Marginalized Predictive Likelihood Comparisons of Linear Gaussian State-Space Models with Applications to DSGE, DSGE-VAR, and VAR Models

Anders Warne, Günter Coenen and Kai Christoffel*

April 15, 2015

Summary: In a Bayesian setting, the predictive likelihood is of particular relevance when the objective is to rank models in a forecast comparison exercise. We discuss how the predictive likelihood can be estimated, by means of marginalization, for any subset of the observable variables in linear Gaussian state-space models and propose to utilize a missing observations consistent Kalman filter for that purpose. Based on this convenient and simple approach, we analyze euro area data and compare the density forecast performance of a DSGE model to a DSGE-VAR, a large BVAR, and a multivariate random walk model over the forecast sample 1999Q1–2011Q4. While the BVAR generally provides superior density forecasts, its performance deteriorates substantially with the onset of the Great Recession. This is particularly notable for longer-horizon real GDP forecasts, where the DSGE and DSGE-VAR models with stronger economic foundations perform better. In fact, for longer horizons the ranking of models changes over the forecast sample when the focus is on a subset of variables comprising real GDP growth, GDP deflator inflation, and the nominal interest rate.

Keywords: Bayesian inference, density forecasting, Kalman filter, missing data, Monte Carlo integration, predictive likelihood.

1. Introduction

In Bayesian analysis of (discrete) time series, the predictive likelihood can be employed to compare forecast accuracy across models. A problem occurs when the models included in

Note: This paper has previously been circulated with the titles "Predictive Likelihood Comparisons with DSGE and DSGE-VAR Models" and "Forecasting with DSGE-VAR Models". We are particularly grateful for comments from the editor and two referees, and to Marta Bańbura who specified the large Bayesian VAR model we use in the paper. We are also very grateful for discussions with Gianni Amisano (ECB and Board of Governors), Michal Andrle (IMF), Jan Brůha (Czech National Bank), Herman van Dijk (Tinbergen Institute), Juha Kilponen (Suomen Pankki), Bartosz Maćkowiak (ECB), Frank Schorfheide (University of Pennsylvania), Mattias Villani (Linköping University), and comments from members of the Working Group on Econometric Modelling, and participants of the Tinbergen Institute workshop on "Recent Theory and Applications of DSGE Models" at Erasmus University Rotterdam in June 2012, and the CEF 2012 conference in Prague. The opinions expressed in this paper are those of the authors and do not necessarily reflect views of the European Central Bank or the Eurosystem. Any remaining errors are the sole responsibility of the authors.

* All authors: Directorate General Research, European Central Bank, Sonnemannstrasse 20, 60314 Frankfurt am Main, Germany.
Corresponding author: Anders Warne, e-mail: anders.warne@ecb.europa.eu, phone: +49-6913448737.

the exercise predict different observables, but where some of these variables are shared by all models. For example, suppose there are two models to compare, where the first can predict the future values of the observables y_1, y_2, and y_3, while the second can predict the future values of y_1 and y_4. To compare the forecast performance of these two models with the predictive likelihood requires that the likelihoods can be marginalized with respect to the non-shared variables, i.e. that the future values of y_2 and y_3 are integrated out from the predictive likelihood of the first model, while y_4 is integrated out for the second model. The same basic marginalization problem arises when a forecaster or policy maker is concerned with only some of the variables that a set of models can predict.

The marginalization problem may be solved by applying textbook results when the joint predictive likelihood of a model has a known distributional form, for instance a Student t. However, models with a known predictive distribution are rare and in the typical case when this distribution is unknown, we can instead make use of the fact that the predictive likelihood is equal to the integral of the conditional likelihood times the posterior density with respect to the model parameters. Here, by conditional likelihood we refer to the predictive likelihood *conditional* on a value for the parameters. If this likelihood is based on a distribution where marginalization can be handled analytically, then the marginalization problem of the predictive likelihood can be solved at this stage. What thereafter remains to be done is to integrate out the dependence on the parameters, a topic closely related to the estimation of the marginal likelihood; see, e.g., Geweke (2005).

In this paper we suggest a convenient and simple recursive approach to marginalizing the conditional likelihood in linear Gaussian discrete-time state-space models, namely to use a missing observations consistent Kalman filter. Since this approach builds up the marginalized parts of only the relevant arrays, it is computationally simpler than the approach in, e.g. Andersson and Karlsson (2008), which involves first calculating the mean and the covariance matrix of the joint conditional distribution and thereafter reducing these to the entries relevant for the marginalized conditional likelihood; see also Karlsson (2013, Section 8.2.1), and Amisano and Geweke (2013).[1]

In Christoffel, Coenen, and Warne (2011)—henceforth, CCW—we followed the normal approximation approach suggested by Adolfson, Lindé, and Villani (2007b) for computing the marginalized predictive likelihood[2] when comparing the density forecasts of the ECB's New Area-Wide Model (NAWM) to various alternatives.[3] This pseudo out-of-sample forecast comparison exercise covered the period after the introduction of the euro and ended in 2006Q4, focusing on three nested partitions of the 12 (out of 18) observable variables that are endogenously determined in the NAWM.

In the novel application of the current paper, we enhance the forecast comparison exercise in CCW in several ways. First, while the forecast sample begins in 1999Q1, as in CCW, the end point is extended from 2006Q4 to 2011Q4, thereby allowing us to study how the models compare in terms of forecasting performance during and after the onset of the Great Recession. Second, we assess the results from using the normal approximation of the predictive likelihood to those obtained from an estimator of the predictive likelihood based on Monte Carlo integration of the marginalized conditional likelihood with respect to posterior draws of the model parameters. Furthermore, we include in this setting a DSGE-VAR with the NAWM as prior and compare the density forecast performance of

[1] The Kalman filter based approach suggested in our paper has been applied in a more recent paper by Del Negro, Hasegawa, and Schorfheide (2014).

[2] In their empirical forecasting study, Adolfson, Lindé, and Villani (2007b) discuss this marginalization problem in the context of marginal likelihoods. Specifically, the authors first note that the joint predictive likelihood (for all observables in a model over $T+1$ until $T+h$) is equal to the ratio of the marginal likelihood of the historical sample (up to time T) and the forecast sample (between time $T+1$ and $T+h$), and the marginal likelihood of the historical sample only; see equation (5.3) in Adolfson, Lindé, and Villani. While the marginal likelihood for the whole sample may be decomposed into a sequence of intermediate non-overlapping joint predictive likelihoods, they also mention that the marginal likelihood cannot be decomposed into terms of the marginal h-step-ahead predictive likelihood when $h > 1$. They therefore conclude that the marginal likelihood cannot detect whether some models perform well on certain forecast horizons while other models do better on other horizons; see the last paragraph on page 324 of Adolfson, Lindé, and Villani. Since the number of variables included in their density forecast comparison exercise is large, the authors claim that kernel density estimation of the predictive likelihood is not practical, and instead they assume that the predictive likelihood for marginal h-step-ahead forecasts is multivariate normal and estimate the mean vector and the covariance matrix from the predictive sample.

[3] See Christoffel, Coenen, and Warne (2008) for details about the NAWM, an open-economy DSGE model of the euro area.

the DSGE model to this DSGE-VAR model, as well as to a BVAR from CCW, and a multivariate random walk model estimated with Bayesian methods.

The remainder of the paper is organized as follows. Section 2 introduces notation and presents concepts related to the predictive likelihood. Section 3 focuses on linear state-space models with Gaussian innovations and shows how the conditional likelihood can be marginalized via a Kalman filter that takes into account missing observations. Section 4 presents the empirical density forecast comparison exercise, while Section 5 summarizes the main findings of the paper.

2. The Predictive Likelihood

To establish notation, let $\theta_m \in \Theta_m$ be a vector of unknown parameters of a model indexed by m, while $\mathcal{Y}_T = \{y_1, y_2, \dots, y_T\}$ is a real-valued time series of an n-dimensional vector of observables y_t. The observed value of this vector of random variables is denoted by y_t^o, while the sample of observations is similarly denoted by $\mathcal{Y}_T^o$. The observables' density function is given by $p(\mathcal{Y}_T|\theta_m, m)$, while the likelihood function is denoted by $p(\mathcal{Y}_T^o|\theta_m, m)$. Bayesian inference is based on combining a likelihood function with a prior distribution, $p(\theta_m|m)$, in order to obtain a posterior distribution of the model parameters, $p(\theta_m|\mathcal{Y}_T^o, m)$. From Bayes theorem we know that the posterior is equal to the posterior kernel (the product of the likelihood and the prior) divided by the marginal likelihood, denoted by

$$p(\mathcal{Y}_T^o|m) = \int_{\Theta_m} p(\mathcal{Y}_T^o|\theta_m, m)p(\theta_m|m)d\theta_m. \tag{1}$$

The marginal likelihood is a standard measure to compare how well different models fit the data in Bayesian analysis and for each model m it is a joint assessment of how well the prior and likelihood agrees with the observations.

Point and density forecasts are determined from the predictive density of model m. For a sequence of future values of the observable variables $\mathcal{Y}_{T,h} = \{y_{T+1}, \dots, y_{T+h}\}$, this density can be expressed as

$$p\big(\mathcal{Y}_{T,h}\big|\mathcal{Y}_T^o, m\big) = \int_{\Theta_m} p\big(\mathcal{Y}_{T,h}\big|\mathcal{Y}_T^o, \theta_m, m\big) p\big(\theta_m\big|\mathcal{Y}_T^o, m\big) d\theta_m, \tag{2}$$

where T increases as additional data points are added to the historical sample in a recursive forecast exercise. The joint predictive likelihood of model m is equal to the predictive density in (2) evaluated at the observed values $\mathcal{Y}_{T,h}^o = \{y_{T+1}^o, \ldots, y_{T+h}^o\}$.

Suppose that we are only interested in forecasting a subset of the variables $\mathcal{Y}_{T,h}$, denoted by $\mathcal{Y}_{s,T,h} = \{y_{s,T+1}, \ldots, y_{s,T+h}\}$, where $y_{s,T+i} = S_i' y_{T+i}$, and S_i is an $n \times n_i$ selection matrix with $n_i \in \{0, 1, \ldots, n\}$ for $i = 1, \ldots, h$. Hence, the subset of variables in $y_{s,T+i}$ is the same for each time period $T + i$ when i is fixed and T varies, while the subset of variables can vary for different i.[4] In addition, let $y_{-s,T+i}$ denote the remaining variables in y_{T+i}, i.e. $y_{-s,T+i} = S_{i,\perp}' y_{T+i}$ with $S_{i,\perp}' S_i = 0$, while $S_i' S_i = I_{n_i}$ and $S_{i,\perp}' S_{i,\perp} = I_{n-n_i}$.

The marginalized predictive density of $\mathcal{Y}_{s,T,h}$ can be expressed as

$$p\big(\mathcal{Y}_{s,T,h}\big|\mathcal{Y}_T^o, m\big) = \int_{\Theta_m} p\big(\mathcal{Y}_{s,T,h}\big|\mathcal{Y}_T^o, \theta_m, m\big) p\big(\theta_m\big|\mathcal{Y}_T^o, m\big) d\theta_m, \tag{3}$$

where we have made use of (2), and where

$$p\big(\mathcal{Y}_{s,T,h}\big|\mathcal{Y}_T^o, \theta_m, m\big) = \int \cdots \int p\big(\mathcal{Y}_{T,h}\big|\mathcal{Y}_T^o, \theta_m, m\big) dy_{-s,T+1} \cdots dy_{-s,T+h}. \tag{4}$$

The marginalized predictive likelihood is given by (3) when evaluated at the observed values $\mathcal{Y}_{s,T,h}^o$, while equation (4) is then called the marginalized *conditional* likelihood.

The order in which the parameters θ_m and the variables $y_{-s,t}$ are integrated out can of course be reversed when determining the left hand side of (3). However, the precise density function which determines the conditional likelihood is typically an integral part of

[4] Hence, S_i does not need to be the same as *or* have the same dimension as S_j when $i \neq j$, with the consequence that different subsets of observables may be included in $y_{s,T+i}$ and $y_{s,T+j}$ when i and j are fixed, while T may vary.

the model assumptions in a Bayesian setting, suggesting that this function may be "easier" to operate on when attempting to integrate out the dependence on the variables $y_{-s,t}$ than the joint predictive density, which will generally not be of a known type.

The marginalized predictive likelihood can alternatively be expressed as the ratio of two marginal likelihoods:

$$p\left(\mathcal{Y}^o_{s,T,h} \middle| \mathcal{Y}^o_T, m\right) = \frac{p\left(\mathcal{Y}^o_{s,T,h}, \mathcal{Y}^o_T \middle| m\right)}{p\left(\mathcal{Y}^o_T \middle| m\right)}. \tag{5}$$

The denominator is here given by equation (1), while the numerator is

$$p\left(\mathcal{Y}^o_{s,T,h}, \mathcal{Y}^o_T \middle| m\right) = \int_{\Theta_m} p\left(\mathcal{Y}^o_{s,T,h} \middle| \mathcal{Y}^o_T, \theta_m, m\right) p\left(\mathcal{Y}^o_T \middle| \theta_m, m\right) p\left(\theta_m \middle| m\right) d\theta_m. \tag{6}$$

Again we see that the computation of the marginalized predictive likelihood depends on being able to evaluate the marginalized conditional likelihood.

In the next section we suggest an approach to marginalizing the conditional likelihood in equation (4) for $\mathcal{Y}^o_{s,T,h}$ which applies to linear Gaussian state-space models and which is simple, fast, and robust. Once this likelihood has been determined, the problem of calculating the marginalized predictive likelihood in (3) for $\mathcal{Y}^o_{s,T,h}$, or via (5), generally depends on selecting an appropriate numerical method for integrating out the dependence on the parameters. Consequently, estimation of the predictive likelihood is closely related to the problem of estimating the marginal likelihood; see, e.g., Geweke (2005).

3. Marginalizing the Predictive Likelihood in Linear Gaussian State-Space Models

3.1. Marginalizing the Conditional Likelihood

Standard linear time series models—including VAR models, VARMA models, dynamic factor models, and other unobserved component models—may be cast in state-space form.

Structural models, such as log-linearized DSGE models and other linear rational expectations models, also have such a representation provided that a unique and convergent solution exists for a given value of the underlying parameter vector.[5]

To establish some further notation, let the observable variables y_t be linked to a vector of state variables ξ_t of dimension r through equation

$$y_t = \mu + H'\xi_t + w_t, \quad t = 1, \ldots, T. \tag{7}$$

The errors, w_t, are assumed to be i.i.d. $N(0, R)$, with R being an $n \times n$ positive semidefinite matrix, while the state variables are determined from a first-order VAR system:

$$\xi_t = F\xi_{t-1} + B\eta_t, \quad t = 1, \ldots, T. \tag{8}$$

The state shocks, η_t, are of dimension q and i.i.d. $N(0, I_q)$, while F is an $r \times r$ matrix, and B is $r \times q$. The parameters of this model, (μ, H, R, F, B), are all uniquely determined by θ_m. Provided that $H'\xi_t$ is stationary, the vector μ is the population mean of y_t conditional on θ_m.

The system in (7) and (8) is a state-space model, where equation (7) gives the measurement or observation equation and (8) the state or transition equation. Provided that the number of measurement errors and state shocks is large enough and an assumption about the initial conditions is added, we can calculate the likelihood function with a suitable Kalman filter; see, e.g., Harvey (1989) and Durbin and Koopman (2012) for details.[6]

Suppose that we are interested in forecasting the subset of variables $\mathcal{Y}_{s,T,h}$ with the state-space system and that we have the observed values $\mathcal{Y}^o_{s,T,h}$. The log of the marginalized conditional likelihood can now be determined via a missing observations consistent Kalman

[5] Sargent (1989) was among the first to recognize that linear rational expectations models can be expressed in state-space form.

[6] It is straightforward to generalize the system in (7) and (8) to allow for, e.g., time variation of the matrices and apply the ideas presented in this paper.

filter. We here find for $h \geq 1$

$$\log p\big(\mathcal{Y}_{s,T,h}^{o} \big| \mathcal{Y}_{T}^{o}, \theta_m, m\big) = \sum_{i=1}^{h} \log p\big(y_{s,T+i}^{o} \big| \mathcal{Y}_{s,T,i-1}^{o}, \mathcal{Y}_{T}^{o}, \theta_m, m\big), \tag{9}$$

where $\mathcal{Y}_{s,T,0}^{o}$ is empty by definition. If $n_i \geq 1$ the marginalized conditional log-likelihood value at $T + i$ is

$$\begin{aligned} \log p\big(y_{s,T+i}^{o} \big| \mathcal{Y}_{s,T,i-1}^{o}, \mathcal{Y}_{T}^{o}, \theta_m, m\big) &= -\frac{n_i}{2} \log(2\pi) - \frac{1}{2} \log \big|\Sigma_{y_s,T+i|T+i-1}\big| \\ -\frac{1}{2} \left(y_{s,T+i}^{o} - y_{s,T+i|T+i-1}\right)' \Sigma_{y_s,T+i|T+i-1}^{-1} &\left(y_{s,T+i}^{o} - y_{s,T+i|T+i-1}\right), \end{aligned} \tag{10}$$

while it is zero when $n_i = 0$.

The n_i-dimensional vector of forecasts $y_{s,T+i|T+i-1}$ and its forecast error covariance matrix $\Sigma_{y_s,T+i|T+i-1}$ are directly obtained from the missing observations consistent Kalman filter. This filter is well known and we therefore refer the interested reader to, e.g., Harvey (1989) or Durbin and Koopman (2012); see also the online appendix (Appendix C). Such an approach generates a *bottom-up* evaluation of the marginalized conditional likelihood. Since the joint conditional likelihood for $\mathcal{Y}_{T,h}^{o}$ is normal, the marginalized conditional likelihood for $\mathcal{Y}_{s,T,h}^{o}$ is also normal with mean and covariance obtained by selecting the appropriate elements from the mean vector and covariance matrix of the joint predictive distribution conditional on the parameters. This property is used by, for instance, Andersson and Karlsson (2008), when estimating the marginalized predictive likelihood for VAR models via the conditional likelihood. When h is large, such a *top-down* approach to evaluating the marginalized conditional likelihood is expected to be slower than the bottom-up Kalman filter evaluation presented above, especially when the evaluations are repeated many times. Note also that the bottom-up and top-down approaches coincide only in the case of one-step-ahead forecasts.

3.2. Estimating the Marginalized Predictive Likelihood

Once the problem of evaluating the marginalized conditional likelihood has been addressed, we proceed with the second step in the estimation of the marginalized predictive likelihood, which involves integrating out the dependence on the parameters. It is assumed below that an ergodic sequence of parameter draws are available from the posterior distribution.

In the empirical application in Section 4 we apply a simple Monte Carlo (MC) integration method to estimate the predictive likelihood. Specifically, we use:

$$\hat{p}_{MC}\big(\mathcal{Y}^o_{s,T,h}|\mathcal{Y}^o_T, m\big) = \frac{1}{N}\sum_{j=1}^{N} p\big(\mathcal{Y}^o_{s,T,h}|\mathcal{Y}^o_T, \theta^{(j)}_m, m\big), \tag{11}$$

where $\theta^{(j)}_m$ is a draw from the posterior density $p(\theta_m|\mathcal{Y}^o_T, m)$ for $j = 1, \ldots, N$.[7] Under certain regularity conditions (Tierney, 1994), the right hand side of (11) converges almost surely to the expected value of $p(\mathcal{Y}^o_{s,T,h}|\mathcal{Y}^o_T, \theta_m, m)$ with respect to $p(\theta_m|\mathcal{Y}^o_T, m)$, i.e. to the predictive likelihood $p(\mathcal{Y}^o_{s,T,h}|\mathcal{Y}^o_T, m)$. Hence, equipped with the posterior draws, the marginalized predictive likelihood can be consistently estimated from the sample average of the marginalized conditional likelihood. A further property of this estimator is that it is unbiased (see Chan and Eisenstat, 2015, Proposition 1).[8]

The MC estimator in (11) is expected to work well in practise when the posterior draws cover well enough the parameter region where the marginalized conditional likelihood is large. This is more likely to be the case when the dimension of $\mathcal{Y}_{s,T,h}$ is fairly small *and* h is not too large, but it is not obvious when one or both of these properties is not met. Clearly, when $S_i = I_n$ for all h and the latter is sufficiently large, the situation resembles the case where the marginal likelihood is estimated by averaging the likelihood over draws from the prior distribution, and such an estimator is expected to be poor.

[7] Andersson and Karlsson (2008) also use the estimator in (11) and refer to it as the Rao-Blackwellized estimate of $p(\mathcal{Y}_{s,T,h}|\mathcal{Y}^o_T, m)$; see also Karlsson (2013).

[8] The numerical standard error of (11) can be calculated using the Newey and West (1987) estimator when the posterior draws $\theta^{(j)}_m$ and $\theta^{(j+1)}_m$ are correlated.

As an alternative, one may therefore consider standard methods for calculating the marginal likelihood, such as harmonic mean estimators; see Gelfand and Dey (1994), the truncated normal version in Geweke (2005), or the extension to a truncated elliptical in Sims, Waggoner, and Zha (2008). Such estimators require two sets of posterior draws to be consistent. Namely, apart from the $\theta_m^{(j)}$ draws based on the data $\mathcal{Y}_T^o$, we would also need posterior draws $\theta_m^{(i)}$ using the data $(\mathcal{Y}_{s,T,h}^o, \mathcal{Y}_T^o)$. The former posterior draws are used to estimate the marginal likelihood in (1), the latter posterior draws are needed to estimate the marginal likelihood in (6), while the marginalized predictive likelihood is estimated by the ratio of these two estimates; see equation (5). As a consequence, harmonic mean estimators of the marginalized predictive likelihood are more time consuming than the MC estimator in (11). However, since the harmonic mean estimators take the data $\mathcal{Y}_{s,T,h}^o$ into account, they are less likely to be hampered by a large dimension of the set of predicted values or a long forecast horizon. Other methods, such as bridge sampling may also be considered; see, e.g., Meng and Wong (1996). Evaluating these options is beyond the scope of this paper, but it is an important topic for future reasearch.

4. Comparing Forecast Accuracy: An Application to Euro Area Data

In this section we compare marginalized h-step-ahead density forecasts for three subsets of observables and across four linear Gaussian state-space models for euro area data using the approach discussed in Section 3. In Section 4.1 we discuss the models: a DSGE model, a DSGE-VAR model, a large BVAR model, as well as a multivariate random walk model. In Section 4.2, we present the forecast sample and summarize the empirical results of the exercise. The calculations below have to a large extent been performed with the help of YADA, a Matlab program for Bayesian estimation and evaluation of DSGE and DSGE-VAR models; see Warne (2015) for details.

4.1. Empirical Models

4.1.1. The New Area-Wide Model of the Euro Area

Since the turn of the century, we have witnessed the development of a new generation of DSGE models that build on explicit micro-foundations with optimizing agents. Major methodological advances allow the estimation of variants of these models that are able to compete, in terms of data coherence, with more standard time series models, such as VARs; see, among others, the empirical models in Christiano, Eichenbaum, and Evans (2005), Smets and Wouters (2003, 2007), and Adolfson, Laséen, Lindé, and Villani (2007a). Efforts have also been undertaken to bring these models to the forecasting arena with promising results; see, for example, CCW, Del Negro and Schorfheide (2013), Smets, Warne, and Wouters (2014), Wolters (2015), and references therein.

In our application, we extend the forecasting comparison exercise in CCW using the New Area-Wide Model (NAWM) of the ECB. With E_t being the rational expectations operator, the NAWM, like other log-linearized DSGE models, can be written as:

$$A_{-1}\xi_{t-1} + A_0\xi_t + A_1 E_t \xi_{t+1} = D\eta_t, \quad t = 1, 2, \dots, T, \tag{12}$$

where η_t is a q-dimensional vector with i.i.d. standard normal structural shocks ($\eta_t \sim N(0, I_q)$), while ξ_t is an r-dimensional vector of model variables. The matrices A_i $(r \times r)$, with $i = -1, 0, 1$, and D $(r \times q)$ are functions of the vector of DSGE model parameters. Provided that a unique and convergent solution of the system (12) exists at a particular value of these parameters,[9] we can express the model as the first order VAR system in (8).

The NAWM is a micro-founded open-economy model of the euro area designed for use in the ECB/Eurosystem staff projections and for policy analysis; see Christoffel, Coenen, and Warne (2008) for details. The development of this DSGE model has been guided by a

[9] See, e.g., Anderson (2010), Klein (2000), or Sims (2002) for methods of solving linear rational expectations models.

principal consideration, namely to provide a comprehensive set of core projection variables, including a number of foreign variables, which, in the form of exogenous assumptions, play an important role in the projections. As a consequence, the scale of the NAWM—compared with a typical DSGE model—is rather large.

Christoffel et al. (2008) adopt the empirical approach outlined in Smets and Wouters (2003) and An and Schorfheide (2007) and estimate the NAWM with Bayesian methods, using time series for 18 macroeconomic variables. The estimation sample in Christoffel et al. (2008) is given by the period 1985Q1 until 2006Q4, with 1980Q2–1984Q4 serving as training sample.

In our application, we extend the sample to 2011Q4, also taking into account the changes in country composition of the euro area, and reestimate the model with the new data, using the same training sample as in CCW and Christoffel et al. (2008). The time series for the sample 1985Q1–2011Q4 are displayed in Figure 1, where real GDP, private consumption, total investment, exports, imports, the GDP deflator, the consumption deflator, the import deflator, nominal wages, foreign demand, and foreign prices are all expressed as 100 times the first difference of their logarithm. All other variables are expressed in logarithms except for the short-term nominal domestic and foreign interest rates. A number of further transformations are made to ensure that variable measurement is consistent with the properties of the NAWM's balanced-growth path and in line with the underlying assumption that all relative prices are stationary; see Christoffel et al. (2008) for details.

4.1.2. A DSGE-VAR Model with the NAWM as Prior

VAR models have played a central role in the development of empirical macroeconomics since the seminal article by Sims (1980). One reason for this success is that they highlight the importance of a multivariate dynamic specification for macroeconomic analysis, letting all observable variables be treated as endogenous. The VAR model of the NAWM observables y_t can be written as:

$$y_t = \Phi_0 + \sum_{j=1}^{p} \Phi_j y_{t-j} + \epsilon_t, \quad t = 1, \ldots, T, \tag{13}$$

where $\epsilon_t \sim N(0, \Sigma_\epsilon)$ and with Σ_ϵ being an $n \times n$ positive definite matrix. The vector Φ_0 is $n \times 1$, while Φ_j is $n \times n$ for $j = 1, \ldots, p$. We assume that initial values of y_t exists for $t = 0, \ldots, 1 - p$.

The parameters of a VAR model are given by $(\Phi_0, \Phi_1, \ldots, \Phi_p, \Sigma_\epsilon)$, provided that the prior distribution of the VAR does not include additional unknown parameters. BVAR models (see, e.g., Del Negro and Schorfheide, 2011, or Karlsson, 2013) typically include a number of hyperparameters that are calibrated by the researcher and are therefore included in the model index m rather than among θ_m.[10] However, a well-known example of a VAR model which includes additional parameters through the prior is a DSGE-VAR.

An early attempt to combine DSGE models with VARs is Ingram and Whiteman (1994), where the VAR parameters are expressed as a function of the DSGE model parameters. A prior for the DSGE model parameters then implies a prior for the VAR parameters through a first-order Taylor expansion of the mapping. This idea was considerably enriched by Del Negro and Schorfheide (2004), where the prior distribution of the VAR model parameters conditional on the DSGE model parameters is specified such that the conditional moments of y_t are determined through the implied first and second population moments for a given value of the DSGE model parameters. A prior distribution for *all* parameters is thereafter obtained by multiplying this conditional prior by the marginal prior of the DSGE model parameters.

In addition, DSGE-VARs are indexed by the parameter λ, which determines the weight on the prior relative to the data. The VAR approximation of the DSGE model resides

[10] See Giannone, Lenza, and Primiceri (2012) for a novel approach to making inference about the informativeness of the prior distribution of BVARs.

at one end of its range ($\lambda = \infty$), an unrestricted VAR at the other end ($\lambda = 0$), and in between these two extremes a large number of models exist. Consequently, the DSGE-VAR provides an approach for assessing the degree of misspecification, by relaxing the cross-equation restrictions of the DSGE model, and where lower values of λ suggest a greater degree of misspecification; see Del Negro, Schorfheide, Smets, and Wouters (2007).

In our application, the DSGE-VAR model is taken from Warne, Coenen, and Christoffel (2013) and has the largest marginal likelihood among all pairs (λ, p) that they considered, i.e. we let $(\lambda, p) = (2.5, 2)$.[11]

4.1.3. BVAR and Random Walk Models

We also consider a Bayesian VAR model for the same observable variables y_t as in the NAWM. The usefulness of BVARs of the Minnesota-type for forecasting purposes has long been recognized, as documented early on by Litterman (1986), and such models are therefore natural benchmarks in forecast comparisons. Below, we employ the same large BVAR as in CCW, estimated using the methodology in Bańbura, Giannone, and Reichlin (2010). This approach relies on using dummy observations when implementing the normal-inverted Wishart version of the Minnesota prior. Moreover, the prior mean of the parameters on the first *own* lag of the endogenous variables (diagonal of Φ_1) are either unity, if the variable is measured in log-levels or levels, or zero if it is measured in log first differences. That is, the prior mean supports random walks for all variables in log-levels or

[11] The common unit-root technology trend in the NAWM implies a number of cointegration relations among the levels of the observables, such as the consumption-output and investment-output "great ratios". In principle, it may be worthwhile to include these theoretically founded cointegration relations in the VAR model, as in Del Negro et al. (2007), thus forming a vector error correction model (VECM) instead of a VAR. However, the euro area data of the great ratios are trending and, thus, suggest that adding the cointegration relations will result in a worse fit for a DSGE-VECM than for a DSGE-VAR. With this in mind, we have opted to not include a DSGE-VECM in the forecast comparison exercise below. The interested reader may also consider the forecasting exercise in Adolfson, Lindé, and Villani (2007b) based on euro area data. In their study, the DSGE model which includes the cointegration relations as observables and a Bayesian VECM based on the same cointegration relations lead to inferior forecasting performance compared with models that exclude these long-run relations, although the deterioration is more pronounced for the DSGE model; see also Adolfson, Laséen, Lindé, and Villani (2008).

levels. In CCW, this large BVAR is referred to as the model with a mixed prior. A more detailed description of this BVAR is also found in the online appendix (Appendix B).

The last model we shall consider is a multivariate random walk for the vector y_t with the NAWM observables. We employ a standard diffuse prior for the covariance matrix of the random walk innovations. That is, the vector $y_t - y_{t-1} = \varepsilon_t$ is i.i.d. $N(0,\Omega)$, where Ω is an $n \times n$ positive definite matrix of unknown parameters, and $p(\Omega) \propto |\Omega|^{-(n+1)/2}$. One advantage of this model is that it allows for an analytical determination of the predictive density. For marginal h-step-ahead forecasts of $y_{s,T+h}$ the predictive density is given by a n_h-dimensional Student t-distribution with mean $S_h' y_T^o$, covariance matrix

$$\frac{h}{T-n-1} \sum_{t=1}^{T} S_h' \big(y_t^o - y_{t-1}^o\big)\big(y_t^o - y_{t-1}^o\big)' S_h,$$

and $T - n + n_h$ degrees of freedom; see the online appendix (Appendix A) for details.

4.2. Density Forecasts

A forecast comparison exercise is naturally cast as a decision problem within a Bayesian setting and therefore needs to be based on a particular preference ordering. Scoring rules can be used to compare the quality of probabilistic forecasts by giving a numerical value using the predictive distribution and an event or value that materializes.

A widely used scoring rule that was suggested by, e.g., Good (1952) is the log predictive score. Based on the predictive likelihood of $\mathcal{Y}_{s,T,h}^o$, it can be expressed as

$$S_h(m) = \sum_{t=T}^{T+T_h-1} \log p(\mathcal{Y}_{s,t,h}^o | \mathcal{Y}_t^o, m), \quad h = 1, \ldots, h^*, \tag{14}$$

where T_h is the number of time periods the h-step-ahead predictive likelihood is evaluated. The log predictive score is optimal in the sense that it uniquely determines the model ranking among all local and proper scoring rules; see Gneiting and Raftery (2007) for a survey on scoring rules. However, there is no guarantee that it will pick the same model as the forecast horizon or the selected subset of variables changes.

When comparing the density forecasts of the NAWM, the DSGE-VAR, the large BVAR, and the multivariate random walk model below we will evaluate the log predictive score in (14) with realizations for different subsets of the observables $\mathcal{Y}_{s,T,h}^{o} = y_{s,T+h}^{o}$. Hence, the predictive likelihood for each model and time period is marginalized with respect to the forecast horizon and the variables included in a subset.

The first pseudo out-of-sample forecasts are computed for 1999Q1—the first quarter after the introduction of the euro—while the final period is 2011Q4. The maximum forecast horizon is eight quarters, yielding 52 quarters with one-step-ahead forecasts and 45 quarters with eight-step-ahead forecasts. We shall only consider forecasts of quarterly growth rates for the variables in first differences, while CCW also studied forecasts of annual growth rates for such variables. The Kalman filter based calculations can be adjusted to handle such transformations of the observables; see Warne (2015, Section 12.6.1).

Concerning the selection of variables in the subsets of the observables we follow CCW and exclude the variables which are essentially exogenous in the NAWM. That is, we do not compare density forecasts which include the five foreign variables (foreign demand, foreign prices, foreign interest rate, competitors' export prices, and oil prices) and government consumption. For the remaining 12 variables we examine three nested subsets. The smallest subset is called the *small selection* and given by real GDP growth, GDP deflator inflation, and the short-term nominal interest rate. This selection may be regarded as the minimum set of variables relevant to a meaningful analysis of monetary policy, and we shall also study density forecasts of the individual variables in this selection. The second case covers a *medium selection* with the seven variables studied in Smets and Wouters (2003). In addition to the variables in the small selection, this selection covers private consumption, total investment, employment, and nominal wages. Finally, the *large selection* has 12 variables, given by the medium selection plus exports, imports, the import price deflator, the private consumption deflator, and the real effective exchange rate.

4.2.1. Empirical Evidence using the MC Estimator

The log predictive scores based on the MC estimator of the marginal h-step-ahead predictive likelihood are shown in the upper part of Figure 2 for our entire forecast sample from 1999Q1 to 2011Q4, the three variable selections defined above, eight forecast horizons $h = 1, \ldots, 8$, and our four models.[12,13] For the NAWM and the DSGE-VAR model we have used 10,000 posterior draws among the available 500,000 post burn-in draws for each model and time period when calculating the log predictive likelihood. These draws have been selected as draw number $1, 51, \ldots, 499951$ to combine modest computational costs with a lower correlation between the draws and a sufficiently high estimation accuracy. This procedure yields estimates of the log predictive likelihood that typically are accurate up to and including the first decimal. We shall discuss the numerical standard errors in Section 4.2.3.

Direct sampling is possible for the BVAR model through its normal-inverted Wishart posterior and we have used 50,000 draws from its posterior distribution when computing the predictive likelihood with the MC estimator. In the case of the random walk model, the predictive likelihood for a selection of variables is multivariate Student t and can therefore be computed from its analytical expression.

When comparing the NAWM with the DSGE-VAR, it is noteworthy that the latter model generally obtains higher log scores for all horizons and variable selections. At the longer horizons, the NAWM obtains values that are near those of the DSGE-VAR and, in the case of the small selection, even slightly higher for the eight-quarter-ahead forecasts.

[12] The log predictive scores estimated with the normal approximation are shown in the lower part of Figure 2 and are discussed below in Section 4.2.2.

[13] While the BVAR and the random walk are reestimated each quarter, the NAWM and the DSGE-VAR are both reestimated on an annual basis over the forecast sample. Specifically, the latter two models are reestimated once the end point of the historical sample is Q4 in a given calendar year.

Hence, it seems that taking into account possible misspecification of the NAWM through a DSGE-VAR improves forecasting performance, especially at the shorter horizons.

It is also worth pointing out that the random walk model is competitive with the NAWM for the one-step-ahead forecasts, especially for the small selection. As the forecast horizon increases, however, the random walk model's performance worsens considerably.

Compared with the BVAR model, the NAWM is outperformed for the large and medium selections and all forecast horizons, while in the case of the small selection the forecast performance of the BVAR deteriorates relative to the NAWM (and the DSGE-VAR) as the forecast horizon increases. In fact, in this case the BVAR performs worse than the NAWM for the five- to eight-quarter-ahead density forecasts. At the same time, the BVAR is the best performer over all horizons when using the large and the medium selection, with the DSGE-VAR in second place. However, it performs worse than the DSGE-VAR for the small selection and all forecast horizons beyond two quarters.

CCW identify two main factors that may explain the relative strengths and weaknesses of the NAWM. On the one hand, its explicit microfoundations give rise to a parsimoniously parameterized structure with a large number of cross-equation restrictions. This is potentially an advantage for achieving forecast accuracy. On the other hand, the embedded balanced growth path assumption means that the model's ability to deal with differing trends in the observables is limited compared to VAR models. In fact, this assumption may induce a bias in the forecasts and CCW report that this bias is particularly important in the case of variables connected with the labor share, which is trending downward over the forecast sample. Specifically, the NAWM systematically overpredicts nominal wage growth and underpredicts the private consumption deflator when the forecast sample ends in 2006Q4. Since real wage growth is systematically overpredicted, real private consumption growth is also overpredicted. These systematic forecast errors primarily affect the NAWM's performance for the large and medium selections, and these findings are also valid in our application where the forecast sample ends in 2011Q4. By relaxing the cross-equation restrictions of the NAWM, the DSGE-VAR model is consequently able to achieve better density forecasts than the NAWM for both of these selections.[14]

In Figure 3 the recursive average log predictive scores until 2011Q4 are displayed for all models based on the large, medium and small selections and the one, two, four, and the eight-quarter-ahead forecasts. For the large and medium selections we find that the ranking of models over the various horizons is not greatly affected by the choice of the sample end point. It is interesting to note that the average log predictive scores of the NAWM and the DSGE-VAR are fairly constant. In view of the Great Recession in late 2008 and early 2009, the drop in forecast performance of these two models is quite small for all the selections.

By constrast, the forecasting performance of the BVAR deteriorates substantially with the onset of the Great Recession. This loss in performance is quite remarkable, especially in the case of the small selection with the result that the model looses its first rank position over the longer forecast horizons. In addition, there are several time periods during 2002–2003 where the performance of the BVAR drops for the longer horizons and for all selections. At the onset of both these periods, the largest modulus of the BVAR at the posterior mean increases from values around 0.98 to values above unity. In the first instance, this sign of explosiveness also lasts until 2005, thus covering a period over which the forecasting performance of the BVAR is actually improving.[15]

To gain more insight into possible explanations for the reversal in model ranking for the small selection, we next turn to Figure 4. The focus of the graphs in this figure is a simple decomposition of the log predictive score (likelihood) into a conditional score for GDP deflator inflation and the short-term nominal interest rate given real GDP growth,

[14] Plots of the mean forecast paths from all models and all variables in the large selection are available from the authors on request.

[15] A graph of the recursive posterior mean based largest modulus values of the BVAR is available from the authors on request.

and the marginal score for real GDP growth. It can be seen from these graphs that the drop in forecasting performance for the BVAR in connection with the Great Recession is mainly explained by a loss in real GDP growth forecasting performance.

To confirm this claim, we also display all the average marginal log predictive scores for the three variables in the small selection in Figure 5. The graphs in this figure show that the loss in forecasting performance of the BVAR at the onset of the Great Recession in 2008Q4 is most severe in the case of real GDP growth.[16]

For the other three models, the average forecast performance of the small selection is also affected by the Great Recession, but here the impact is less striking than for the BVAR. In the case of the NAWM and the DSGE-VAR, it is interesting to note that the performance loss also seems mainly related to the real GDP growth forecasts. In view of the sharp fall in euro area real GDP growth in 2008Q4 and 2009Q1 compared with the behavior of inflation and short-term nominal interest rates, this finding is hardly surprising, but the fact that the "more structural models" are less sensitive than the reduced form BVAR model to such an event is indeed an interesting result.

This outcome may be explained by the fact that the BVAR prior of the constant term (Φ_0) is diffuse; see the online appendix (Appendix B). By contrast, the constant vector of the NAWM (μ) is calibrated, while the DSGE-VAR has a proper prior for the constant term. In the case of the BVAR, the posterior mean of the constant term for real GDP growth fluctuates around 0.3 before the Great Recession, falls sharply to a minimum close to zero in 2009Q2, and gradually recovers thereafter to about 0.25 at the end of the sample. It is possible that a proper prior for the constant term may render its posterior mean to be less sensitive to the recorded output loss in late 2008 and early 2009 and could therefore result in better density forecasts of real GDP growth over this period. While it may be

[16] Concerning the loss in performance for the BVAR around 2002-2003, this appears to be mainly related to poor forecasts of GDP deflator inflation, although also for this variable density forecasts of the BVAR drop at the onset of the Great Recession, albeit not by as much as in the case of real GDP growth.

tempting to change the prior of the BVAR and thereby possibly reduce the sensitivity of its estimated constant term to the Great Recession, it is also conceivable that the diffuse prior has its advantages during the more tranquil periods of the forecast sample. In any event, we prefer to restrict the specification of the BVAR model in our application to follow the methodology of Bańbura, Giannone, and Reichlin (2010) for large BVARs.

4.2.2. Evidence Based on the Normal Approximation

It was suggested by Adolfson, Lindé, and Villani (2007b) to approximate the marginalized predictive likelihood with a normal density with mean and covariance matrix taken from the predictive density. While such an approximation is not necessary when we know how to estimate the marginalized predictive likelihood, it can nevertheless serve as a tool for enhancing our understanding of the results in Section 4.2.1. In view of the assumption that the conditional likelihood is normal, the normal approximation is a natural reference point as any deviation from normality is due to the impact of the posterior parameter distribution on the predictive likelihood. Consequently, the size of the errors from using a normal approximation relative to the MC estimator is a relevant matter when the latter is accurate.[17]

The mean and covariance matrix of the predictive density in (2) can be estimated directly from the posterior draws when the mean and covariance matrix of the predicted variables conditional on the historical data and the parameters exist. Let these moments be denoted by $E[\mathcal{Y}_{T,h}|\mathcal{Y}_T^o, \theta_m, m]$ and $C[\mathcal{Y}_{T,h}|\mathcal{Y}_T^o, \theta_m, m]$, respectively. The mean of the predictive density is then given by

$$E\big[\mathcal{Y}_{T,h}\big|\mathcal{Y}_T^o, m\big] = E_T\Big[E\big[\mathcal{Y}_{T,h}\big|\mathcal{Y}_T^o, \theta_m, m\big]\Big], \tag{15}$$

[17] See Section 4.2.3 for some details on the numerical accuracy of the MC estimator.

where E_T denotes the expectation with respect to the posterior $p(\theta_m|\mathcal{Y}_T^o, m)$. The covariance matrix can likewise be expressed as

$$C\big[\mathcal{Y}_{T,h}|\mathcal{Y}_T^o, m\big] = E_T\Big[C\big[\mathcal{Y}_{T,h}|\mathcal{Y}_T^o, \theta_m, m\big]\Big] + C_T\Big[E\big[\mathcal{Y}_{T,h}|\mathcal{Y}_T^o, \theta_m, m\big]\Big], \tag{16}$$

and C_T denotes the covariance with respect to the posterior. It should be stressed that these estimates of the mean and the covariance of the predictive distribution do not rely on the assumption that the latter distribution is normal.[18]

The normal approximation of the predictive likelihood for the observables $\mathcal{Y}_{s,T,h}^o$ can be computed from sample estimates of the moments in (15) and (16). This approximation also provides a simple way of decomposing the predictive likelihood into a term reflecting forecast errors and a term driven by forecast uncertainty. The mean and covariance matrix of $\mathcal{Y}_{s,T,h}$ are determined by selecting the proper elements of (15) and (16), respectively. Next, notice that

$$\log \hat{p}_N\big(\mathcal{Y}_{s,T,h}^o|\mathcal{Y}_T^o, m\big) = -\frac{\bar{n}}{2}\log\big(2\pi\big) + D_{s,T,h}(m) + Q_{s,T,h}(m), \tag{17}$$

where $\bar{n} = \sum_{i=1}^h n_i$, whereas

$$D_{s,T,h}(m) = -\frac{\log\Big|C\big[\mathcal{Y}_{s,T,h}|\mathcal{Y}_T^o, m\big]\Big|}{2}, \tag{18}$$

$$Q_{s,T,h}(m) = -\frac{\epsilon_{s,T,h}^{o\prime}(m) C\big[\mathcal{Y}_{s,T,h}|\mathcal{Y}_T^o, m\big]^{-1}\epsilon_{s,T,h}^o(m)}{2}, \tag{19}$$

and $\epsilon_{s,T,h}^o(m)$ is the vector of prediction errors for the realizations $\mathcal{Y}_{s,T,h}^o$. The forecast uncertainty term is given by $D_{s,T,h}(m)$ in (18), while $Q_{s,T,h}(m)$ in (19) gives the impact

[18] It is also worth noting that, according to equation (16), the covariance matrix of the predictive density is decomposed into two terms, where the first term on the right hand side reflects shock uncertainty over the forecast horizon (as well as uncertainty about unobserved variables up to period T) and the second term parameter uncertainty; see Adolfson, Lindé, and Villani (2007b). Geweke and Amisano (2014) more generally refer to the first term as the intrinsic variance of $\mathcal{Y}_{T,h}$ and the second term as the extrinsic variance.

of the quadratic standardized forecast errors on the normal approximation of the log predictive likelihood. This decomposition may be of particular interest when the difference between the normal approximation and the MC estimator is small, and it may then reveal whether forecast uncertainty (18) or forecast errors (19) is responsible for the ranking of models.

The log predictive scores estimated with the normal approximation are displayed in the lower part of Figure 2. It is noteworthy how similar these graphs look when compared with the log predictive scores computed with the MC estimator in the upper part of the figure. The numerical differences between the MC estimator in (11) and the normal approximation in (17) for all models, forecast horizons, and selections of variables over the entire forecast sample are documented in Table 1. The differences between the MC estimator and the normal approximation of the log predictive score for the NAWM and the DSGE-VAR are positive for all forecast horizons and variable selections. In this table we also report, within parentheses, the differences between the two estimators when the quarters 2008Q4 and 2009Q1 are excluded from the calculations. As can be seen from the table these two quarters are particularly detrimental for the overall approximation errors for the BVAR model, but they also have a notable impact on approximation errors for the NAWM and the DSGE-VAR. In view of our findings above regarding the onset of the Great Recession, these results suggest that the accuracy of the normal approximation suffers when the value of the predictive likelihood is very low. Furthermore, a large fraction of the differences are positive suggesting that the normal approximation tends to be downwardly biased relative to the MC estimator (and the analytically determined value for the random walk model).[19]

Since the normal approximation overall provides a good approximation of the MC estimator of the predictive likelihood, it may be of interest to utilize equations (17)–(19) to assess if the ranking of the models is driven by forecast uncertainty or by forecast errors.

[19] The numerical values for the log predictive likelihood for all periods, models, and horizons are available from the authors on request.

To this end, the contribution of the forecast uncertainty term to the recursive estimates of the average log predictive score are depicted in Figure 6. Analogously, the quadratic standardized forecast error part of the recursive average log predictive scores are displayed in Figure 7.

Starting with the forecast uncertainty term in Figure 6 it can be seen that for all depicted models and forecast horizons it is weakly upward sloping over the forecast sample and that the slope is roughly equal across the four models. This indicates that overall forecast uncertainty is slowly decreasing as data points are added to the historical sample. In general, the values for the BVAR model are roughly a few log-units higher in each period than for the second group of models, given by the DSGE-VAR and the NAWM. Overall, the ranking of models based on the forecast uncertainty term is constant across time periods, variable selections, and forecast horizons, with the exception of the small selection for the NAWM and the random walk model.

Turning to the quadratic standardized forecast error term in Figure 7, it can be seen that the time variation of the recursively estimated average log predictive score is primarily due to the forecast errors. This is not surprising since the covariance matrix of the predictive distribution changes slowly and smoothly over time while the forecast errors are more volatile. Second, the ranking of the models is to some extent reversed, particularly with the BVAR having much larger quadratic standardized forecast errors than the other models. The reversal in rankings for the forecast error term can also be understood from the behavior of second moments, where a given squared forecast error yields a larger value for this term when the uncertainty linked to the forecast is smaller. Nevertheless, when compared with the forecast uncertainty term in Figure 6 the differences between the models are generally smaller for the forecast error term. This suggests that the model ranking based on the log predictive score is primarily determined by the second moments of the predictive distribution in this application. However, in the case of the small selection, at the onset of the Great Recession the h-quarters-ahead forecast errors with $h \geq 4$ of the BVAR are so severe that the model looses out to the two DSGE-based models.

4.2.3. Numerical Precision of the MC Estimates

The precision of the MC estimator of the log predictive likelihood may be assessed with the usual numerical standard error based on within posterior sample information.[20] Posterior sampling of the parameters of the NAWM and of the DSGE-VAR has been conducted with the random walk Metropolis sampler (see, e.g., An and Schorfheide, 2007) and the draws are therefore (highly) correlated, while the BVAR allows for direct sampling so that its parameter draws are independent. Accordingly, the numerical standard errors of the point estimates for the NAWM and the DSGE-VAR can be obtained with the Newey and West (1987) estimator, while those of the BVAR exclude autocorrelation.[21]

We typically find that the standard errors of the log predictive likelihood are the largest for the large selection of the observed variables and the smallest for the small selection. The standard errors of the log predictive likelihood for the NAWM, DSGE-VAR, and the BVAR before 2008Q4 are small and generally below 0.05 also for the eight-quarter-ahead forecasts and the large selection. At the onset of the Great Recession when the log predictive likelihood is small, these standard errors increase, especially for the BVAR model. On average, the standard errors are the smallest for the NAWM and the largest for the BVAR and they increase with the forecast horizon. For instance, the average standard

[20] As suggested by one of the referees, an alternative is to compute the predictive likelihood estimates from multiple posterior samples and estimate the numerical standard errors across these samples.

[21] We have used the standard weights $w(l, L) = 1 - l/(L + 1)$, $l = 1, 2, \ldots, L$, for the autocovariances in the expression of the Newey and West (1987) estimator, where $L = 97$ when $N = 10000$. Our results are robust with respect to the choice of L with only minor fluctuations of the standard errors over the range $\{50, 51, \ldots, 500\}$. The standard error of the log predictive likelihood is computed with the delta method and is therefore based on the standard error of the predictive likelihood. The estimated autocovariances in the expression of the Newey and West estimator therefore rely on the conditional likelihood values of the N posterior draws. In our case, the N draws are not consecutive but taken 50 steps apart. The autocorrelation among the selected draws from the Markov chain is thereby considerably reduced compared with the case of consecutive draws. We have chosen to use a subset of the posterior draws for the NAWM and the DSGE-VAR since it significantly speeds up the calculations, without leading to substantial problems for the numerical precision. Our estimates are also robust to very small values for the conditional likelihood.

error of the eight-step-ahead forecasts of the NAWM is about 0.02 for the large selection, 0.03 for the DSGE-VAR, and 0.08 for the BVAR.

To further document our findings, Table 2 reports the point estimates of the log predictive likelihood and the numerical standard errors for the case when the historical data ends in 2007Q4 and the forecasts are performed over the horizon 2008Q1–2009Q4. It is striking that the largest standard errors are obtained at the onset of the Great Recession in 2008Q4 and, in particular, in 2009Q1. The estimated log predictive likelihood values are also the smallest for these quarters. For example, in the case of the large selection the estimated value for the BVAR is approximately -61.2 in 2009Q1 with a standard error of close to unity, which is also the largest recorded standard error over all selections, models, time periods, and forecast horizons. By contrast, the log predictive likelihood of the NAWM is estimated at -40.0 for this quarter, with a standard error of about 0.1. These two predictive likelihood values may be compared with the average log predictive score for five-quarter-ahead forecasts which is -15.7 for the BVAR and -17.9 for the NAWM.

In summary, we find that the numerical precision of the MC estimator of the log predictive likelihood is satisfactory. For very small values of the log predictive likelihood, the precision falls, especially for models with a large number of parameters, but it is unlikely that it will impair the validity of the ranking of models since such small values are rare. At the same time, the ranking is based on the log predictive score, i.e., the sum of the log predictive likelihood, and we have not attempted to assess the numerical accuracy of this measure. As a simple but rough rule of thumb for the large selection one may multiply the average standard error (approximated to be at most 0.1) by the number of forecasts (about 50), suggesting that models with a log predictive score differing by less than 5 log-units are approximately as good or as bad in this study. This strongly supports the view that our model ranking results, as documented in e.g. Figure 2, are robust to estimation error.

5. Summary and Conclusions

As pointed out by Geweke and Amisano (2010, p. 217), the predictive likelihood function

> "... lies at the heart of Bayesian calculus for posterior model probabilities, reflecting the logical positivism of the Bayesian approach: a model is as good as its predictions."

Accordingly, the predictive likelihood can be applied to rank models in a forecast comparison exercise via the log predictive score.

This paper discusses how the predictive likelihood can be computed, by means of marginalization, for any subset of the observable variables in linear Gaussian discrete-time state-space models estimated with Bayesian methods. Moreover, it is shown that the marginalized predictive likelihood can be computed recursively via a missing observations consistent Kalman filter. Such an approach builds up the marginalized parts of only the relevant arrays and is therefore—except when dealing with one-step-ahead forecasts—simpler and faster than first calculating the mean and the covariance matrix of the joint predictive likelihood conditional on the parameters (conditional likelihood), and thereafter reducing these arrays to the entries of the set of variables to be predicted before evaluating the marginalized conditional likelihood (as in, e.g., Andersson and Karlsson, 2008). Based on a value of the conditional likelihood for each posterior draw of the model parameters, the paper considers simple Monte Carlo (MC) integration over the posterior draws to estimate the marginalized predictive likelihood.

In the empirical application with four linear Gaussian time series models, the MC estimator of the predictive likelihood is compared with a normal approximation, constructed from the mean vector and the covariance matrix of the predictive distribution. The application is an extension of the CCW study for euro area data and compares the results for the NAWM, a DSGE-VAR model with the NAWM as prior, a large BVAR, and a multivariate random walk model. The DSGE-VAR model was not included in CCW and is used to relax the possibly misspecified cross-equation restrictions of the NAWM, while the random walk model is an extension of the random walk model in CCW to a Bayesian framework. In addition, the forecast sample is extended by five years to include the Great Recession and ends in 2011Q4.

In terms of model ranking, the log predictive score (the sum of the log predictive likelihood over the forecast sample 1999Q1–2011Q4) typically favors the BVAR model, with the DSGE-VAR model improving somewhat on the density forecasts of the NAWM, especially at the shorter horizons. The random walk model, on the other hand, is only competitive with the NAWM at the one-step-ahead horizon, especially for the variable selection with real GDP growth, GDP deflator inflation, and the short-term nominal interest rate only, a subset of variables called the small selection above.

It is noteworthy that for the longer-term forecasts and the small selection, the BVAR not only looses its first rank to the DSGE-VAR at the onset of the Great Recession in 2008Q4, but also the second rank to the NAWM. The main reason for this appears to be the deterioration in the BVAR density forecasts of real GDP growth, compared with those of the NAWM and the DSGE-VAR. In other words, the "more structural" models seem to cope better with the substantial loss in output growth observed during the Great Recession than the reduced form BVAR model.

Finally, we find that the assumption of a normal predictive density provides a good approximation of the predictive likelihood when examining the density forecasts for the four models. One useful aspect of the normal approximation is that it can be employed to decompose the average log predictive score into a term that represents forecast uncertainty and another term that represents the quadratic standardized forecast errors. Using this decomposition, we show that the model ranking in our forecast comparison exercise is generally determined by forecast uncertainty. However, in the case of the change in model ranking for the small selection at the onset of the Great Recession, the decomposition reveals that this change is driven by the forecast error term.

While we have shown that the MC estimator is numerically reliable in our application, alternative estimators may be preferable in other density forecasting situations. For example, one may consider harmonic mean estimators of the marginalized predictive likelihood, cross entropy, or bridge sampling methods. Evaluating these options is beyond the scope of this paper, but it ought to be addressed in future research.

Although we have only considered linear Gaussian models that can be written in state-space form, this already covers a large number of the models frequently used in applied macroeconomics. The basic idea that has been presented for computing the conditional likelihood through a missing observations consistent Kalman filter can, in principle, be extended to nonlinear and nonnormal models. For such models, the marginalized conditional likelihood may be estimated with a suitable missing observations consistent particle filter (sequential Monte Carlo); see, e.g., Giordani, Pitt, and Kohn (2011) for a survey on filtering in state-space models, or Durbin and Koopman (2012, Chapter 12) for an introduction to particle filtering. Whether or not this leads to a reliable approach for computing the marginalized conditional likelihood in such models, however, is an open and important question for future research.

TABLE 1. Differences in log predictive score between the MC estimator and the normal approximation for the large, medium, and small selections of variables over the evaluation period 1999Q1–2011Q4.

	Large selection (12 variables)				Medium selection (7 variables)				Small selection (3 variables)			
Horizon	NAWM	DSGE-VAR	BVAR	RW	NAWM	DSGE-VAR	BVAR	RW	NAWM	DSGE-VAR	BVAR	RW
1	2.54	4.81	9.24	14.32	1.70	2.06	2.57	6.56	1.00	0.88	2.13	2.38
	(1.79)	(2.82)	(3.49)	(12.18)	(1.87)	(2.06)	(0.73)	(5.98)	(1.05)	(0.70)	(0.20)	(1.63)
2	3.28	10.62	14.29	19.82	2.28	4.30	7.41	9.26	1.61	2.07	5.35	3.29
	(1.44)	(4.38)	(2.33)	(15.93)	(2.25)	(3.29)	(2.06)	(8.19)	(1.44)	(1.10)	(0.02)	(2.34)
3	3.76	8.93	13.07	18.38	2.77	4.62	8.82	9.20	1.84	2.09	7.59	2.80
	(1.43)	(4.58)	(2.74)	(16.18)	(2.56)	(3.81)	(3.02)	(8.67)	(1.57)	(1.33)	(0.52)	(2.45)
4	3.99	8.85	12.59	15.88	3.19	4.81	10.35	8.88	1.90	1.99	6.44	2.32
	(1.58)	(4.47)	(2.12)	(14.70)	(2.86)	(4.04)	(2.76)	(8.47)	(1.55)	(1.36)	(0.36)	(2.16)
5	4.53	8.02	18.62	13.61	3.47	4.81	10.96	8.34	1.83	1.97	6.00	1.82
	(2.21)	(4.10)	(1.70)	(13.10)	(3.29)	(3.88)	(2.79)	(8.25)	(1.57)	(1.26)	(−0.19)	(1.47)
6	5.41	8.99	12.32	11.85	3.51	4.25	10.15	7.93	1.74	1.77	6.72	1.48
	(2.84)	(4.18)	(1.46)	(11.76)	(3.32)	(3.69)	(2.64)	(7.90)	(1.46)	(1.21)	(−0.04)	(1.47)
7	5.36	8.99	11.04	10.39	3.54	4.12	10.50	7.14	1.69	1.75	8.02	1.13
	(2.80)	(3.89)	(0.68)	(10.37)	(3.36)	(3.50)	(2.62)	(7.20)	(1.41)	(1.15)	(−0.39)	(1.18)
8	5.54	8.94	10.69	9.76	3.50	3.90	8.58	6.67	1.66	1.78	6.60	0.94
	(2.98)	(3.53)	(−0.64)	(9.70)	(3.30)	(3.26)	(0.75)	(6.67)	(1.36)	(1.13)	(−0.77)	(0.98)

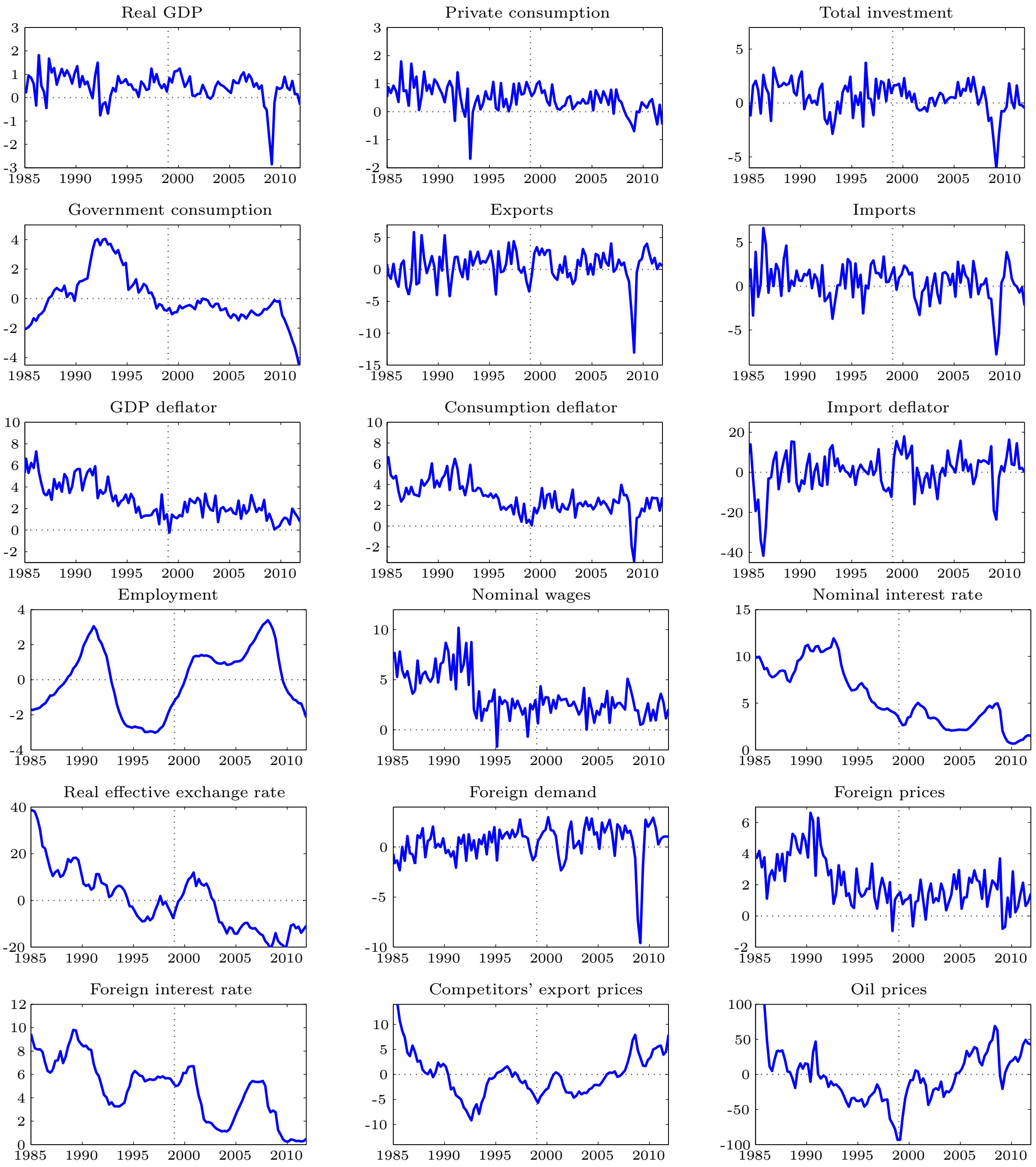
Notes: The results within parentheses are based on dropping the density forecasts of the quarters 2008Q4 and 2009Q1 from the calculations. The log predictive likelihood for the random walk model is calculated with its analytical expression; see the online appendix (Appendix A) for details. For the NAWM and the DSGE-VAR models, 10,000 posterior draws have been taken from the available 500,000 post burn-in draws obtained via the random walk Metropolis sampler. The draws have been selected as draw number $1, 51, 101, \ldots, 499951$. For the BVAR direct sampling is possible and 50,000 posterior draws have been used; see the online appendix (Appendix B).

TABLE 2. Log predictive likelihood values with numerical standard errors based on the MC estimator for density forecasts over the forecast horizon 2008Q1–2009Q4 using historical data until 2007Q4.

	Large selection (12 variables)			Medium selection (7 variables)			Small selection (3 variables)		
Period	NAWM	DSGE-VAR	BVAR	NAWM	DSGE-VAR	BVAR	NAWM	DSGE-VAR	BVAR
2008Q1	−10.3279	−8.5661	−6.1484	−3.6295	−2.4129	−0.6359	−1.1646	−1.1503	−0.5581
	(0.0078)	(0.0127)	(0.0070)	(0.0074)	(0.0102)	(0.0050)	(0.0037)	(0.0070)	(0.0031)
2008Q2	−12.4433	−10.9317	−7.8271	−4.8907	−3.7637	−2.1863	−2.1381	−1.4539	−2.2022
	(0.0075)	(0.0123)	(0.0079)	(0.0067)	(0.0099)	(0.0064)	(0.0041)	(0.0053)	(0.0048)
2008Q3	−16.7118	−13.5161	−14.7368	−7.0085	−5.6422	−7.8644	−3.5873	−3.0935	−6.6163
	(0.0100)	(0.0125)	(0.0319)	(0.0084)	(0.0113)	(0.0240)	(0.0053)	(0.0097)	(0.0136)
2008Q4	−28.7516	−29.8833	−39.3785	−9.4763	−9.5283	−17.3812	−6.1870	−6.2509	−14.1912
	(0.0443)	(0.1210)	(0.8129)	(0.0162)	(0.0220)	(0.1335)	(0.0141)	(0.0179)	(0.0866)
2009Q1	−40.0379	−44.0610	−61.1553	−13.8284	−17.8481	−32.9921	−10.0689	−12.9329	−28.1760
	(0.0928)	(0.2685)	(0.9934)	(0.0264)	(0.1582)	(0.6053)	(0.0252)	(0.0951)	(0.2400)
2009Q2	−19.6019	−19.7989	−30.3377	−9.2210	−9.8641	−19.3259	−3.7704	−4.9328	−9.4277
	(0.0207)	(0.0382)	(0.4501)	(0.0191)	(0.0249)	(0.2059)	(0.0091)	(0.0169)	(0.0525)
2009Q3	−19.4668	−18.7053	−25.5558	−10.2396	−10.1233	−18.1275	−3.8423	−4.8870	−8.2950
	(0.0214)	(0.0295)	(0.3619)	(0.0220)	(0.0240)	(0.1948)	(0.0093)	(0.0154)	(0.0377)
2009Q4	−19.4219	−19.7333	−24.4852	−9.1823	−9.6789	−17.4570	−3.8024	−4.7036	−7.1406
	(0.0173)	(0.0401)	(0.3509)	(0.0183)	(0.0218)	(0.2381)	(0.0091)	(0.0148)	(0.0302)

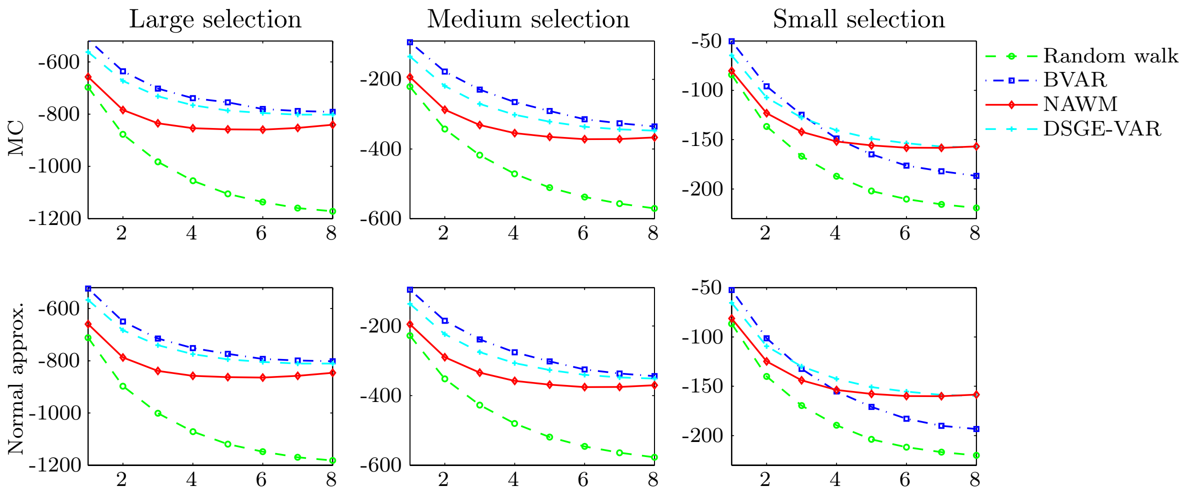
Notes: Numerical standard errors are reported within parentheses. The standard errors of the NAWM and the DSGE-VAR have been computed with the Newey and West (1987) estimator. For these models, 10,000 posterior draws have been taken from the available 500,000 post burn-in draws obtained via the random walk Metropolis sampler. The draws have been selected as draw number $1, 51, 101, \ldots, 499951$. For the BVAR direct sampling is possible and 50,000 posterior draws have been computed; see the online appendix (Appendix B).

Figure 1. The euro area data for the sample 1985Q1–2011Q4.



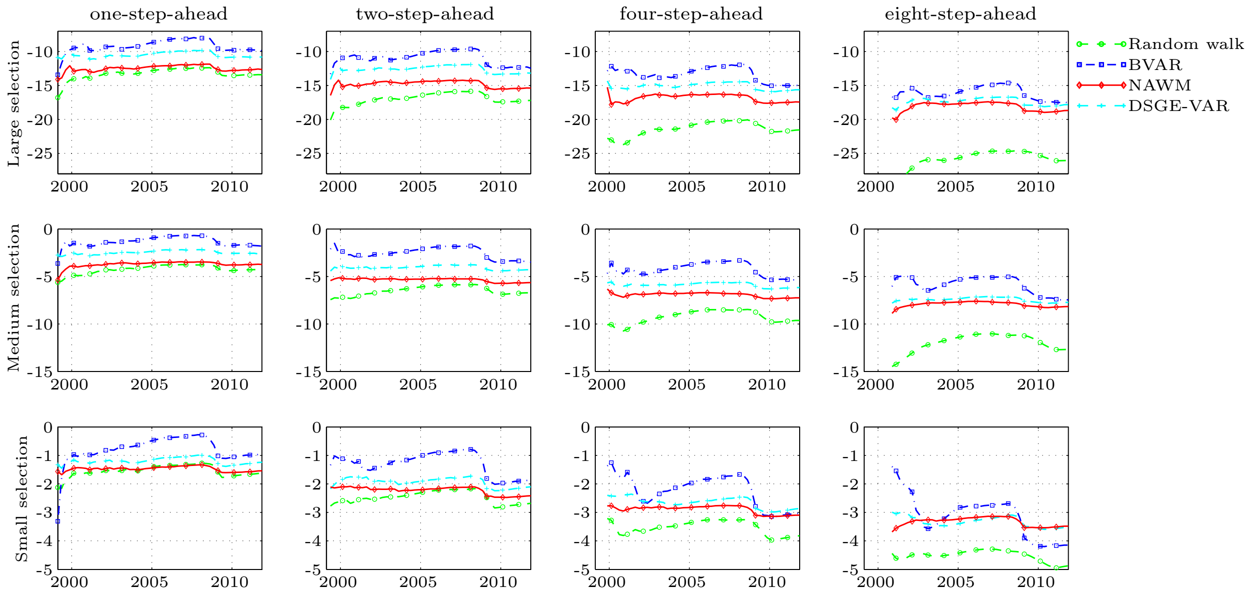
Note: This figure shows the time series of the observable variables used in the estimation of the NAWM. Details on the variable transformations are provided in Christoffel, Coenen, and Warne (2008, Section 3.2) or Section 2.3 in CCW. Price and wage inflation as well as the interest rates are reported in annualized percentage terms.

FIGURE 2. Log predictive scores using the MC estimator of the predictive likelihood for the entire sample 1999Q1–2011Q4.



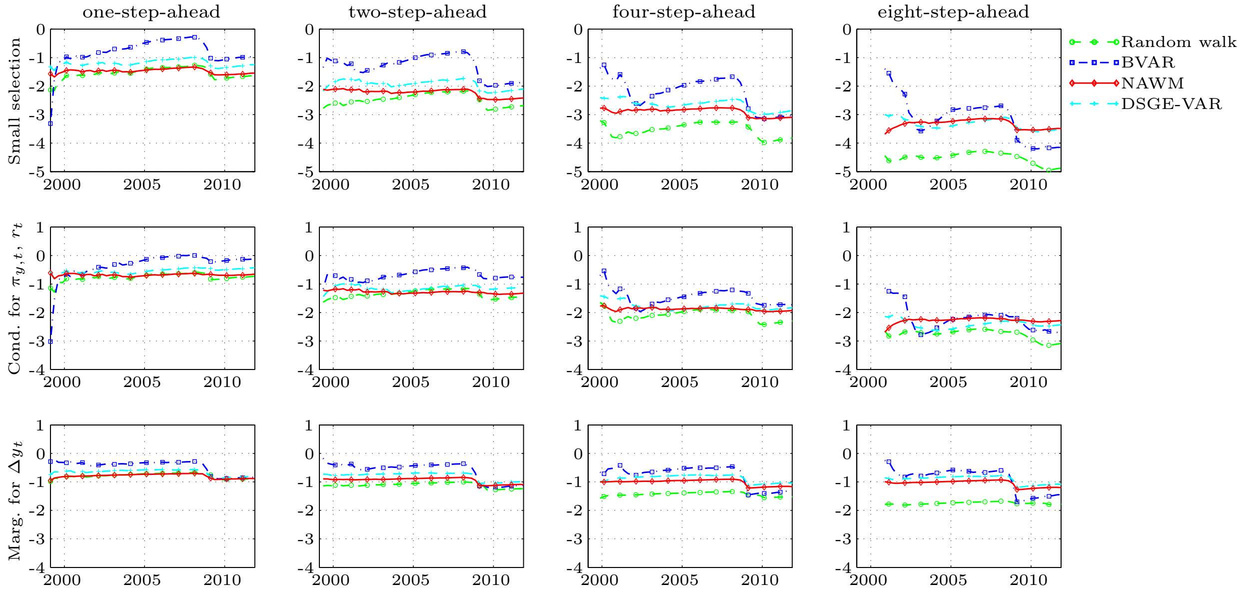
Note: The log predictive likelihood of the random walk model is calculated with its analytical expression.

FIGURE 3. Recursive estimates of the average log predictive score using the MC estimator.



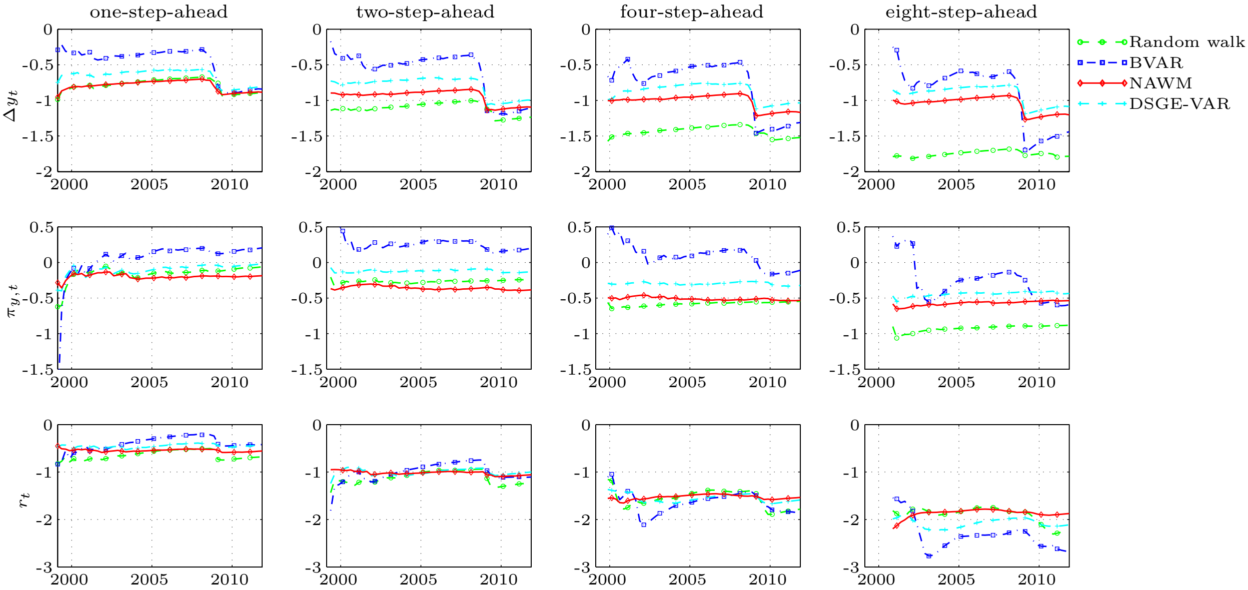
Note: The log predictive likelihood of the random walk model is calculated with its analytical expression. The forecast sample is 1999Q1–2011Q4.

FIGURE 4. Recursive estimates of a decomposition of the average log predictive score for the small selection.



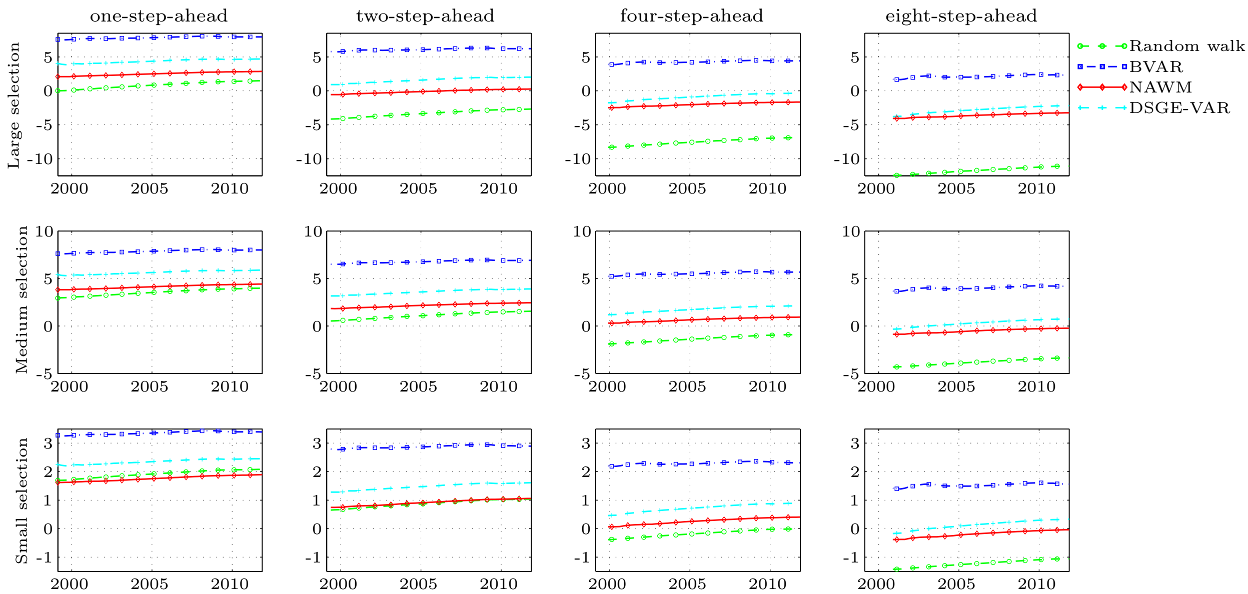
Note: The predictive likelihood of the small selection is decomposed into (i) the conditional predictive likelihood of GDP deflator inflation and the short-term nominal interest rate conditional on real GDP growth, and (ii) the marginal predictive likelihood of real GDP growth. See also Table 3.

FIGURE 5. Recursive estimates of the average log predictive score for real GDP growth, GDP deflator inflation, and the short-term nominal interest rate using the MC estimator.



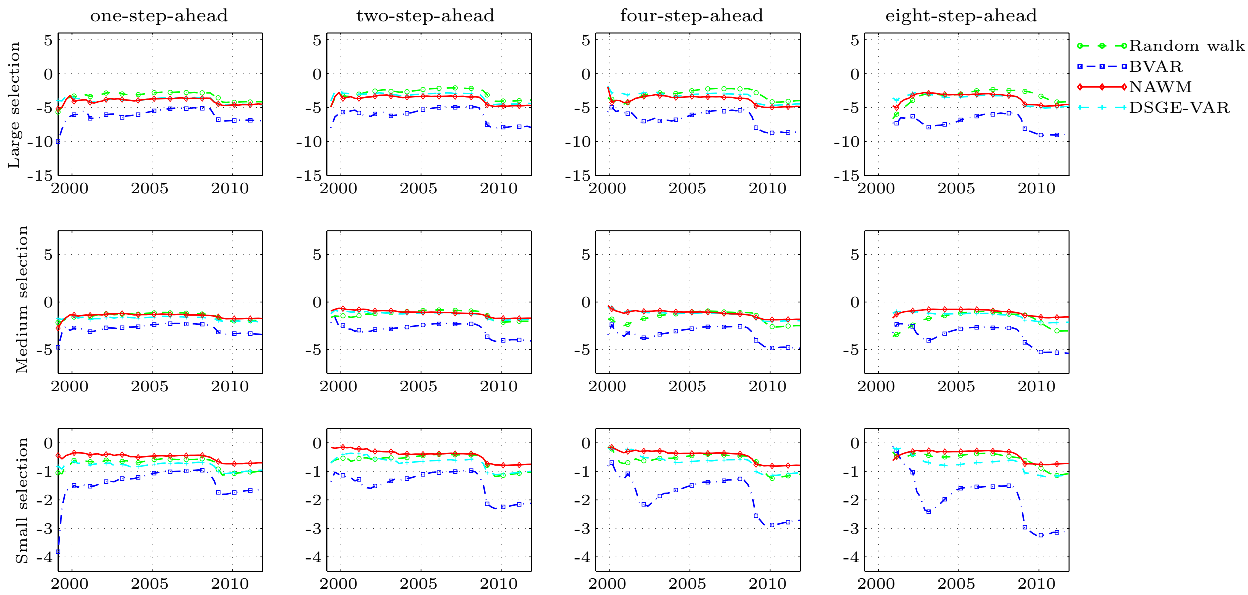
Note: See also Table 3.

Figure 6. Recursive estimates of the forecast uncertainty term of the average log predictive score.



Note: The forecast uncertainty term of the average log predictive score is computed from the normal approximation of the log predictive likelihood; see equation (18). See also Table 3.

FIGURE 7. Recursive estimates of the quadratic standardized forecast error term of the average log predictive score.



Note: The quadratic standardized forecast error term of the average log predictive score is computed from the normal approximation of the log predictive likelihood; see equation (19). See also Table 3.

References

Adolfson, M., Laséen, S., Lindé, J., and Villani, M. (2007a), "Bayesian Estimation of an Open Economy DSGE Model with Incomplete Pass-Through," *Journal of International Economics*, 72, 481–511.

Adolfson, M., Laséen, S., Lindé, J., and Villani, M. (2008), "Evaluating an Estimated New Keynesian Small Open Economy Model," *Journal of Economic Dynamics and Control*, 32, 2690–2721.

Adolfson, M., Lindé, J., and Villani, M. (2007b), "Forecasting Performance of an Open Economy DSGE Model," *Econometric Reviews*, 26, 289–328.

Amisano, G. and Geweke, J. (2013), "Prediction Using Several Macroeconomic Models," ECB Working Paper No. 1537.

An, S. and Schorfheide, F. (2007), "Bayesian Analysis of DSGE Models," *Econometric Reviews*, 26, 113–172, with discussion, p. 173–219.

Anderson, G. S. (2010), "A Reliable and Computationally Efficient Algorithm for Imposing the Saddle Point Property in Dynamic Models," *Journal of Economic Dynamics and Control*, 34, 472–489.

Andersson, M. K. and Karlsson, S. (2008), "Bayesian Forecast Combinations for VAR Models," in S. Chib, G. Koop, W. Griffith, and D. Terrell (Editors), *Bayesian Econometrics*, 501–524, Emerald Group Publishing, Bingley, Volume 23 of *Advances of Econometrics.*

Bańbura, M., Giannone, D., and Reichlin, L. (2010), "Large Bayesian Vector Auto Regressions," *Journal of Applied Econometrics*, 25, 71–92.

Chan, J. C. C. and Eisenstat, E. (2015), "Marginal Likelihood Estimation with the Cross-Entropy Method," *Econometric Reviews*, 34, 256–285.

Christiano, L. J., Eichenbaum, M., and Evans, C. (2005), "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy," *Journal of Political Economy*, 113, 1–45.

Christoffel, K., Coenen, G., and Warne, A. (2008), "The New Area-Wide Model of the Euro Area: A Micro-Founded Open-Economy Model for Forecasting and Policy Analysis," ECB Working Paper Series No. 944.

Christoffel, K., Coenen, G., and Warne, A. (2011), "Forecasting with DSGE Models," in M. P. Clements and D. F. Hendry (Editors), *The Oxford Handbook of Economic Forecasting*, 89–127, Oxford University Press, New York.

Del Negro, M., Hasegawa, R. B., and Schorfheide, F. (2014), "Dynamic Prediction Pools: An Investigation of Financial Frictions and Forecasting Performance," NBER Working Paper No. 20575.

Del Negro, M. and Schorfheide, F. (2004), "Priors from General Equilibrium Models," *International Economic Review*, 45, 643–673.

Del Negro, M. and Schorfheide, F. (2011), "Bayesian Macroeconometrics," in J. Geweke, G. Koop, and H. K. van Dijk (Editors), *The Oxford Handbook of Bayesian Econometrics*, 293–389, Oxford University Press, New York.

Del Negro, M. and Schorfheide, F. (2013), "DSGE Model-Based Forecasting," in G. Elliott and A. Timmermann (Editors), *Handbook of Economic Forecasting*, volume 2, 57–140, North Holland, Amsterdam.

Del Negro, M., Schorfheide, F., Smets, F., and Wouters, R. (2007), "On the Fit of New-Keynesian Models," *Journal of Business & Economic Statistics*, 25, 123–143, with discussion, p. 143–162.

Durbin, J. and Koopman, S. J. (2012), *Time Series Analysis by State Space Methods*, Oxford University Press, Oxford, 2nd edition.

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

Geweke, J. (2005), *Contemporary Bayesian Econometrics and Statistics*, John Wiley, Hoboken.

Geweke, J. and Amisano, G. (2010), "Comparing and Evaluating Bayesian Predictive Distributions of Asset Returns," *International Journal of Forecasting*, 26, 216–230.

Geweke, J. and Amisano, G. (2014), "Analysis of Variance for Bayesian Inference," *Econometric Reviews*, 33, 270–288.

Giannone, D., Lenza, M., and Primiceri, G. E. (2012), "Prior Selection for Vector Autoregressions," ECB Working Paper Series No. 1494.

Giordani, P., Pitt, M., and Kohn, R. (2011), "Bayesian Inference for Time Series State Space Models," in J. Geweke, G. Koop, and H. van Dijk (Editors), *The Oxford Handbook of Bayesian Econometrics*, 61–124, Oxford University Press, New York.

Gneiting, T. and Raftery, A. E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378.

Good, I. J. (1952), "Rational Decisions," *Journal of the Royal Statistical Society Series B*, 14, 107–114.

Harvey, A. C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

Ingram, B. F. and Whiteman, C. H. (1994), "Supplanting the 'Minnesota' Prior — Forecasting Macroeconomic Time Series Using Real Business Cycle Model Priors," *Journal of Monetary Economics*, 34, 497–510.

Karlsson, S. (2013), "Forecasting with Bayesian Vector Autoregressions," in G. Elliott and A. Timmermann (Editors), *Handbook of Economic Forecasting*, volume 2, 791–897, North Holland, Amsterdam.

Klein, P. (2000), "Using the Generalized Schur Form to Solve a Multivariate Linear Rational Expectations Model," *Journal of Economic Dynamics and Control*, 24, 1405–1423.

Litterman, R. B. (1986), "Forecasting with Bayesian Vector Autoregressions — Five Years of Experience," *Journal of Business & Economic Statistics*, 4, 25–38.

Meng, X.-L. and Wong, W. H. (1996), "Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration," *Statistica Sinica*, 6, 831–860.

Newey, W. K. and West, K. D. (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Sargent, T. J. (1989), "Two Models of Measurement and the Investment Accelerator," *Journal of Political Economy*, 97, 251–287.

Sims, C. A. (1980), "Macroeconomics and Reality," *Econometrica*, 48, 1–48.

Sims, C. A. (2002), "Solving Linear Rational Expectations Models," *Computational Economics*, 20, 1–20.

Sims, C. A., Waggoner, D. F., and Zha, T. (2008), "Methods for Inference in Large Multiple-Equation Markov-Switching Models," *Journal of Econometrics*, 146, 255–274.

Smets, F., Warne, A., and Wouters, R. (2014), "Professional Forecasters and Real-Time Forecasting with a DSGE Model," *International Journal of Forecasting*, 30, 981–995.

Smets, F. and Wouters, R. (2003), "An Estimated Stochastic Dynamic General Equilibrium Model for the Euro Area," *Journal of the European Economic Association*, 1, 1123–1175.

Smets, F. and Wouters, R. (2007), "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," *American Economic Review*, 97, 586–606.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, 22, 1701–1728, with discussion, p. 1728–1762.

Warne, A. (2015), "YADA Manual — Computational Details," Manuscript, European Central Bank. Available with the YADA distribution.

Warne, A., Coenen, G., and Christoffel, K. (2013), "Predictive Likelihood Comparisons with DSGE and DSGE-VAR Models," ECB Working Paper Series No. 1536.

Wolters, M. H. (2015), "Evaluating Point and Density Forecasts of DSGE Models," *Journal of Applied Econometrics*, 30, 74–96.