IHS Economics Series
Working Paper 292
October 2012

Forecast Combination Based on Multiple Encompassing Tests in a Macroeconomic DSGE-VAR System

Mauro Costantini
Ulrich Gunter
Robert M. Kunst

Impressum

Author(s):
Mauro Costantini, Ulrich Gunter, Robert M. Kunst

Title:
Forecast Combination Based on Multiple Encompassing Tests in a Macroeconomic DSGE-VAR System

ISSN: Unspecified

2012 Institut für Höhere Studien - Institute for Advanced Studies (IHS)
Josefstädter Straße 39, A-1080 Wien
E-Mail: office@ihs.ac.at
Web: www.ihs.ac.at

292

Reihe Ökonomie

Economics Series

Forecast Combination Based on Multiple Encompassing Tests in a Macroeconomic DSGE-VAR System

Mauro Costantini, Ulrich Gunter, Robert M. Kunst

Institut für Höhere Studien
Institute for Advanced Studies
Vienna

292
Reihe Ökonomie
Economics Series

Forecast Combination Based on Multiple Encompassing Tests in a Macroeconomic DSGE-VAR System

Mauro Costantini, Ulrich Gunter, Robert M. Kunst

October 2012

Institut für Höhere Studien (IHS), Wien
Institute for Advanced Studies, Vienna

Contact:

Mauro Costantini
Department of Economics and Finance
Brunel University London
Uxbridge, UB8 3PH, United Kingdom
email: mauro.costantini@brunel.ac.uk

Ulrich Gunter
Austrian National Bank
Otto-Wagner-Platz 3
1090 Vienna, Austria

Robert M. Kunst – **Corresponding Author**
Department of Economics and Finance
Institute for Advanced Studies
Stumpergasse 56
1060 Vienna, Austria
email: kunst@ihs.ac.at
and
Department of Economics
University of Vienna

Abstract

We study the benefits of forecast combinations based on forecast-encompassing tests relative to uniformly weighted forecast averages across rival models. For a realistic simulation design, we generate multivariate time-series samples of size 40 to 200 from a macroeconomic DSGE-VAR model. Constituent forecasts of the combinations are formed from four linear autoregressive specifications, one of them a more sophisticated factor-augmented vector autoregression (FAVAR). The forecaster is assumed not to know the true data-generating model. Results depend on the prediction horizon. While one-step prediction fails to support test-based combinations at all sample sizes, the test-based procedure clearly dominates at prediction horizons greater than two.

Keywords

Combining forecasts, encompassing tests, model selection, time series, DSGE-VAR model

JEL Classification

C15, C32, C53

Contents

1 Introduction

Forecast combination is often used to improve forecast accuracy. A linear combination of two or more predictions often yields more accurate forecasts than a single prediction when useful and independent information is taken into account (see Bates and Granger, 1969; Clemen, 1989; Timmermann, 2006). In this paper we evaluate the gains in terms of predictive accuracy that can be achieved by combining forecasts on the basis of a multiple encompassing test developed by Harvey and Newbold (2000) as compared to combinations based on simple uniform weights. We focus on predicting real gross domestic product (GDP), the output variable of central interest in macroeconomic analysis. A novelty of this paper is that we use a DSGE-VAR model as the generating mechanism for our data. This hybrid model builds on the DSGE (dynamic stochastic general equilibrium) model suggested by Smets and Wouters (2003) and fuses it to a VAR (vector autoregressive) model following Del Negro and Schorfheide (2004).

Forecast comparisons are often conducted for samples of accounts data taken from databases for several countries. Costantini and Kunst (2011) use French and U.K. data in order to investigate whether and to what extent combined forecasts with weights determined by multiple encompassing tests help in improving prediction accuracy. Results show some benefits for test-based weighting in one of their two data sets. This approach, however, has some limitations as the data-generating mechanism remains unknown and the performance of prediction methods may be affected by sample-specific features such as extraordinary recessions and booms or abrupt policy changes. For this reason, Monte

Carlo simulations play a crucial role in assessing the empirical value of forecast techniques. Unlike empirical applications, comparisons based on simulations are immune to sampling variation, their precision is limited by the number of replications only. Simulation designs, however, must be carefully chosen if the results are to be relevant for typical empirical situations. Linear vector autoregressions (VAR), as they were used as designs in parametric bootstraps by Costantini and Kunst (2011), may miss important aspects of the complex dynamic interaction at work in actual economies. To address this point, the present paper simulates data from a DSGE-VAR structure that builds on a DSGE model, which in turn has been suggested for Euro area data by Smets and Wouters (2003).

Our interest in using a DSGE model for generating data arises from the ubiquitous usage of this modelling approach in current macroeconomic practice, which makes it plausible to view designs of this type as approximating a realistic macroeconomic world. Over the past two decades, these so-called New Keynesian models have been spreading out in the macroeconomic literature, varying in their levels of complexity as well as in the specific focus of application. For example, customized models are used nowadays by virtually every central bank in the world. These institutions are mainly interested in empirical policy analysis (see, e.g., Smets and Wouters 2003), forecasting (see, e.g., Smets and Wouters 2004), or both. In those applications of DSGE models, Bayesian estimation techniques play a major role (see An and Schorfheide, 2007, for a survey).

In contrast to previous work (Costantini *et al.*, 2010), we here rely on a hybrid DSGE-VAR specification due to Del Negro and Schorfheide (2004), as it has evolved that these DSGE-VAR models attain a more realistic representation of actual data than the pure

DSGE variant. We also outline the differences between the results from the two variants.

Our forecasting evaluation assumes that the forecaster has no knowledge of the underlying DSGE or DSGE-VAR model. She considers four time-series specifications as potential approximations to the generating mechanism: a univariate autoregression; two bivariate autoregressions that contain the target variable and one of two main indicator variables, the (nominal) interest rate and the rate of inflation; and a factor-augmented VAR (FAVAR) model that adds two or three estimated common factors to output in order to form a three- or four-dimensional VAR.

The question whether VAR models mimic the dynamic behavior of DSGE models is not completely settled in the literature. Among others, Kascha and Mertens (2009) show that vector autoregressions can be good approximations to the dynamic behavior of DSGE models, while Ravenna (2007) criticizes the quality of this approximation. Boivin and Giannoni (2006) interpret the FAVAR as the reduced form of a DSGE model in the context of short-run forecasting. Gupta and Kabundi (2011) forecast South African data using a DSGE model and FAVAR variants as rival models and find that the FAVAR models outperform the DSGE model. We emphasize, however, that we do not address the issue of DSGE as a prediction device—which is a topic of current interest in the forecasting literature—but we rather use the DSGE-type structures as realistic simulation designs.

From the four models, the forecaster is assumed to form weighted averages for the target variable of output. To this aim, forecast-encompassing regressions (see Section 2) are run in all directions, encompassed models are eliminated as determined by F-statistics and a specific significance level, and the surviving models are averaged uniformly. The multiple encompassing test of Harvey and Newbold (2000) is also considered by Costantini and Pappalardo (2010), who use it to corroborate their hierarchical procedure for forecast combinations that is based on a simple encompassing test of Harvey *et al.* (1998). By contrast, the procedure considered here attains complete symmetry with respect to all rival forecasting models, as the multiple encompassing test is run in all directions.

We evaluate the forecasts for various sample sizes ranging from 40 to 200 observations, i.e. for a range that may be typical for macroeconomic forecasting, on the basis of the traditional moment-based criteria MSE (mean squared error) and MAE (mean absolute error) and also by the incidence of better predictions. For the test procedure, we consider significance levels ranging in 1% steps from 0—which corresponds to uniform weighting—to 10%. The results support testing at sharp levels, mainly at 1%. We also find that simple uniform weighting is difficult to beat and that sample sizes of 200 or more may be needed to firmly establish the relative merits of test-based weighting in single-step prediction. Results in favor of uniform weighting relative to more sophisticated methods are well in line with the forecasting literature (see de Menezes and Bunn, 1993; Clements and Hendry, 1998; Timmermann, 2006). At larger horizons, however, our results tend to support test-based weighting even in smaller samples.

In summary, our experiment is of interest with regard to two aspects: first, it assesses the value of forecast combinations based on multiple encompassing in a realistic DSGE-VAR design; second it assesses the effects on forecast accuracy by dimension reduction in the spirit of FAVAR models.

The plan of this paper is as follows. Section 2 outlines all methods: the forecast-encompassing test, the weighting scheme based on that test, and the rival prediction models that are to be combined. Section 3 details the DSGE-VAR model specification and the simulation design. Section 4 presents the results of the prediction evaluation. Section 5 concludes.

2 Methodology

2.1 Encompassing test procedure for forecasting combination

This section presents the encompassing test procedure used to determine the weights in the combination forecast. The procedure is based on the multiple forecast encompassing F–test developed by Harvey and Newbold (2000).

Consider M forecasting models that deliver out-of-sample prediction errors $e_t^{(k)}, k = 1, \ldots, M$ for a given target variable Y, with t running over an evaluation sample that is usually a portion of the sample of available observations. Then, the encompassing test procedure uses M encompassing regressions:

$$
\begin{aligned}
e_t^{(1)} &= a_1(e_t^{(1)} - e_t^{(2)}) + a_2(e_t^{(1)} - e_t^{(3)}) + \ldots + a_{M-1}(e_t^{(1)} - e_t^{(M)}) + u_t^{(1)}, \\
e_t^{(2)} &= a_1(e_t^{(2)} - e_t^{(1)}) + a_2(e_t^{(2)} - e_t^{(3)}) + \ldots + a_{M-1}(e_t^{(2)} - e_t^{(M)}) + u_t^{(2)}, \\
&\ldots \\
e_t^{(M)} &= a_1(e_t^{(M)} - e_t^{(1)}) + a_2(e_t^{(M)} - e_t^{(2)}) + \ldots + a_{M-1}(e_t^{(M)} - e_t^{(M-1)}) + u_t^{(M)}. \qquad (1)
\end{aligned}
$$

These homogeneous regressions yield M regression F statistics. A model k is said to forecast-encompass its rivals if the F statistic in the regression with dependent variable $e_t^{(k)}$ is insignificant at a specific level of significance. Following the evidence of the forecast-encompassing tests, weighted average forecasts are obtained according to the following rule. If F–tests reject or accept their null hypotheses in all M regressions, a new forecast will be formed as a uniformly weighted average of all model-based predictions. If some, say $m < M$, F–tests reject their null, only those $M - m$ models that encompass their rivals are combined. In this case, each of the surviving models receives a weight of $(M - m)^{-1}$.

2.2 The forecasting models

Forecasts are based on four classes of time-series models and on combinations of representatives from these four classes that have been estimated from the data by least squares after determining lag orders by information criteria. As information criteria, we employ the AIC criterion by Akaike and the BIC criterion by Schwarz (see Lütkepohl, 2005).

The first model class (model #1) is a univariate autoregressive model for the targeted output series. The second and third model are two bivariate vector autoregressive models (VAR). Model #2 contains output and inflation, and model #3 contains output and the nominal interest rate. This choice of added variables has been motivated by the fact that inflation and the interest rate are often viewed as main economic business-cycle indicators and they are also more often reported in the media than the remaining variables of the DSGE system.

The fourth and last model class (model #4) is a factor-augmented VAR (FAVAR) model. Suppose that Y_t is the target variable to be predicted (GDP), while F_t is a vector of unobserved factors that are assumed as related to a matrix of observed variables X by the linear identity $F = X\Lambda$ with unknown Λ, such that the column dimension of F is considerably smaller than that of X. A FAVAR model can be described as follows:

$$\Phi(L) \begin{bmatrix} Y_t \\ F_t \end{bmatrix} = \varepsilon_t, \tag{2}$$

where $\Phi(L) = I - \Phi_1 L - \ldots - \Phi_p L^p$ is a conformable lag polynomial of finite order p. L denotes the lag operator, and I denotes the identity matrix. Equation (2) defines a VAR in $(Y_t, F_t')'$. This system reduces to a standard univariate autoregression for Y_t if the terms in $\Phi(L)$ that relate Y_t to $F_{t-j}, j = 1, \ldots, p$ are all zero. Equation (2) cannot be estimated directly, as the factors F_t are unobserved.

The proper estimation of the models requires the use of factor analysis (see Stock and Watson, 1998, 2002). The estimation procedure consists of two steps. In the first step, the factors F are estimated using principal component analysis. The minimum of the BIC(3) criterion developed by Bai and Ng (2002) determines the number of factors, i.e. the dimension of F. In the second step, the FAVAR model is estimated by a standard VAR method with F_t replaced by the estimate $\hat{F}_t$ that is available from the first step.

Thus, in our forecast experiments, the FAVAR forecasts rely on VAR models for the target output series and two or three additional factors that have been selected automatically from combinations of the nine remaining observable variables of the DSGE system that is detailed in Section 3. The choice of the numbers two and three has been motivated by the fact that it is customary not to use more than a maximum of three factors if nine series are available. In fact, we use two or three as upper bounds on the factor dimension but the information criterion BIC(3) always selects the maximum dimension. This indicates that the variables in the DSGE system are quite heterogeneous and that the information in the system cannot be easily condensed to a low dimension.

It follows that the FAVAR formed using this procedure has a dimension of three or four. On the whole, we consider four variants of our simulation design: AIC and BIC selection of lag orders and two or three additional factors in the FAVAR.

The four rival model classes are incompletely nested, with models #2 to #4 representing generalizations of model #1 and models #2 and #3 representing special cases of #4. Due to the lag selection that tends to choose larger lag orders for the univariate model, however, the general situation is to be seen as non-nested.

For a given considered sample size of N, all models are estimated for samples of size $3N/4$ to $N-h-1$ using expanding windows, with $h=1,\ldots,4$ denoting the prediction horizon. Then, the next observation at position $t=3N/4+h,\ldots,N-1$ is forecasted. In the following, these prediction experiments will be referred to as the predictions using the basic rival models. Note that an original sample of size $N=200$ yields one-step forecasts based on 150 observations up to 198 observations. Thus, the reported accuracy measures average estimates of different quality. However, our design represents the action taken by a forecaster who observes 199 data points and targets the forecast for the observation at $N=200$ by optimizing her combinations of the basic rival forecasts to this aim. In other words, the report of the forecasts from the basic rival models is to be seen as an intermediate step.

For each replication, we consider combinations of forecasts based on weighted averages of the four basic rival models for the observations at time points $t = N$. These combinations are determined by the forecast-encompassing tests outlined above. For the F tests, we consider significance levels of $k * 0.01$ with $k = 0, \ldots, 10$. Note that $k = 0$ corresponds to a uniform average, as no F statistic can be significant at the 0% level and hence models always encompass all other models. By contrast, $k = 10$ corresponds to a significance level of 10%. At sharp levels, forecast encompassing remains often unrejected, and many combinations will be uniform. At looser levels, rejections become more common, and some models will be excluded from the average. At extreme levels, no model will encompass and weights will again tend to be uniform. We do not consider levels beyond 10%, however, as these are unlikely to be of practical use, and some unreported experiments insinuate that they do not improve predictive accuracy.

3 The data-generating process

3.1 A medium-scale DSGE model

Smets and Wouters (2003) originally developed a medium-scale DSGE model of the Euro area and estimated it based on quarterly data and Bayesian techniques. Our objective, however, is to use this closed-economy model in order to create artificial data.

We decided for the model by Smets and Wouters (2003) due to the following two properties. First, the model remains present in the empirical DSGE literature. Besides its original application for policy analysis and forecasting in the Euro area (see Smets and Wouters, 2003, 2004), it was also successfully adapted to US data (see Smets and Wouters, 2005, 2007). Second, it achieves an attractive level of complexity, as it concentrates on the main features of a realistic macroeconomy and avoids being too country-specific. Onatski and Williams (2010) investigate the effects of introducing an assumption of prior uncertainty. While parameter estimates differ from the original values, many of the qualitative features of the Smets and Wouters (2003) model remain intact.

The subsequent ten expectational difference equations constitute the log-linear representation of this fully micro-founded model. For a detailed derivation of these equations see Smets and Wouters (2003). All variables are given in percentage deviations from the non-stochastic steady state, denoted by hats. The endogenous variables are consumption $\hat{C}$, real wage $\hat{w}$, capital $\hat{K}$, investment $\hat{I}$, real value of installed capital $\hat{Q}$, output $\hat{Y}$, labor $\hat{L}$, inflation $\hat{\pi}$, rental rate of capital $\hat{r}^k$, and gross nominal interest rate $\hat{R}$. For a description of all model parameters appearing below see Table 1.

The economy is inhabited by a continuum of measure 1 of infinitely-lived households who maximize the present value of expected future utilities. The optimal intertemporal allocation of consumption characterized by external habit formation is given by:

$$\hat{C}_t = \frac{h}{1+h}\hat{C}_{t-1} + \frac{1}{1+h}\mathrm{E}_t\{\hat{C}_{t+1}\} - \frac{1-h}{(1+h)\sigma_c}\{\hat{R}_t - \mathrm{E}_t(\hat{\pi}_{t+1})\} + \frac{1-h}{(1+h)\sigma_c}\varepsilon_t^b. \qquad (3)$$

Households are monopolistically competitive suppliers of labor and face nominal rigidities in terms of Calvo (1983) contracts when resetting their nominal wage. These assumptions imply a New Keynesian Phillips curve for the real wage, which is characterized by partial indexation:

$$\begin{aligned} \hat{w}_t &= \frac{\beta}{1+\beta}\mathrm{E}_t(\hat{w}_{t+1}) + \frac{1}{1+\beta}\hat{w}_{t-1} + \frac{\beta}{1+\beta}\mathrm{E}_t(\hat{\pi}_{t+1}) - \frac{1+\beta\gamma_w}{1+\beta}\hat{\pi}_t + \frac{\gamma_w}{1+\beta}\hat{\pi}_{t-1} \\ &- \frac{1}{1+\beta}\frac{(1-\beta\xi_w)(1-\xi_w)}{\{1+\frac{(1+\lambda_w)\sigma_l}{\lambda_w}\}\xi_w}\{\hat{w}_t - \sigma_l\hat{L}_t - \frac{\sigma_c}{1-h}(\hat{C}_t - h\hat{C}_{t-1}) + \varepsilon_t^l\} + \eta_t^w. \qquad (4) \end{aligned}$$

Capital is also owned by households and accumulates according to:

$$\hat{K}_t = (1-\tau)\hat{K}_{t-1} + \tau\hat{I}_{t-1}. \qquad (5)$$

Investment, which is subject to adjustment costs, evolves as follows:

$$\hat{I}_t = \frac{1}{1+\beta}\hat{I}_{t-1} + \frac{\beta}{1+\beta}\mathrm{E}_t(\hat{I}_{t+1}) + \frac{\varphi}{1+\beta}\hat{Q}_t + \varepsilon_t^i. \qquad (6)$$

The corresponding equation for the real value of installed capital reads:

$$\hat{Q}_t = -\{\hat{R}_t - \mathrm{E}_t(\hat{\pi}_{t+1})\} + \frac{1-\tau}{1-\tau+\bar{r}^k}\mathrm{E}_t(\hat{Q}_{t+1}) + \frac{\bar{r}^k}{1-\tau+\bar{r}^k}\mathrm{E}_t(\hat{r}_{t+1}^k) + \eta_t^q. \qquad (7)$$

Moreover, there is also a continuum of measure 1 of monopolistically competitive intermediate goods producers who maximize the present value of expected future profits while facing the subsequent production function:

$$\hat{Y}_t = \phi\varepsilon_t^a + \phi\alpha\hat{K}_{t-1} + \phi\alpha\psi\hat{r}_t^k + \phi(1-\alpha)\hat{L}_t. \tag{8}$$

Their labor demand equation is therefore given by:

$$\hat{L}_t = -\hat{w}_t + (1+\psi)\hat{r}_t^k + \hat{K}_{t-1}. \tag{9}$$

Similar to households, intermediate goods producers face nominal rigidities in terms of Calvo (1983) contracts when resetting their price. These assumptions imply the standard New Keynesian Phillips curve, which again is characterized by partial indexation:

$$\begin{aligned} \hat{\pi}_t &= \frac{\beta}{1+\beta\gamma_p}\mathrm{E}_t\{\hat{\pi}_{t+1}\} + \frac{\gamma_p}{1+\beta\gamma_p}\hat{\pi}_{t-1} \\ &+ \frac{1}{1+\beta\gamma_p}\frac{(1-\beta\xi_p)(1-\xi_p)}{\xi_p}\{\alpha\hat{r}_t^k + (1-\alpha)\hat{w}_t - \varepsilon_t^a\} + \eta_t^p. \end{aligned} \tag{10}$$

Using data from 13 OECD countries, Korenok *et al.* (2010) showed that this way of modelling firms' price-setting behaviour—sticky prices in combination with indexation—represents actually observed behaviour quite well.

The goods market equilibrium condition reads:

$$\hat{Y}_t = (1-\tau k_y - g_y)\hat{C}_t + \tau k_y\hat{I}_t + \varepsilon_t^g. \tag{11}$$

Finally, monetary policy is assumed to be implemented by the following Taylor-type interest-rate rule:

$$\hat{R}_t = \rho\hat{R}_{t-1} + (1-\rho)\{\bar{\pi}_t + r_\pi(\hat{\pi}_{t-1} - \bar{\pi}_t) + r_y\hat{Y}_t\} + r_{\Delta\pi}(\hat{\pi}_t - \hat{\pi}_{t-1}) + r_{\Delta y}(\hat{Y}_t - \hat{Y}_{t-1}) + \eta_t^r. \quad (12)$$

Differing from the original article, we assume that the interest-rate rule depends on actual output only, but not on hypothetical potential output.

Equations (3)–(12) contain six macroeconomic shocks that are assumed to follow independent stationary AR(1) processes of the form $\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$ with $\rho \in (0,1)$ and η i.i.d. $\sim N(0, \varsigma_\eta^2)$. More specifically, there is a consumption preference shock ε^b in equation (3), a labor supply shock ε^l in equation (4), an investment shock ε^i in equation (6), a productivity shock ε^a in equation (10), a government spending shock ε^g in equation (11), and an inflation objective shock $\bar{\pi}$ in equation (12).

In addition, there are four shocks assumed to follow i.i.d. processes $\sim N(0, \varsigma_\eta^2)$: there is a real-wage mark-up shock η^w in equation (4), an equity-premium shock η^q in equation (7), a price mark-up shock η^p in equation (10), and an interest-rate shock η^r in equation (12).

The parameter values given in Table 1 correspond to the modes of the posterior distributions of the parameters in case those were estimated in Smets and Wouters (2003) or, otherwise, to the values that were kept fixed during Bayesian estimation. All parameter values guarantee that the Blanchard and Kahn (1980) conditions are satisfied, which means that there are six eigenvalues of the coefficient matrix of the equation system (3)–(12) larger than 1 in modulus for its six forward-looking variables $(\hat{C}, \hat{w}, \hat{I}, \hat{Q}, \hat{\pi}, \hat{r}^k)$. Hence, there is a unique stationary solution to the equation system (3)–(12).

3.2 The DSGE-VAR simulation design

Smets and Wouters (2003) originally developed a medium-scale DSGE model of the Euro area and estimated it based on quarterly data and Bayesian techniques. At first sight, this closed-economy model has the two desirable properties for creating artificial data that we mentioned above, i.e. relevance in macroeconomics due to widespread usage and an attractive level of complexity.

Nonetheless, whereas Smets and Wouters (2003) find that their DSGE models attain a higher marginal likelihood than VARs—which would promise a better out-of-sample predictive performance—Del Negro *et al.* (2007) warned that such findings crucially hinge on the observation sample. Even relatively sophisticated DSGE models are not robust against small changes in the sample period, hence a non-negligible degree of misspecification in DSGE models is apparent. In consequence, policy recommendations and forecasts based on this model class could be biased, and the empirical plausibility of artificial data generated by such a model may be impaired. Moreover, Smets and Wouters (2007) find that the estimates for some of the model parameters differ considerably between the 'Great Inflation' (1966Q1–1979Q2) and 'Great Moderation' (1984Q1–2004Q4) subsamples in US data, which casts doubt on the validity of approximating the actual economy by a DSGE model with time-constant parameters.

One way of addressing this misspecification issue is to replace the pure DSGE data-generating process by a hybrid DSGE-VAR that is known to be much less sensitive to changes in the observation period and also typically attains a higher marginal likelihood than both VAR and DSGE specifications (see Del Negro *et al.*, 2007). The DSGE-VAR developed by Del Negro and Schorfheide (2004) and Del Negro *et al.* (2007) is a Bayesian VAR (BVAR) that uses the information provided by a DSGE model as an informative prior for BVAR estimation. The impact of the DSGE information relative to the actual sample information is measured by some hyper-parameter $\aleph \in (0, \infty]$, which can either be kept fixed during estimation or estimated together with the DSGE model parameters (see Adjemian *et al.*, 2008). A value of $\aleph$ close to 0 corresponds to an unrestricted VAR at the one extreme, whereas a value of $\aleph$ equal to ∞ corresponds to the VAR approximation of the DSGE model at the other extreme (see Del Negro *et al.*, 2007).

The misspecification of the DSGE model class also shows in the optimal weight of the DSGE information for constructing the DSGE-VAR prior of $\aleph^* = 1.25$ for the sample from 1974Q2–2004Q1 in Del Negro *et al.* (2007), which reflects an optimal impact of the information provided by the DSGE model of around 55% for DSGE-VAR estimation. For the derivation of the DSGE-VAR prior and posterior distributions as well as for a more technical description of the DSGE-VAR methodology, see Del Negro and Schorfheide (2004) and Del Negro *et al.* (2007).

Our aim is to generate artificial data that are empirically plausible across countries and sample periods. We therefore apply the subsequent three-step DSGE-VAR procedure while employing the Dynare preprocessor for Matlab, which is downloadable in its current version from `http://www.dynare.org`:

Step 1. We generate 2,000 time series, each of length 1,100, for ten key macroeconomic variables (consumption $\hat{C}$, real wage $\hat{w}$, capital $\hat{K}$, investment $\hat{I}$, real value of installed capital $\hat{Q}$, output $\hat{Y}$, labor $\hat{L}$, inflation $\hat{\pi}$, rental rate of capital $\hat{r}^k$, and gross nominal interest rate $\hat{R}$) from the original Smets and Wouters (2003) model as laid out in Section 3.1.

Step 2. Discarding the first 100 observations of each of the 2,000 time series as burn-in draws, the remaining $T = 1,000$ observations serve as the data sample for estimating a DSGE-VAR of lag order $p = 2$ via Bayesian techniques. The posterior distribution of a DSGE-VAR model cannot be determined analytically, hence a Monte-Carlo Markov chain sampling algorithm has to be invoked to simulate the distribution of the vector of DSGE-VAR model parameters (for a survey on Bayesian inference in DSGE models see An and Schorfheide, 2007). In particular, we adopt a Metropolis-Hastings algorithm with two parallel Monte-Carlo Markov chains, each consisting of 55,000 draws. The first half of the draws are discarded before computing the posterior simulations. A desirable acceptation rate of draws according to Roberts *et al.* (1997) of about one third is met across simulations.

The DSGE model used for constructing the DSGE-VAR prior again is the model by Smets and Wouters (2003) as given by equations (3)–(12). Due to the computational burden associated with 2,000 full-fledged Metropolis-Hastings simulations, we declare our target variable of interest ($\hat{Y}$), the two additional variables used in the bivariate VAR forecast models ($\hat{R}, \hat{\pi}$), and two auxiliary variables ($\hat{w}, \hat{r}^k$) as the only observed variables, i.e. there are $m = 5$ observed variables altogether. We further restrict the number of free parameters to those listed in Table 2. Concerning the prior probability density functions as well as the prior means and standard deviations of the DSGE model parameters, we again follow Smets and Wouters (2003).

In line with Adjemian *et al.* (2008), we choose to estimate the hyper-parameter $\aleph$ along with the deep parameters of the DSGE model. Similar to Adjemian *et al.* (2008), we assume a uniform distribution for the hyper-parameter between $\aleph = 0.1$, corresponding to an impact of the DSGE model information of about 10%, and $\aleph = 10$, corresponding to an impact of the DSGE model information of about 90% (note that, as in Adjemian *et al.*, 2008, the minimum value to obtain a proper prior $\aleph^{min} \geq (mp + m)/T = 0.015$ is satisfied). All other parameters listed in Table 1 are kept fixed at the indicated values during estimation.

Step 3. After retrieving the posterior distributions of the model parameters for each of the 2,000 replications, we generate time series of length 1,100 for the ten macro variables using the pure perturbation algorithm of Schmitt-Grohé and Uribe (2004). In this step, parameter values are set at the means of the posterior distributions.

This step is repeated 2,000 times to obtain 2,000 new time series of artificial data. Whereas the first 100 observations of each time series are discarded as starting values, the remaining 1,000 observations are separated into shorter non-overlapping time series. Thus, the number of replications depends on the sample size. For the largest sample size of $N = 200$, 10,000 replications are available for our forecasting experiments. At the other extreme, for the smallest considered sample size of $N = 40$, the number of available replications increases to 50,000. The sample size N is varied over $20 * j$ for $j = 2, 3, \ldots, 10$. Samples smaller than $N = 40$ would admit no useful forecasting evaluation, due to the relatively high dimension of the system.

4 Results

This section consists of three parts. First, we focus on the relative forecasting performance of the four basic rival models. The second subsection looks at the weights that these rival models obtain in the test-based forecast combinations. The third part considers the performance of the combined forecast in detail.

4.1 Performance of the rival models

Based on the evaluation of mean squared errors, Figure 1 shows that the factor VAR model dominates at larger sample sizes in all designs, that is for AIC as well as BIC and for two as well as three factors. Unreported control experiments have shown that this is not true for FAVAR variants that do not explicitly include the predicted variable as an additional factor. In other words, the factor search is unable to locate the most important variable for forecasting, the target variable itself, as it concentrates on variance contributions. However, it is successful in adding factors to the list, while we do not have proof that this choice is optimal with regard to prediction.

The two or three factors identified by the FAVAR algorithm vary considerably across replications. A rough inspection of the average weights of observed variables shows that the first factor tends to incorporate investment I and the capital stock K. Even the second factor tends to assign large weights to I and to K, with some contributions from wages W

and the rental rate r^k. The third factor focuses on consumption C and on wages W, with some further contributions from the labor force L and the real value of capital Q.

In small samples, the univariate autoregression dominates but it loses ground as the sample size increases. Among the two bivariate VAR models, a clear ranking is recognizable. Model #3 with output and nominal interest rate achieves a more precise prediction for output than model #2 with output and inflation. This ranking is due to the structure of the DSGE model that assumes stronger links between output and the interest rate than between output and inflation. In fact, model #3 is pretty good for intermediate samples and can compete with the FAVAR specification at all but the largest samples, particularly in the AIC variants.

By contrast, the FAVAR performance is extremely poor in small samples, with the worst impression for AIC and three factors. This is the most profligate specification, with the largest number of free parameters to be estimated. For BIC order selection, FAVAR overtakes its rivals for good around $N = 100$, whereas slightly larger samples are needed for the AIC cases.

All four graphs use a common scale, which admits a rough comparative visual assessment of the four variants. An obvious feature is the inferior performance of the combination of AIC and three factors in small samples, due to parameter profligacy. In large samples, AIC and BIC perform similarly for the FAVAR model, with three-factor specifications overtaking two-factor specifications. BIC selection, however, becomes less attractive for its less informative rival models that would need longer lags for optimizing their predictions.

Figure 1 restricts attention to single-step prediction. Results for longer horizons are qualitatively similar and are not reported. They are available upon request.

4.2 Weights in the combination forecasts

The univariate model is best for small samples, the FAVAR is best for large samples. Thus, one may expect that the FAVAR model receives a stronger weight in the encompassing-test weighting procedure, as the samples get larger.[1] Figure 2 shows that this is indeed the case. There are slight differences between the AIC and BIC search. AIC implies a share of FAVAR of less than the proportional 25% for $N = 40$, meaning that the FAVAR is often encompassed. BIC, on the other hand, chooses the proportional share even for $N = 60$ and $N = 80$. While for BIC order selection the less informative rivals outperform the FAVAR model in small samples with respect to the MSE criterion (see Figure 1), this behavior does not entail forecast encompassing, due to the heavily penalized and thus typically low lag orders. Otherwise, reaction is fairly monotonic in the sense that the FAVAR share increases with rising N and also with looser significance level.

As the significance level increases, weights diverge from the uniform pattern. We note, however, that even at 10% and $N = 200$ the weight allotted to the FAVAR model hardly exceeds 40%. This value is an average over many replications with uniform weighting and comparatively few where weights of 1/3, of 1/2, or even of one are allotted to FAVAR.

Whereas the weights for the univariate model and the bivariate VAR with inflation monotonically decrease with increasing N, weights for the bivariate model #3 with the

[1] Ericsson (1992) showed that the null hypothesis of the forecast encompassing test is a sufficient condition for forecast MSE dominance.

interest rate peak for intermediate samples and are overtaken by FAVAR as N exceeds 120. Contrary to the FAVAR weights, they rise fast at small significance levels and then level out. Model #3 captures the essence of the DSGE-VAR dynamics at reasonable sample sizes well, and if it wins the encompassing tournament it does so typically at sharp significance levels. Figure 3 provides a summary picture of the weight allotted to model #3 and demonstrates that this model remains competitive in larger samples, with weights decreasing only slowly as N approaches 200, particularly in the AIC variant, thus confirming the impression from Figure 1.

When the prediction horizon grows, the main features of Figure 2 and 3 continue to hold, with one noteworthy exception. For larger samples, Figure 2 shows a smooth increase of the weight allotted to the FAVAR model with rising significance level. At larger horizons, this slope steepens, such that even at the 1% level a considerable weight is assigned to FAVAR. The larger weight allotted to FAVAR coincides with a lesser weight assigned to the bivariate model #3. This stronger discrimination among rival models affects the accuracy comparison to be reported in the next subsection.

4.3 Performance of test-based weighting

In order to evaluate the implications of the test-based method for forecasting, we use three criteria: the mean squared error (MSE), the mean absolute error (MAE), and the winning incidence. Generally, the MAE yields similar qualitative results as the MSE and we do not show the MAE results in detail. The qualitative coincidence of MAE and MSE naturally reflects the normal distribution used in the simulation design.

Figures 4 to 7 shows ratios of the MSE achieved by the test-based weighting relative to the benchmark of a uniformly weighted forecast, depending on the sample size. Values below one indicate an advantage for the test-based procedure. In order not to overload the graphs, they contain results for $N \in \{40, 80, 120, 160, 200\}$ only, while all simulation results are available for $N = 40 + k * 20, k = 0, \ldots, 8$. The intermediate values always correspond to roughly interpolated curves, so little information is lost here.

For single-step prediction (see Figure 4), differences among models are small and remain in the range of 2% at the maximum values. For the smallest sample size $N = 40$, testing at strict levels is competitive with uniform weighting, while testing at looser levels is not. For sample sizes $60 \leq N \leq 100$, uniform weighting outperforms the test-based weighting scheme. Performance at $N = 80$ is even monotonic in the sense that a looser significance level and thus a greater divergence from uniformity implies further deterioration. We note, referring to Figure 2, that these differences to uniformity are small. At larger samples, test-based weighting becomes again preferable, with optimum significance levels at 1% or occasionally 2%. The experiments with two factors are slightly more friendly to the test-based rules than those with three factors, while there is hardly any difference between the AIC and BIC versions.

Costantini *et al.* (2010) reported simulations for a pure DSGE rather than a DSGE-VAR design. Generally, test-based weighting has a better relative performance in the hybrid model investigated here. Even at $N = 200$, uniform weighting dominates test-based weighting for one-step predictions if data stem from a pure DSGE model, while the DSGE-VAR—which we view as the empirically more relevant and realistic one—shows small but clear advantages for test-based weighting.

The non-monotonic reaction, with test-based weighting preferable for very small and for large samples and inferior for the empirically quite relevant intermediate range, asks for an explanation. For $N = 40$, test-based weighting tends to hardly deviate from uniform weighting, and, if it does, it tends to eliminate the FAVAR structure. This relatively complex FAVAR, however, incurs a larger prediction error than its simple rivals, thus it pays to eliminate it at $N = 40$. As the sample size grows, the weight of the FAVAR model increases, while poorer models are more often eliminated, which however becomes profitable only when this decision is reliable. To this aim, relatively large samples are needed.

If the step size increases, the occurrence of ties among the procedures becomes less prominent. This, in turn, leads to a clearer separation with regard to the accuracy ensuing from prediction models. The weight allotted to the best model, in large samples the FAVAR model, increases.

The graphs for the case of two-step predictions, Figure 5, show that this stronger emphasis on the FAVAR model now leads to an improvement in accuracy for most sample sizes, with $N = 80$ constituting the only exception. Whereas one-step prediction shows a window between small samples—where any elimination of inferior models is helpful—and large samples—where the best prediction models are properly recognized, this window nearly closes if the prediction horizon is extended to two. The optimum significance level remains mostly at 1%, with only few cases of minima at 2% or 3%.

The impression that larger step sizes benefit the test-based procedure is confirmed for the three-step prediction that yields the graphs given as Figure 6. Test-based weighting dominates uniform weights at all sample sizes and specifications. The significance level of 1%, i.e. the sharpest level, is clearly supported over the looser levels for the encompassing test. These features are confirmed and slightly enhanced by the four-step predictions summarized in Figure 7. Again, test-based weighting dominates in all four variants for all sample sizes. Advantages for test-based weighting are most pronounced in the smallest and largest samples. Relative gains in precision attain around 3%.

The criteria MAE and MSE are summary statistics, and they are based on moments of the error distributions. A lower MSE may be attained by a forecast that is actually worse in many replications but wins few of them at a sizeable margin. Therefore, we also consider the direct ranking of absolute errors across significance levels. The incidence of a minimum among all levels could indicate which level is more likely to generate the best forecast. There are many ties among these significance levels, however, so we only report the direct comparison between the 1% test-based weighting and the uniform benchmark in more detail. Figure 8 shows the frequencies of each of these two models of generating the smaller prediction error. Among others, Chatfield (2001) advocated the usage of the winning incidence as a measure of predictive accuracy.

For one-step forecasts, Figure 8 demonstrates that the differences in MSE reported above are due to comparatively few replications. Ties are many even for large samples (around 70%) and are the rule for small samples (around 90%). At small samples, no advantage for the test-based scheme is recognizable. At large samples, test-based weighting gains some margin over its uniform rival but fails to impress.

In line with the MSE graphs, also the 'winning frequency' for the test-based scheme improves at larger forecast horizons. At two steps, the two schemes are still comparable. There is a slight advantage for the encompassing test in the BIC versions, while uniform weighting is remarkably strong in the AIC versions. At three and four steps, however, the test-based procedure gains a sizeable margin. Also note that ties become less frequent and their frequency falls to around 30% at horizon four and larger samples.

In summary, at larger prediction horizons test-based weighting becomes increasingly attractive. At short horizons, the merits of test-based weighting are most pronounced for very small samples, where the accuracy of prediction is low, and at larger samples, where weighting becomes reliable. Many empirical samples may belong to the intermediate region, where the prediction horizon must exceed two in order to provide a clear support for weights based on the encompassing test.

5 Conclusion

Our forecast evaluations generally confirm the traded wisdom in the forecasting literature that uniform weighting of rival model forecasts is difficult to beat in typical forecasting situations. Large sample sizes are needed to reliably eliminate inferior rival models from forecasting combinations. In many situations of empirical relevance, the information contained in slightly worse predictions as marked by individual MSE performance may still be helpful for increasing the precision of the combination.

Forecast-encompassing tests imply a reasonable weighting of individual models in our experiments. Univariate models yield the best forecasts in small samples, and sophisticated higher-dimensional models receive a small weight. With increasing sample size, our experiments clearly show that the factor-augmented VAR achieves superior predictive accuracy and thus it receives the largest weights in test-based combinations. The benefits with respect to an optimized combination forecast, however, turn out to be more difficult to exploit. At the one-step horizon, the test-based combination forecast fails to show a clear dominance over a simple uniform weighting procedure in the range of $N = 60$ to $N = 120$ that is of strong empirical relevance. Only at horizons of three and beyond does the dominance of test-based weighting become convincing. A noteworthy general result is that, for the encompassing test, the sharpest significance level of 1% often yields the best results.

References

Adjemian, S., Darracq Pariès, M., Moyen, S. 2008. Towards a Monetary Policy Evaluation Framework. *ECB Working Paper* No. 942.

An, S., Schorfheide, F., 2007. Bayesian Analysis of DSGE Models. *Econometric Reviews* 26, 113–172.

Bai, J., Ng, S., 2002. Determining the Number of Factors in Approximate Factor Models. *Econometrica* 70, 191–221.

Bates, J.M., Granger, CWJ., 1969. The combination of forecasts. *Operations Research Quarterly* 20, 451–468.

Blanchard, O.J., Kahn, C.M., 1980. The Solution of Linear Difference Models under Rational Expectations. *Econometrica* 48, 1305–1311.

Boivin, J., Giannoni, M.P., 2006. DSGE Models in a Data-Rich Environment. NBER Technical Working Papers No. 0332, National Bureau of Economic Research.

Calvo, G.A., 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12, 383–398.

Chatfield, C., 2001. *Time-series forecasting.* Chapman & Hall.

Clemen, R.T., 1989. Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting* 26, 725–743.

Clements, M., Hendry, D.F., 1998. *Forecasting economic time series.* Cambridge University Press.

Costantini, M., Gunter, U., Kunst, R.M., 2010. Forecast Combination Based on Multiple Encompassing Tests in a Macroeconomic DSGE System. *Economics Series* 251, Institute for Advanced Studies, Vienna.

Costantini, M., Kunst, R.M., 2011. Combining forecasts based on multiple encompassing tests in a macroeconomic core system. *Journal of Forecasting* 30, 579–596.

Costantini, M., Pappalardo, C., 2010. A hierarchical procedure for the combination of forecasts. *International Journal of Forecasting* 26, 725–743.

Del Negro, M, Schorfheide, F. 2004. Priors from General Equilibrium Models for VARs. *International Economic Review* 45, 643–673.

Del Negro, M, Schorfheide, F, Smets, F, Wouters, R. 2007. On the Fit of New Keynesian Models. *Journal of Business & Economic Statistics* 25, 123–162.

de Menezes, L., Bunn, D.W., 1993. Diagnostic Tracking and Model Specification in Combined Forecast of U.K. Inflation. *Journal of Forecasting* 12, 559–572.

Ericsson, N.R., 1992. Parameter constancy, mean square forecast errors, and measuring forecast performance: An exposition, extensions, and illustration. *Journal of Policy Modeling* 14, 465–495.

Gupta, R., Kabundi, A., 2011. A large factor model for forecasting macroeconomic variables in South Africa. *International Journal of Forecasting* 27, 1076–1088.

Harvey, D.I., Leybourne, S., Newbold, P., 1998. Tests for forecast encompassing. *Journal of Business & Economic Statistics* 16, 254–259.

Harvey, D.I., Newbold, P., 2000. Tests for Multiple Forecast Encompassing. *Journal of Applied Econometrics* 15, 471–482.

Kascha, C., Mertens, K., 2009. Business cycle analysis and VARMA models. *Journal of Economic Dynamics and Control* 33, 267–282.

Korenok, O, Radchenko, S, Swanson, N.R., 2010. International evidence on the efficacy of New-Keynesian models of inflation persistence. *Journal of Applied Econometrics* 25, 31–54.

Lütkepohl, H., 2005. *New Introduction to Multiple Time Series*. Springer.

Onatski, A, Williams, N., 2010. Empirical and policy performance of a forward-looking monetary model. *Journal of Applied Econometrics* 25, 145–176.

Ravenna, F., 2007. Vector autoregressions and reduced form representations of DSGE models. *Journal of Monetary Economics* 54, 2048–2064.

Roberts, G.O., Gelman, A., Gilks, W.R., 1997. Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. *The Annals of Applied Probability* 7, 110–120.

Schmitt-Grohé, S., Uribe, M.. 2004. Solving dynamic general equilibrium models using a second-order approximation to the policy function. *Journal of Economic Dynamics and Control* 28, 755–775.

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

Smets, F., Wouters, R., 2004. Forecasting with a Bayesian DSGE Model: An Application to the Euro Area. *Journal of Common Market Studies* 42, 841–867.

Smets, F, Wouters, R., 2005. Comparing Shocks and Frictions in US and Euro Area Business Cycles: a Bayesian DSGE Approach. *Journal of Applied Econometrics* 20, 161–183.

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

Stock, J.H., Watson, M.W., 1998. Diffusion indexes. Working paper No. 6702, NBER.

Stock, J.H., Watson, M.W., 2002. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics* 20, 147–162.

Timmermann, A., 2006. Forecast combinations, in: Elliott, G., Granger, C.W.J., and Timmermann, A. (Eds.), *Handbook of Economic Forecasting*. Elsevier.

Tables and figures

Table 1: Parameters of the DSGE model and their values.

Parameter	Value	Description
β	0.99	Intertemporal discount factor
τ	0.025	Depreciation rate of capital
α	0.3	Capital output ratio
ψ	1/0.169	Inverse elasticity of capital utilization cost
γ_p	0.469	Degree of partial indexation of price
γ_w	0.763	Degree of partial indexation of real wage
λ_w	0.5	Mark-up in real wage setting
ξ_p	0.908	Degree of Calvo price stickiness
ξ_w	0.737	Degree of Calvo real-wage stickiness
σ_l	2.4	Inverse elasticity of labor supply
σ_c	1.353	Coefficient of relative risk aversion in consumption
h	0.573	Degree of habit formation in consumption
ϕ	1.408	1 + share of fixed cost in production
φ	1/6.771	Inverse of investment adjustment cost
$\bar{r}^k$	$1/\beta - 1 + \tau$	Steady-state rental rate of capital
inv_y	0.22	Share of investment to output
k_y	inv_y/τ	Share of capital to output
c_y	0.6	Share of consumption to output
g_y	$1 - c_y - inv_y$	Share of government spending to output
r_π	1.684	Inflation coefficient
$r_{\Delta\pi}$	0.14	Inflation growth coefficient
r_y	0.099	Output coefficient
$r_{\Delta y}$	0.159	Output growth coefficient
ρ	0.961	Degree of interest-rate smoothing
ρ_{ε^l}	0.889	Autocorrelation coefficient for labor supply shock
ρ_{ε^a}	0.823	Autocorrelation coefficient for productivity shock
ρ_{ε^b}	0.855	Autocorrelation coefficient for consumption preference shock
ρ_{ε^g}	0.949	Autocorrelation coefficient for government spending shock
$\rho_{\bar{\pi}}$	0.924	Autocorrelation coefficient for inflation objective shock
ρ_{ε^i}	0.927	Autocorrelation coefficient for investment shock
ς_{η^l}	3.52	Standard deviation of labor supply shock
ς_{η^a}	0.598	Standard deviation of productivity shock
ς_{η^b}	0.336	Standard deviation of consumption preference shock
ς_{η^g}	0.325	Standard deviation of government spending shock
$\varsigma_{\eta^{\bar{\pi}}}$	0.017	Standard deviation of inflation objective shock
ς_{η^i}	0.085	Standard deviation of investment shock
ς_{η^r}	0.081	Standard deviation of interest-rate shock
ς_{η^p}	0.16	Standard deviation of price mark-up shock
ς_{η^w}	0.289	Standard deviation of real-wage mark-up shock
ς_{η^q}	0.604	Standard deviation of equity-premium shock

Table 2: DSGE-VAR prior information.

Parameter	Domain	Prior PDF	Prior Mean	Prior Std. Dev.
γ_p	$[0,1)$	Beta	0.75	0.15
γ_w	$[0,1)$	Beta	0.75	0.15
ξ_p	$[0,1)$	Beta	0.75	0.05
ξ_w	$[0,1)$	Beta	0.75	0.05
r_π	$(-\infty,+\infty)$	Normal	1.7	0.1
$r_{\Delta\pi}$	$(-\infty,+\infty)$	Normal	0.3	0.1
r_y	$(-\infty,+\infty)$	Normal	0.125	0.05
$r_{\Delta y}$	$(-\infty,+\infty)$	Normal	0.0625	0.05
ρ	$[0,1)$	Beta	0.8	0.1
ρ_{ε^l}	$[0,1)$	Beta	0.85	0.1
ρ_{ε^a}	$[0,1)$	Beta	0.85	0.1
ρ_{ε^b}	$[0,1)$	Beta	0.85	0.1
ρ_{ε^g}	$[0,1)$	Beta	0.85	0.1
$\rho_{\bar{\pi}}$	$[0,1)$	Beta	0.85	0.1
ρ_{ε^i}	$[0,1)$	Beta	0.85	0.1
ς_{η^l}	$[0,+\infty)$	Inv. Gamma-1	1	$+\infty$
ς_{η^a}	$[0,+\infty)$	Inv. Gamma-1	0.4	$+\infty$
ς_{η^b}	$[0,+\infty)$	Inv. Gamma-1	0.2	$+\infty$
ς_{η^g}	$[0,+\infty)$	Inv. Gamma-1	0.3	$+\infty$
$\varsigma_{\eta^{\bar{\pi}}}$	$[0,+\infty)$	Inv. Gamma-1	0.02	$+\infty$
ς_{η^i}	$[0,+\infty)$	Inv. Gamma-1	0.1	$+\infty$
ς_{η^r}	$[0,+\infty)$	Inv. Gamma-1	0.1	$+\infty$
ς_{η^p}	$[0,+\infty)$	Inv. Gamma-1	0.15	$+\infty$
ς_{η^w}	$[0,+\infty)$	Inv. Gamma-1	0.25	$+\infty$
ς_{η^q}	$[0,+\infty)$	Inv. Gamma-1	0.4	$+\infty$
$\aleph$	$[0.1,10]$	Uniform		

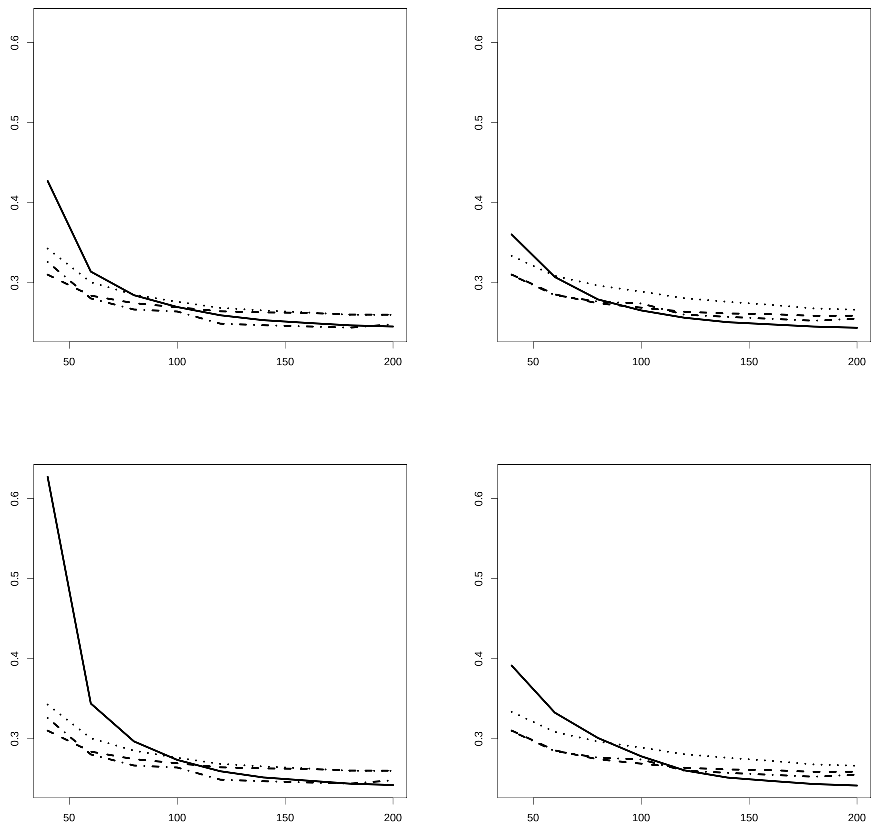

Figure 1: MSE for the four competing forecast models in single-step prediction. Solid curve stands for FAVAR, dashed for the univariate AR model, dotted and dash-dotted for bivariate VAR models. Left graphs for AIC search, right graphs for BIC search, top graphs for two factors FAVAR, bottom row for three factors.

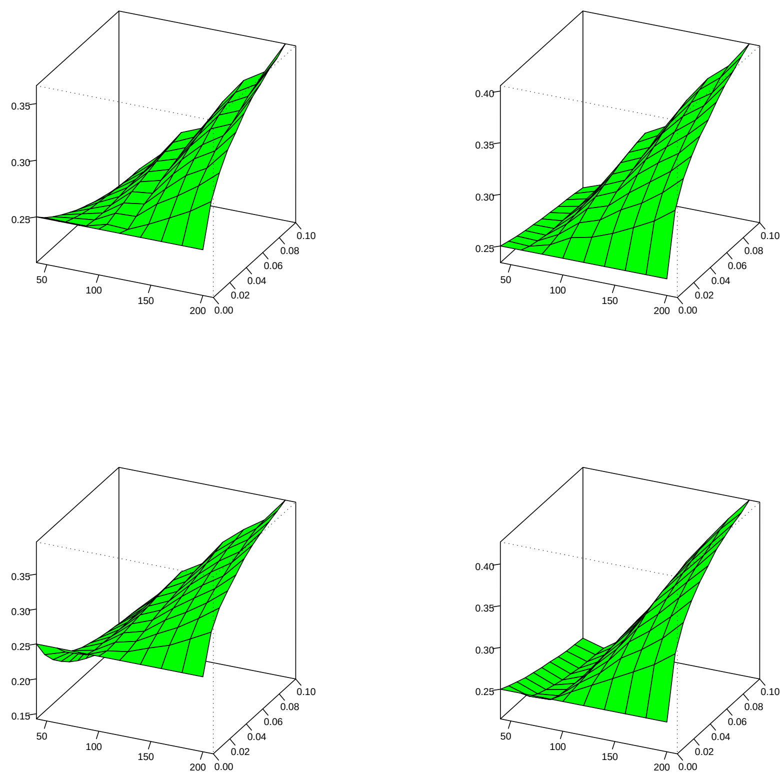

Figure 2: Weights allotted to the FAVAR model in dependence of the sample size and of the significance level for the encompassing test in single-step prediction. Arrangement of graphs as in Figure 1.

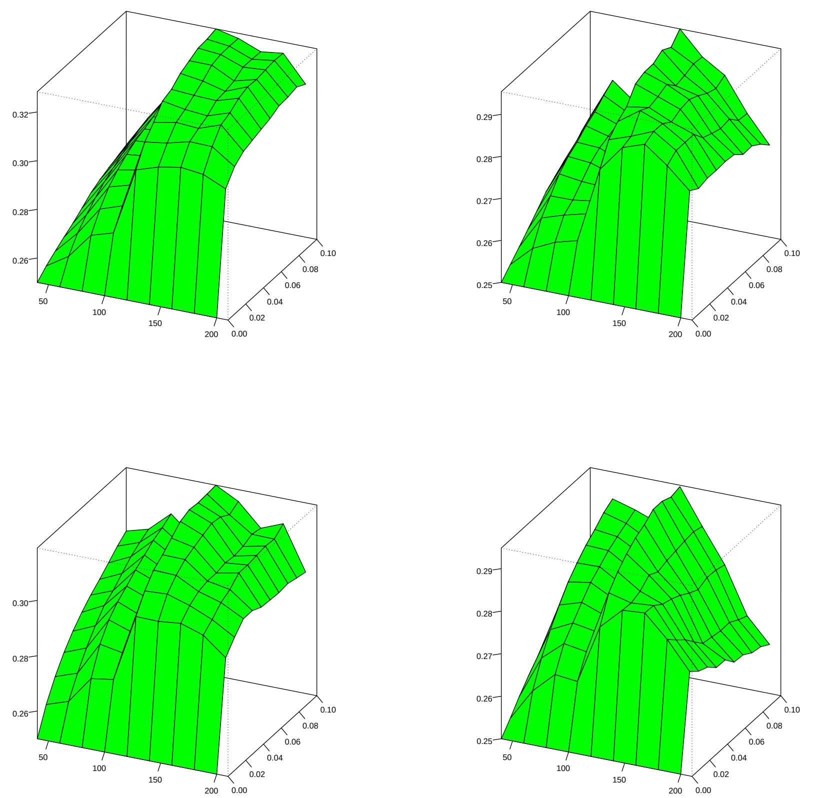

Figure 3: Weights allotted to the bivariate model with interest rate in dependence of the sample size and of the significance level for the encompassing test in single-step prediction. Arrangement of graphs as in Figure 1.

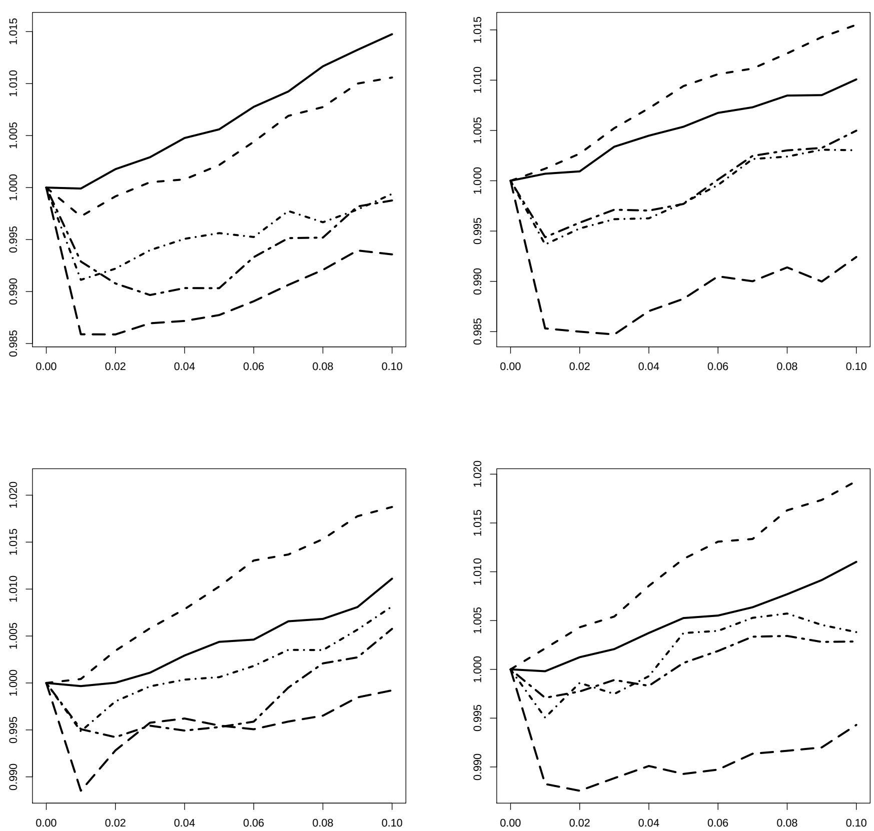

Figure 4: Ratios of test-based weighting MSE divided by the MSE from uniform weighting. Ordering of graphs see Figure 1. Solid curve for $N = 40$, short dashes for $N = 80$, dotted curve for $N = 120$, dash-dotted for $N = 160$, long dashes for $N = 200$.

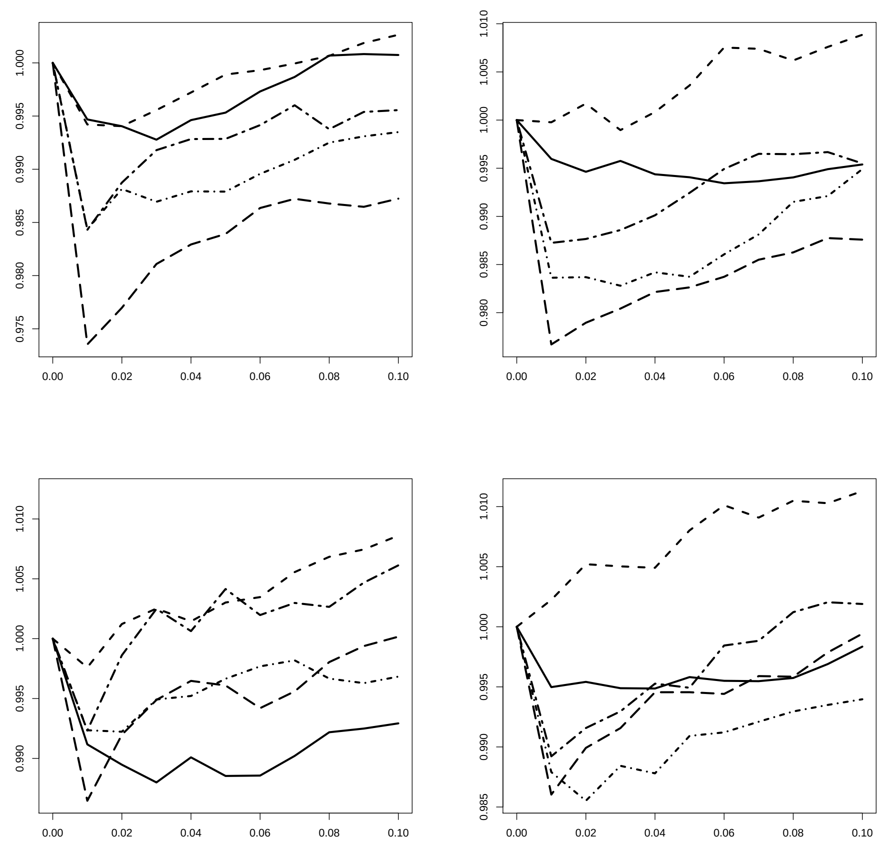

Figure 5: Ratios of test-based weighting MSE to MSE from uniform weighting in two-step prediction. Ordering of graphs see Figure 1. Solid curve for $N = 40$, short dashes for $N = 80$, dotted curve for $N = 120$, dash-dotted for $N = 160$, long dashes for $N = 200$.

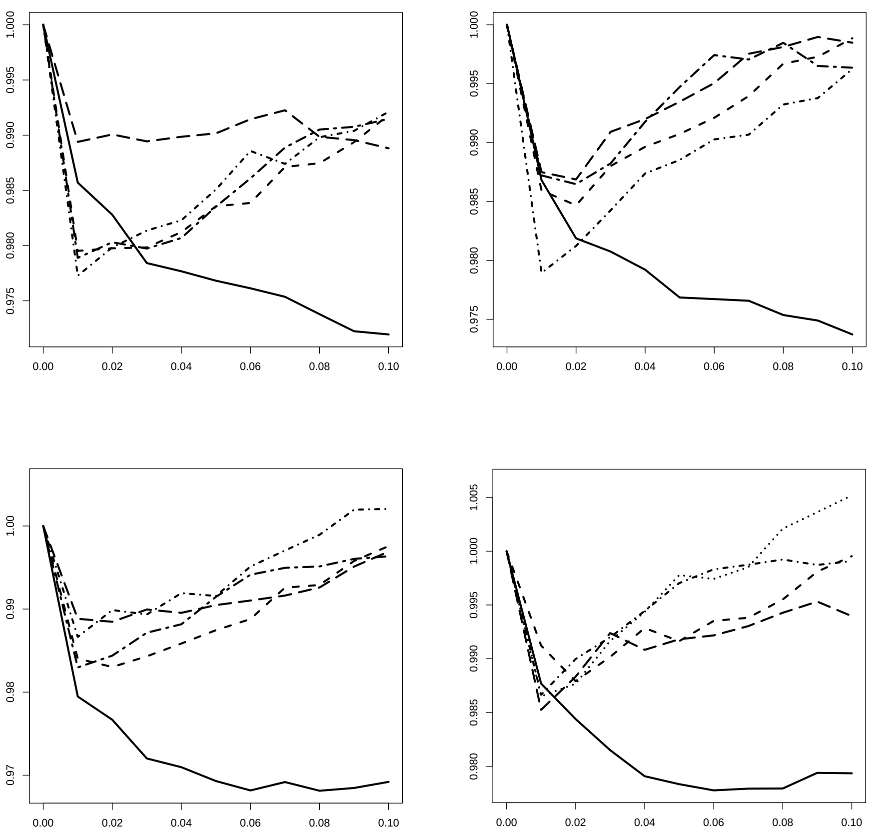

Figure 6: Ratios of test-based weighting MSE to MSE from uniform weighting in three-step prediction. Ordering of graphs see Figure 1. Solid curve for $N = 40$, short dashes for $N = 80$, dotted curve for $N = 120$, dash-dotted for $N = 160$, long dashes for $N = 200$.

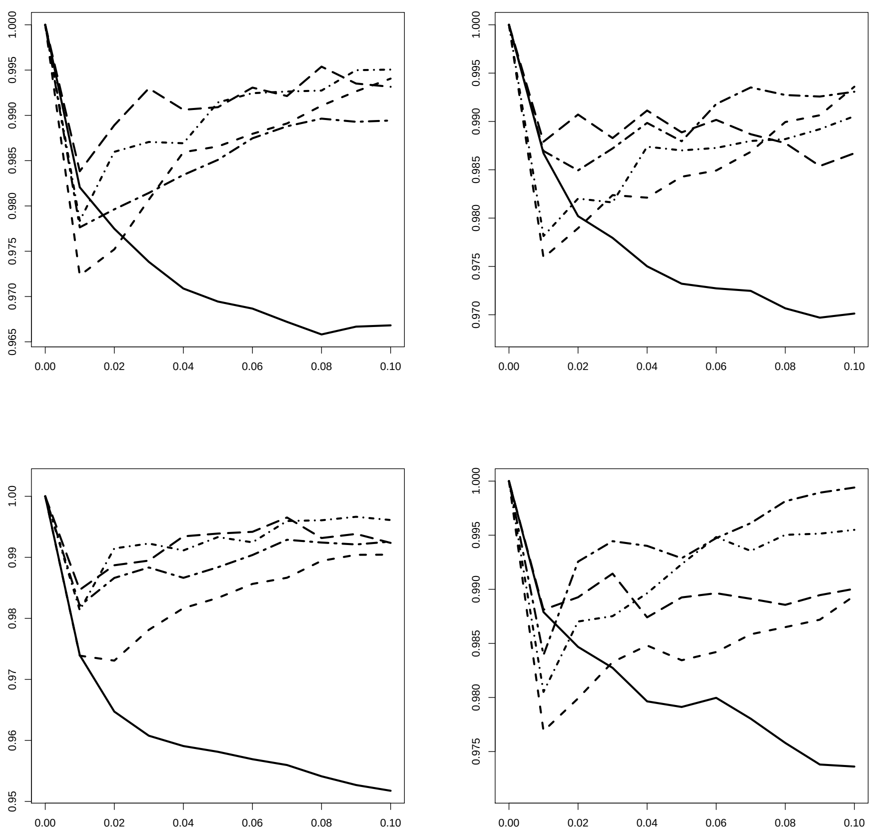

Figure 7: Ratios of test-based weighting MSE to MSE from uniform weighting in four-step prediction. Ordering of graphs see Figure 1. Solid curve for $N = 40$, short dashes for $N = 80$, dotted curve for $N = 120$, dash-dotted for $N = 160$, long dashes for $N = 200$.

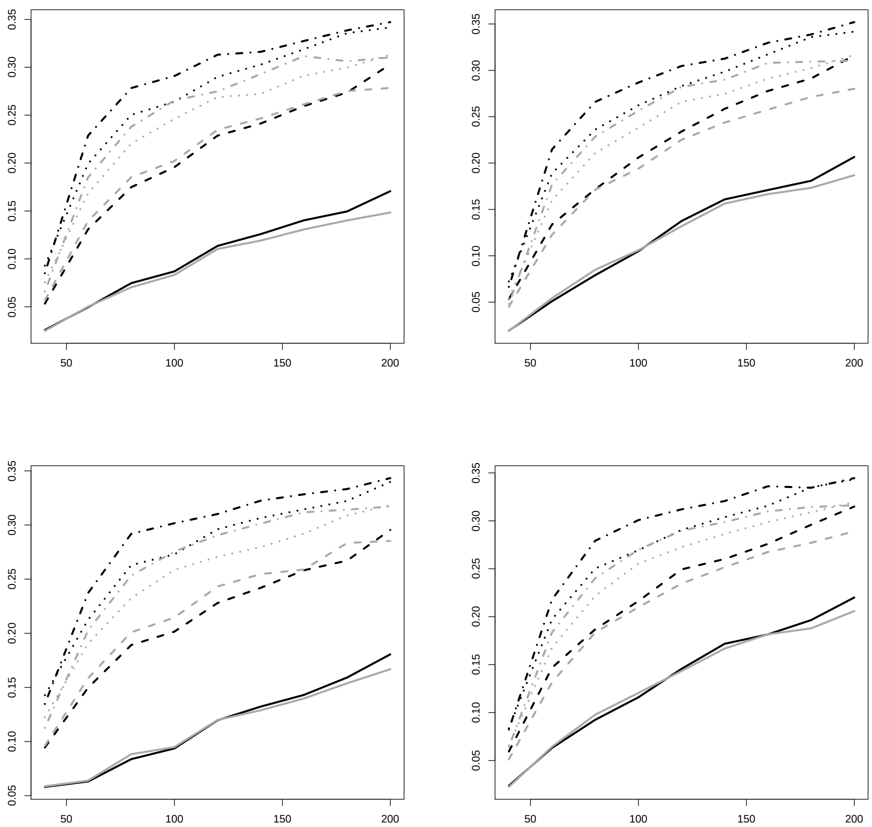

Figure 8: Frequency of a smaller absolute forecast error due to uniform weighting (gray curves) or 1% test-based weighting (black curves). Forecasts at horizons one (solid), two (dashed), three (dotted), and four (dash-dotted). Ordering of graphs see Figure 1.

Authors: Mauro Costantini, Ulrich Gunter, Robert M. Kunst

Title: Forecast Combination Based on Multiple Encompassing Tests in a Macroeconomic DSGE-VAR System

Reihe Ökonomie / Economics Series 292

Editor: Robert M. Kunst (Econometrics)
Associate Editors: Walter Fisher (Macroeconomics), Klaus Ritzberger (Microeconomics)

ISSN: 1605-7996

Stumpergasse 56, A-1060 Vienna • ☎ +43 1 59991-0 • Fax +43 1 59991-555 • http://www.ihs.ac.at

ISSN: 1605-7996