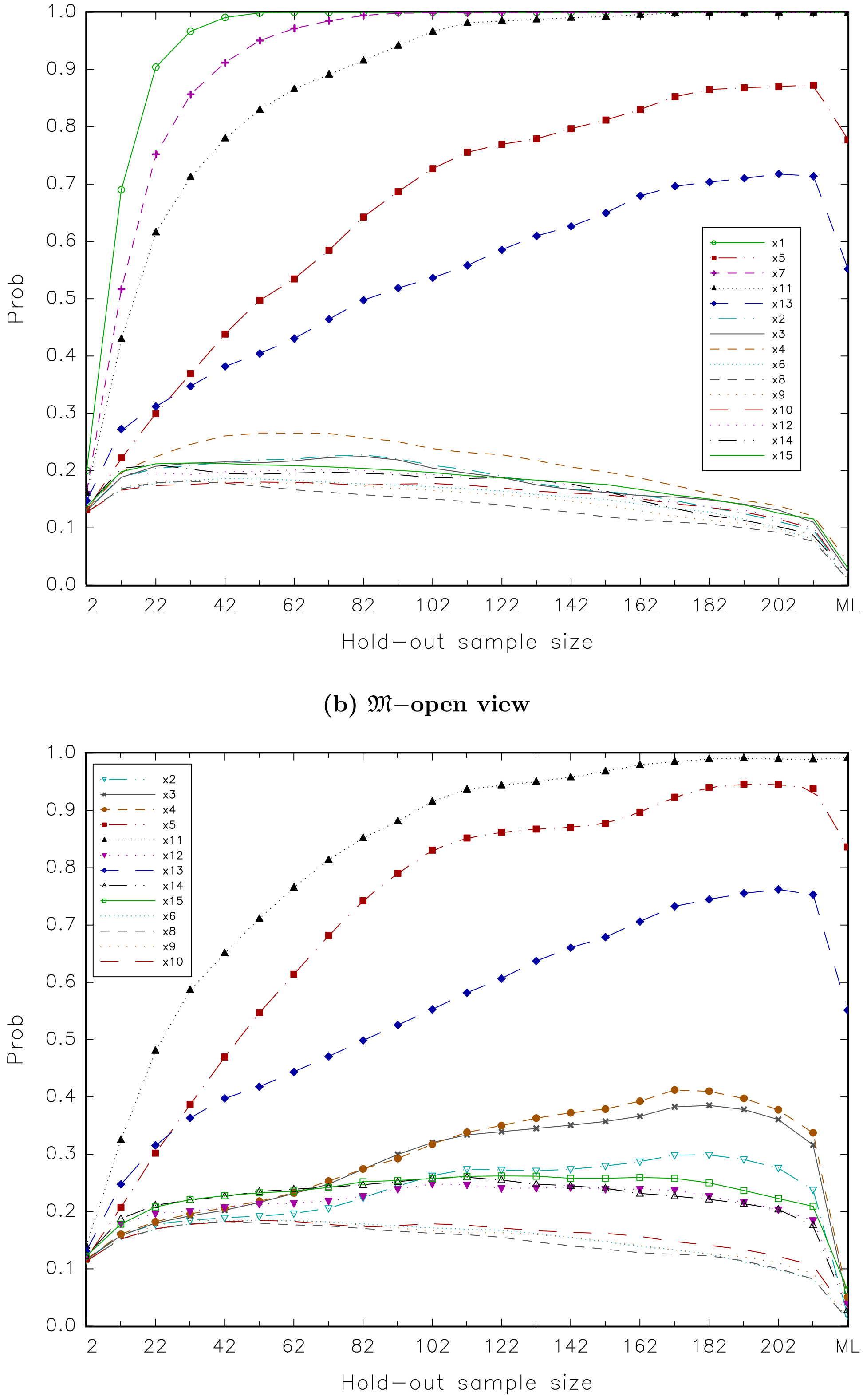
In the case when there is no true model in the model set the predictive likelihood by and large finds the approximation given by the conditional expectation,

$$\begin{aligned} y_t|\,x_{-1,-7} = & -1.034x_{2,t} - 1.448x_{3,t} - 1.862x_{4,t} - 3.276x_{5,t} + 1.414x_{11,t} \\ & + 0.414x_{12,t} + 0.914x_{13,t} + 0.414x_{14,t} + 0.414x_{15,t}. \end{aligned} \tag{33}$$

In contrast, the marginal likelihood in general only selects from the variables originally in the model. Note that the standard deviation of the prediction error from the conditional model (33) is 3.355 compared to 2.5 for the true model.

In the experiments the set of models visited by the primary 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.

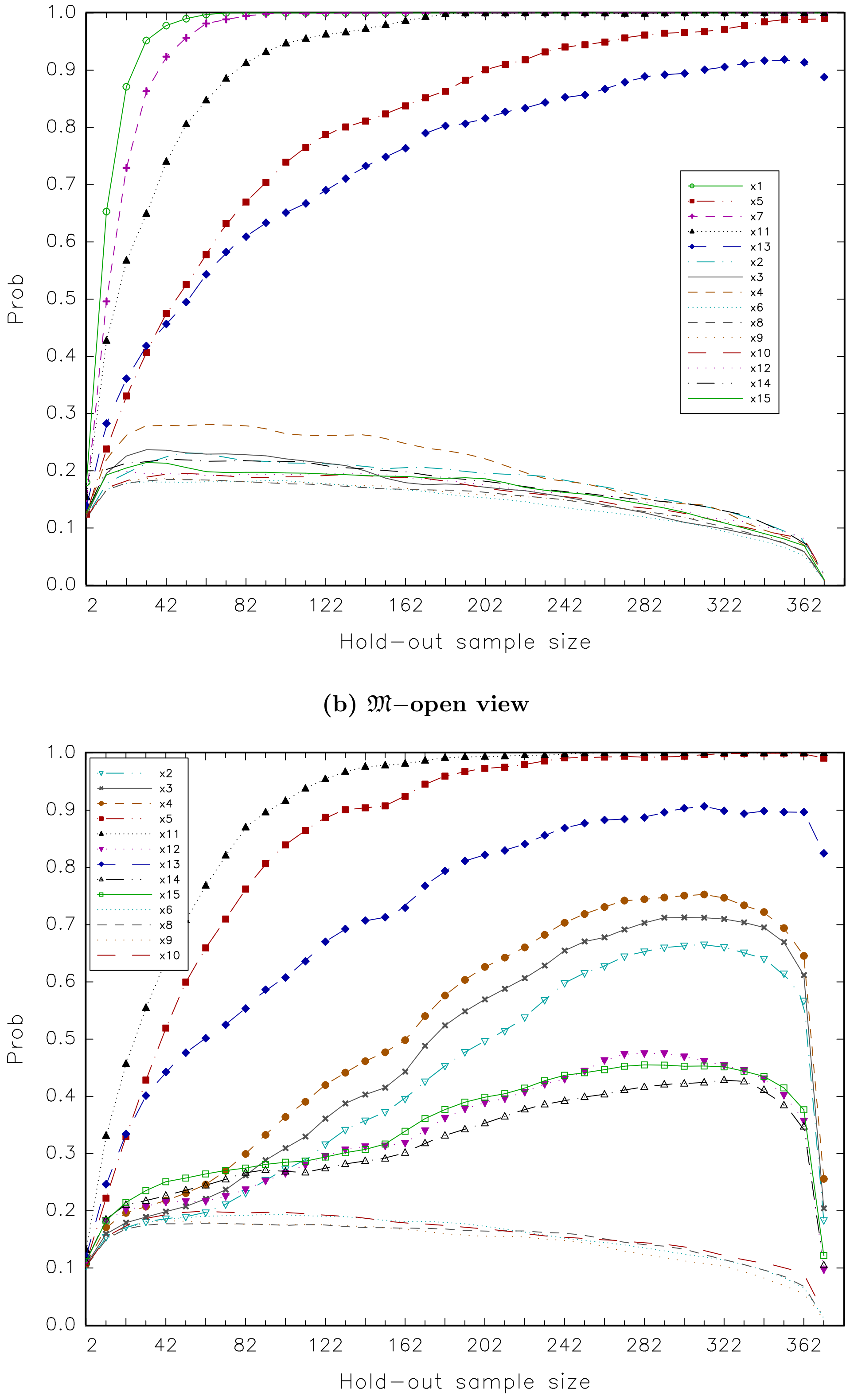
Figure 5 Variable inclusion probabilities (average) for medium data set.



(a) $\mathfrak{M}$−closed view

(b) $\mathfrak{M}$−open view

Figure 6 Variable inclusion probabilities (average) for large data set, where the last point on the horizontal axes denotes the marginal likelihood.



(a) $\mathfrak{M}$−closed view

(b) $\mathfrak{M}$−open view

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

In these experiments, executed only for the medium data set, 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 6(a). The variable inclusion probabilities for the shifting variable x_7 are plotted in Figure 6(b).

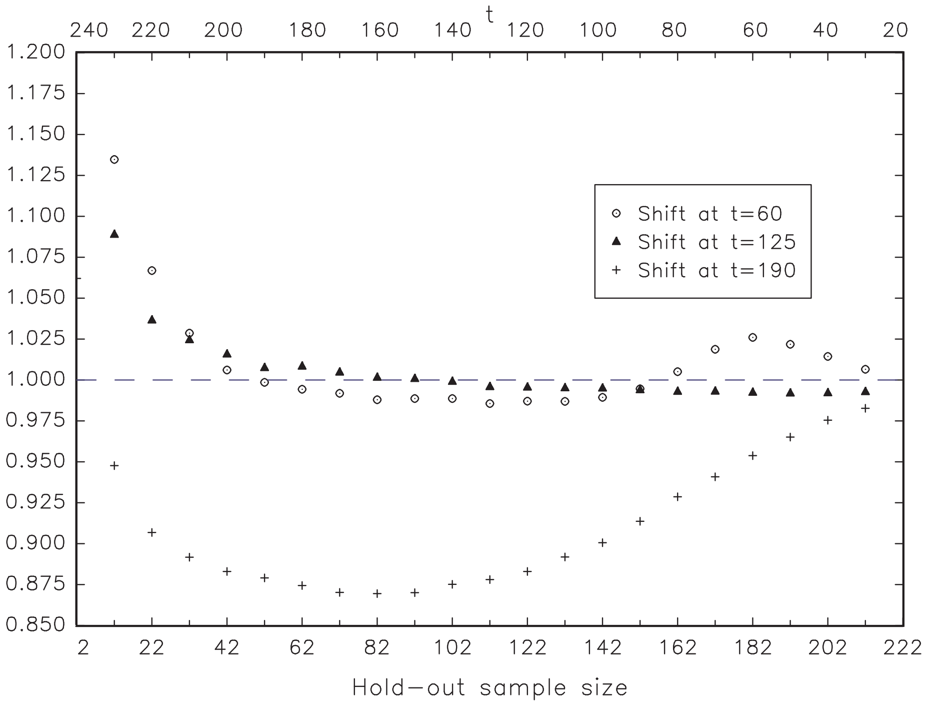
The behavior of the variable inclusion probability of x_7 depends on whether the break is in the training sample or in the hold-out sample, and on its position in the sample. In general, when the break is close to the split between the training and hold-out samples, the variable inclusion probabilities 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 sample indicate a negative value for the parameter associated with 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.

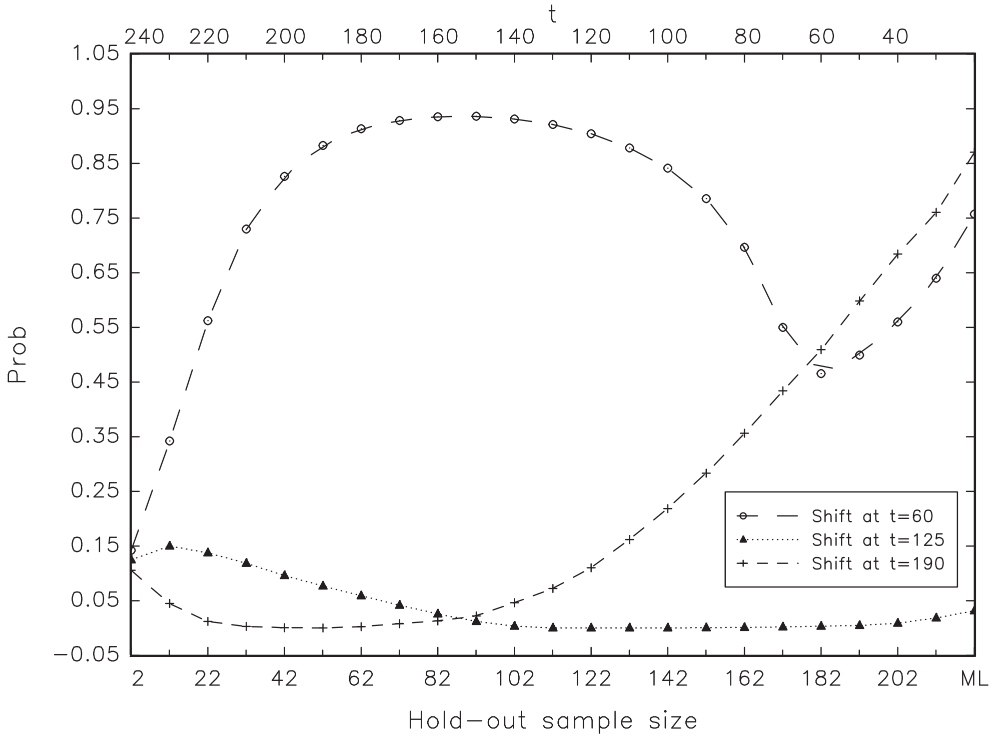
The actual forecasts are calculated using posterior distributions for the parameters that are based on the full sample up to the date of the forecast. That is, the forecasts from a given model are the same for the marginal and predictive likelihoods and the difference in forecasting performance is due to the different weights assigned to the models. 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 2.73 compared to 2.51 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 3.21 but this is still a substantial improvement on the 3.69 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 $72 \leq l \leq 142$ and significantly larger RMSFE for $l \geq 172$. In the intermediate case with a break at $t = 125$ the predictive likelihood gives a smaller RMSFE for $l \geq 152$. 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.

Figure 7 Results for the medium data set 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



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

$$y_{t+h} = \alpha + \omega d_{t+h} + \mathbf{x}_t \beta + \varepsilon_t, \tag{34}$$

with the aim to forecast h time 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[4], 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) for an in-depth discussion of the relative merits of direct forecasts models like (34) and the more traditional dynamic model with forecasts based on the chain rule of forecasting.

In essence we view (34) 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 77 predictor variables listed in Appendix D. 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 quarters ahead forecasts for the period 1999Q1 to 2003Q4.

We use the model prior (31) with the maximum number of variables set to $k' = 15$ and $\delta = 0.1$, corresponding to a prior expected model size of 7.7. For the regression parameters of each model we use the g-type prior (21) 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 data set. After this run we add 1 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) for further details on the variable selection procedure. The chain is run for 5 000 000 replicates in each run.

[4] 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.

Table 1 RMSFE of the Swedish inflation 4 quarters 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

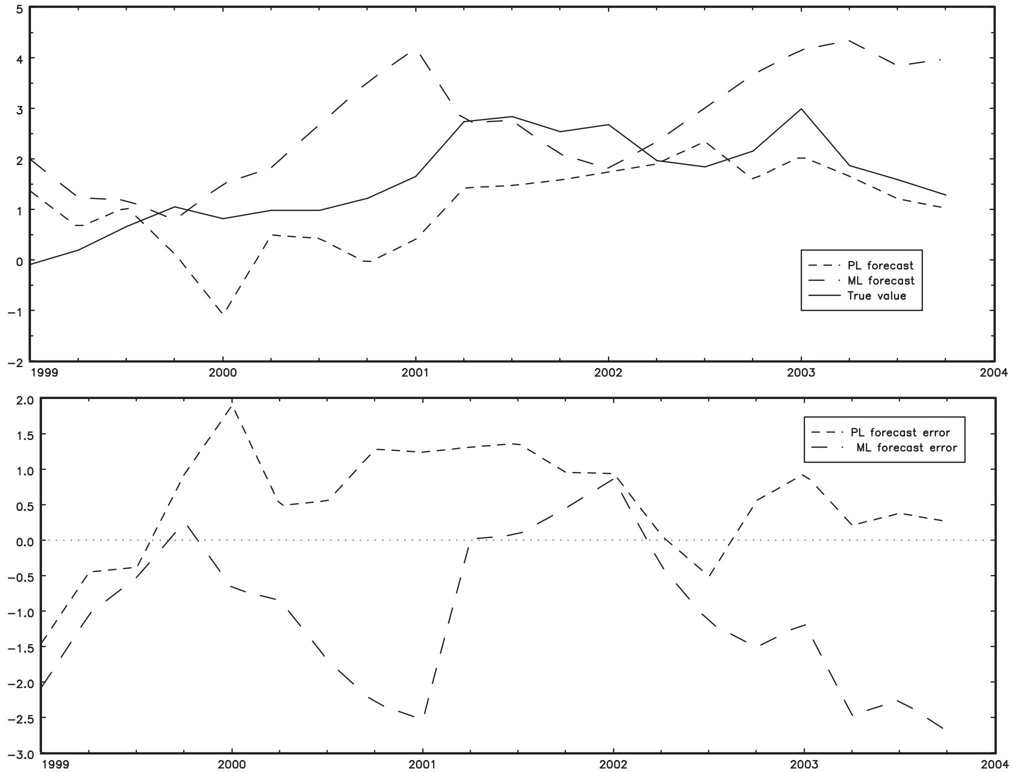
6.1 Results

As the data set is rather short, starting in 1983Q1, we are restricted in our choice of hold-out sample. 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 1. The table also includes forecasts from the 10 models with the highest posterior probabilities 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 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 gains 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 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 data set 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 does. This is confirmed by Table 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 predic-

Figure 8 Swedish inflation rate 4 quarters ahead forecasts and forecast errors, $l = 44$.



tive 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.

As an example consider Table 3 which reports the models with the highest posterior probabilities 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 posterior probability 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. One possible exception is the population variables which might be more difficult to motivate. We note that these are ranked much lower by the predictive likelihood. It is also interesting to note that variables related to current inflation and real activity are ranked higher by the predictive likelihood.

7 Conclusions

This paper 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

Table 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

analysis indicates that the weights based on the predictive likelihood will have better small sample properties than the traditional in-sample marginal likelihood. 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 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.

Table 3 Posterior model probabilities, 4 quarters ahead Swedish inflation forecast for 1999Q1.

(a) Predictive likelihood, $l = 44$

	Model				
Variable	1	2	3	4	5
`InfRel`	×	×	×		×
`InfRel`$_{-1}$				×	×
`ExpInf`	×	×	×	×	×
`R5Y`	×	×	×		×
`InfFl`	×	×		×	×
`InfFl`$_{-1}$			×		
`InfUnd`	×	×	×	×	×
`USD`	×	×	×		×
`GDPTCW`		×			×
`GDPTCW`$_{-1}$				×	
Post. Prob	0.0538	0.0301	0.0218	0.0187	0.0184

(b) Marginal likelihood

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

References

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.

Eklund, J., and S. Karlsson (2005): "Forecasting with Many Predictors," Discussion paper, Stockholm School of Economics.

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.

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.

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.

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

Appendix A Marginal and predictive likelihoods

A.1 Marginal likelihood

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

$$\begin{aligned} \mathbf{y}^* &\sim N_m\left(\mathbf{Z}^*\boldsymbol{\gamma}^*, \sigma^2\mathbf{I}_m\right), & \text{(A.1)}\\ \boldsymbol{\gamma}^* &\sim (\alpha, \boldsymbol{\beta}')', \quad (k+1)\times 1, & \text{(A.2)}\\ \mathbf{Z}^* &= (\boldsymbol{\iota}, \mathbf{X}^*), & \text{(A.3)} \end{aligned}$$

with following priors for the parameters

$$\begin{aligned} p\left(\boldsymbol{\gamma}^*|\sigma^2\right) &\sim N_{k+1}\left(0, c\sigma^2\left(\mathbf{Z}^{*\prime}\mathbf{Z}^*\right)^{-1}\right), & \text{(A.4)}\\ p\left(\sigma^2\right) &\propto \frac{1}{\sigma^2}. & \text{(A.5)} \end{aligned}$$

This yields the Normal-Inverted Gamma-2 posterior density

$$\begin{aligned} \boldsymbol{\gamma}^*|\mathbf{y}^*, \sigma^2 &\sim N_{k+1}\left(\boldsymbol{\gamma}_1, \sigma^2\left(\mathbf{M}^*\right)^{-1}\right), & \text{(A.6)}\\ \sigma^2|\mathbf{y}^* &\sim IG_2\left(S^*, m\right), & \text{(A.7)}\\ \mathbf{M}^* &= \frac{c+1}{c}\mathbf{Z}^{*\prime}\mathbf{Z}^*, & \text{(A.8)}\\ \boldsymbol{\gamma}_1 &= \frac{c}{c+1}\hat{\boldsymbol{\gamma}}^*, & \text{(A.9)}\\ S^* &= \frac{c}{c+1}\left(\mathbf{y}^* - \mathbf{Z}^*\hat{\boldsymbol{\gamma}}^*\right)'\left(\mathbf{y}^* - \mathbf{Z}^*\hat{\boldsymbol{\gamma}}^*\right) + \frac{1}{c+1}\mathbf{y}^{*\prime}\mathbf{y}^*, & \text{(A.10)} \end{aligned}$$

The marginal likelihood is then

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

A.2 Predictive likelihood

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

$$\tilde{\mathbf{y}}|\, \widetilde{\mathbf{Z}}, \mathbf{Z}^*, \mathbf{y}^*, \boldsymbol{\gamma}^*, \sigma^2 \sim N_l\left(\widetilde{\mathbf{Z}}\boldsymbol{\gamma}^*, \sigma^2\mathbf{I}_l\right), \tag{A.12}$$

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 $\tilde{\mathbf{y}}$ conditionally on $\sigma^2, \widetilde{\mathbf{Z}}, \mathbf{Z}^*, \mathbf{y}^*$ is Normal

$$\left(\begin{array}{c} \boldsymbol{\gamma}^* \\ \tilde{\mathbf{y}} \end{array}\right)\Bigg| \widetilde{\mathbf{Z}}, \mathbf{Z}^*, \mathbf{y}^*, \sigma^2 \sim N_{k+1+l}\left(\left(\begin{array}{c} \boldsymbol{\gamma}_1 \\ \widetilde{\mathbf{Z}}\boldsymbol{\gamma}_1 \end{array}\right), \sigma^2\left[\begin{array}{cc} \left(\mathbf{M}^*\right)^{-1} & \left(\mathbf{M}^*\right)^{-1}\widetilde{\mathbf{Z}}' \\ \widetilde{\mathbf{Z}}\left(\mathbf{M}^*\right)^{-1} & \mathbf{I}_l + \widetilde{\mathbf{Z}}\left(\mathbf{M}^*\right)^{-1}\widetilde{\mathbf{Z}}' \end{array}\right]\right). \tag{A.13}$$

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

As $\sigma^2 | \, \mathbf{y}^*, \mathbf{Z}^* \sim IG_2(S^*, m)$ it follows that the predictive density of $\tilde{\mathbf{y}}$ is multivariate Student and defined by

$$\tilde{\mathbf{y}} | \, \widetilde{\mathbf{Z}}, \mathbf{Z}^*, \mathbf{y}^* \sim t_l \left(\widetilde{\mathbf{Z}} \boldsymbol{\gamma}_1, S^*, \left(\mathbf{I}_l + \widetilde{\mathbf{Z}} \left(\mathbf{M}^* \right)^{-1} \widetilde{\mathbf{Z}}' \right)^{-1}, m \right) \tag{A.14}$$

with the density function

$$\begin{aligned} p\left(\tilde{\mathbf{y}} | \, \widetilde{\mathbf{Z}} \right) &= \frac{\Gamma\left(\frac{m+l}{2}\right) \left(S^*\right)^{m/2}}{\pi^{l/2} \Gamma\left(\frac{m}{2}\right) \left| \mathbf{I} + \widetilde{\mathbf{Z}} \left(\mathbf{M}^*\right)^{-1} \widetilde{\mathbf{Z}}' \right|^{1/2}} \qquad\qquad \text{(A.15)} \\ &\times \left[\left(S^*\right) + \left(\tilde{\mathbf{y}} - \widetilde{\mathbf{Z}} \boldsymbol{\gamma}_1 \right)' \left(\mathbf{I} + \widetilde{\mathbf{Z}} \left(\mathbf{M}^*\right)^{-1} \widetilde{\mathbf{Z}}' \right)^{-1} \left(\tilde{\mathbf{y}} - \widetilde{\mathbf{Z}} \boldsymbol{\gamma}_1 \right) \right]^{-T/2} \\ p\left(\tilde{\mathbf{y}} | \, \widetilde{\mathbf{Z}} \right) &\propto \frac{\left(S^*\right)^{m/2} \left| \mathbf{M}^* \right|^{\frac{1}{2}}}{\left| \mathbf{M}^* + \widetilde{\mathbf{Z}}' \widetilde{\mathbf{Z}} \right|^{\frac{1}{2}}} \qquad\qquad \text{(A.16)} \\ &\times \left[\left(S^*\right) + \left(\tilde{\mathbf{y}} - \widetilde{\mathbf{Z}} \boldsymbol{\gamma}_1 \right)' \left(\mathbf{I} + \widetilde{\mathbf{Z}} \left(\mathbf{M}^*\right)^{-1} \widetilde{\mathbf{Z}}' \right)^{-1} \left(\tilde{\mathbf{y}} - \widetilde{\mathbf{Z}} \boldsymbol{\gamma}_1 \right) \right]^{-T/2}. \end{aligned}$$

Appendix B MCMC algorithms

B.1 Predictive likelihood

Algorithm 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 $\dim\left(\boldsymbol{\theta}_{\mathcal{M}}\right)$.

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 $q\left(\mathbf{u}|\,\boldsymbol{\theta}_{\mathcal{M}}, \mathcal{M}, \mathcal{M}'\right)$.

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

4. Accept the proposed move with probability

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

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 $\dim\left(\boldsymbol{\theta}_{\mathcal{M}}\right) = \dim\left(\mathbf{u}'\right)$ 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 $p\left(\boldsymbol{\theta}_{\mathcal{M}'}|\mathbf{y}^*, \mathcal{M}'\right)$ then (B.1) simplifies to

$$\alpha = \min\left\{1, \frac{p\left(\widetilde{\mathbf{y}}|\mathbf{y}^*, \mathcal{M}'\right) p\left(\mathcal{M}'\right) j\left(\mathcal{M}|\mathcal{M}'\right)}{p\left(\widetilde{\mathbf{y}}|\mathbf{y}^*, \mathcal{M}\right) p\left(\mathcal{M}\right) j\left(\mathcal{M}'|\mathcal{M}\right)}\right\} \quad \text{(B.2)}$$

since

$$p\left(\widetilde{\mathbf{y}}|\mathbf{y}^*, \mathcal{M}'\right) = \frac{L\left(\widetilde{\mathbf{y}}|\boldsymbol{\theta}_{\mathcal{M}'}, \mathcal{M}'\right) p\left(\boldsymbol{\theta}_{\mathcal{M}'}|\mathbf{y}^*, \mathcal{M}'\right)}{p\left(\boldsymbol{\theta}_{\mathcal{M}'}|\widetilde{\mathbf{y}}, \mathbf{y}^*, \mathcal{M}'\right)}. \quad \text{(B.3)}$$

Note that this implies that we don't 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 $L\left(\tilde{\mathbf{y}}|\boldsymbol{\theta}_{\mathcal{M}'},\mathcal{M}'\right)\cdot p\left(\boldsymbol{\theta}_{\mathcal{M}'}|\mathbf{y}^{*},\mathcal{M}'\right)$ 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 ration 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	3.5707*	23	2.7788*	44	2.6768*	65	2.6500
5	3.3016*	26	2.7489*	47	2.6742*	68	2.6457
8	3.1063*	29	2.7134*	50	2.6735*	71	2.6429
11	2.9778*	32	2.6958*	53	2.6685*	74	2.6384
14	2.8984*	35	2.6886*	56	2.6623	77	2.6365
17	2.8452*	38	2.6760*	59	2.6637	80	2.6357
20	2.8071*	41	2.6726*	62	2.6569	83	2.6333
						ML	2.6406

* 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	3.6542*	57	2.5652*	112	2.5264	167	2.5111*
7	3.2287*	62	2.5576*	117	2.5246	172	2.5078*
12	2.9753*	67	2.5535*	122	2.5260	177	2.5064*
17	2.8058*	72	2.5490*	127	2.5262	182	2.5081*
22	2.7163*	77	2.5400	132	2.5246	187	2.5090*
27	2.6763*	82	2.5407	137	2.5225	192	2.5104*
32	2.6592*	87	2.5368	142	2.5238	197	2.5111*
37	2.6304*	92	2.5311	147	2.5246	202	2.5128*
42	2.6094*	97	2.5267	152	2.5194	207	2.5138*
47	2.5879*	102	2.5246	157	2.5166	212	2.5125*
52	2.5753*	107	2.5228	162	2.5116*	ML	2.5268

* 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	3.6861*	102	2.5603*	202	2.5415*	302	2.5322
12	2.9563*	112	2.5578*	212	2.5411*	312	2.5323
22	2.7704*	122	2.5573*	222	2.5412*	322	2.5308
32	2.6926*	132	2.5567*	232	2.5390	332	2.5313
42	2.6547*	142	2.5525*	242	2.5377	342	2.5316
52	2.6262*	152	2.5538*	252	2.5359	352	2.5308
62	2.6021*	162	2.5519*	262	2.5341	362	2.5309
72	2.5893*	172	2.5481*	272	2.5341		
82	2.5767*	182	2.5445*	282	2.5319	ML	2.5310
92	2.5685*	192	2.5418	292	2.5325		

* 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	4.0380*	57	3.6365	112	3.6247	167	3.6000*
7	3.9237*	62	3.6331	117	3.6170*	172	3.5978*
12	3.8657*	67	3.6357	122	3.6174*	177	3.5966*
17	3.8043*	72	3.6383	127	3.6176*	182	3.5919*
22	3.7550*	77	3.6296	132	3.6163*	187	3.5966*
27	3.7224*	82	3.6330	137	3.6094*	192	3.5940*
32	3.6996*	87	3.6319	142	3.6072*	197	3.5919*
37	3.6843	92	3.6285	147	3.6072*	202	3.5938*
42	3.6758	97	3.6250	152	3.6059*	207	3.5990*
47	3.6651	102	3.6247	157	3.6051*	212	3.6025*
52	3.6484	107	3.6240*	162	3.6011*	ML	3.6499

* 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	4.0297*	102	3.4832	202	3.4173*	302	3.3956*
12	3.7953*	112	3.4717	212	3.4117*	312	3.4001*
22	3.6960*	122	3.4675	222	3.4078*	322	3.3961*
32	3.6434*	132	3.4611	232	3.4067*	332	3.3977*
42	3.6136*	142	3.4569	242	3.4051*	342	3.3987*
52	3.5864*	152	3.4599	252	3.4021*	352	3.3984*
62	3.5591*	162	3.4629	262	3.4010*	362	3.4111*
72	3.5287*	172	3.4445	272	3.4000*		
82	3.5083*	182	3.4279*	282	3.3997*	ML	3.4605
92	3.4910	192	3.4172*	292	3.3969*		

* 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	3.6698*	62	2.7543	122	2.7344*	182	2.8421*
12	3.1435*	72	2.7478*	132	2.7339*	192	2.8307*
22	2.9555*	82	2.7370*	142	2.7410*	202	2.8101*
32	2.8497*	92	2.7389*	152	2.7554	212	2.7882*
42	2.7871	102	2.7387*	162	2.7842		
52	2.7663	112	2.7303*	172	2.8223*	ML	2.7702

* 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	3.6362*	62	2.8903*	122	2.8534	182	2.8443*
12	3.1208*	72	2.8796	132	2.8523	192	2.8431*
22	2.9711*	82	2.8708	142	2.8519	202	2.8437*
32	2.9365*	92	2.8683	152	2.8490*	212	2.8457*
42	2.9113*	102	2.8637	162	2.8458*		
52	2.8876	112	2.8543	172	2.8463*	ML	2.8660

* 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	3.9198*	62	3.2273*	122	3.2592*	182	3.5203*
12	3.4978*	72	3.2118*	132	3.2917*	192	3.5621*
22	3.3469*	82	3.2090*	142	3.3237*	202	3.5999*
32	3.2915*	92	3.2114*	152	3.3724*	212	3.6271*
42	3.2591*	102	3.2230*	162	3.4276*		
52	3.2443*	112	3.2408*	172	3.4726*	ML	3.6908

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

Appendix D 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 D.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 D.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 D.3 Money supply

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

Table D.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 D.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 D.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 D.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 D.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