ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Berg, Tim Oliver; Henzel, Steffen

Working Paper

Point and Density Forecasts for the Euro Area Using Bayesian VARs

CESifo Working Paper, No. 4711

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Berg, Tim Oliver; Henzel, Steffen (2014) : Point and Density Forecasts for the Euro Area Using Bayesian VARs, CESifo Working Paper, No. 4711, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/96866

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute

CESifo Working Papers

www.cesifo.org/wp

Point and Density Forecasts for the Euro Area Using Bayesian VARs

Tim O. Berg
Steffen R. Henzel

CESIFO WORKING PAPER NO. 4711
CATEGORY 12: EMPIRICAL AND THEORETICAL METHODS
MARCH 2014

An electronic version of the paper may be downloaded
- *from the SSRN website: www.SSRN.com*
- *from the RePEc website: www.RePEc.org*
- *from the CESifo website: www.CESifo-group.org/wp*

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 4711*

Point and Density Forecasts for the Euro Area Using Bayesian VARs

Abstract

Recent articles suggest that a Bayesian vector autoregression (BVAR) with shrinkage is a good forecast device even when the number of variables is large. In this paper we evaluate different variants of the BVAR with respect to their forecast accuracy for euro area real GDP growth and HICP inflation. We consider BVAR averaging, Bayesian factor augmented VARs (BFAVARs), and large BVARs, which differ in the way information is condensed and shrinkage is implemented. We find that: (a) large BVARs produce accurate point forecasts but show a poor performance when the entire density is considered; (b) BVAR averaging shows the opposite pattern; (c) BFAVARs perform well under both evaluation criteria; (d) choosing the degree of shrinkage optimally does not improve forecast accuracy; (e) all variants except the large BVAR tend to be well calibrated for inflation but poorly calibrated for real GDP growth; (f) these findings are robust to several features of the forecast experiment.

JEL-Code: C110, C520, C530, E370.

Keywords: Bayesian vector autoregression, forecasting, model validation, large cross-section, euro area.

Tim O. Berg
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich
Poschingerstr. 5
Germany – 81679 Munich
berg@ifo.de

Steffen R. Henzel
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich
Poschingerstr. 5
Germany – 81679 Munich
henzel@ifo.de

February 28, 2014
We thank Peter Zorn, and participants of the 1stWorkshop on High Dimensional Time Series in Macroeconomics and Finance 2013 (Vienna), the 67th European Meeting of the Econometric Society 2013 (Gothenburg), the Annual Meeting of the German Economic Association 2013 (Düsseldorf), and the Ifo Macro Seminar (Munich) for their helpful comments and suggestions.

1 Introduction

When forecasting economic outcomes, a large set of indicators is wishful to avoid model misspecification. However, forecasting models with large cross-sections are often subject to overparameterization leading to unstable parameter estimates and inaccurate forecasts. In vector autoregressions (VARs) the number of parameters may easily exceed the number of observations which makes classical estimation infeasible in a data-rich environment. Traditionally, factor models have been used to handle large cross-sections and achieve dimension reduction.[1] In a seminal article Bańbura, Giannone, and Reichlin (2010) argue, however, that VARs can forecast better even when the number of variables is large. They propose Bayesian methods and impose additional information in form of a Minnesota-type prior to shrink the overparameterized VAR towards a parsimonious random walk (see also Carriero, Kapetanios, and Marcellino, 2009; Giannone, Lenza, and Primiceri, 2012; D'Agostino, Gambetti, and Giannone, 2013; Koop, 2013; Carriero, Clark, and Marcellino, 2014).

There exist, however, many possible ways to implement a Bayesian VAR (BVAR) for forecasting. In this paper we build on the results in Bańbura et al. (2010) and evaluate different variants of the BVAR which differ in the way information is condensed. In particular, we consider BVAR averaging, Bayesian factor augmented VARs (BFAVARs), as well as large BVARs. We also include the random walk variant and the autoregressive (AR) model as benchmarks. Moreover, we consider different specification choices which may also affect the performance of each variant. Along with the aggregation weights of the BVAR averaging and the number of factors of the BFAVAR, there are several ways to choose the degree of shrinkage. We compare the three predominant approaches of the related literature. First, we follow Bańbura et al. (2010) and select the shrinkage hyperparameter such that the average in-sample fit for our target variables is the same across variants during a training sample period. Second, we calibrate the hyperparameter by maximizing the marginal likelihood in each period as in Carriero et al. (2014). Finally, we follow Giannone et al. (2012) and estimate the tightness parameter by modelling it in a hierarchical fashion. Our study thus evaluates all major specification choices discussed in the prevalent literature. To our best knowledge, no one else has yet compared all these approaches within the same setup. The competing approaches are evaluated according to their out-of-sample forecast performance one step and four steps ahead. Specifically, we forecast the quarterly change in the euro area harmonized index of consumer prices (HICP) and the real

[1] The idea in this literature is that the information contained in a large number of indicator variables can be summarized by a rather small number of factors that are added to the variables of interest (see, e.g., Stock and Watson, 2002, 2005, 2006, 2011; Forni, Hallin, Lippi, and Reichlin, 2003, among others).

gross domestic product (GDP).

A feature of the analyses in Bańbura et al. (2010) and related studies is that the authors compare forecast models of fairly different size. For instance Bańbura et al. (2010) consider systems with 3, 7, 20, and 131 variables. The BVARs hence produce forecasts conditional on different information sets, making it difficult to identify whether the relative performance of the variants is result of an efficient use of information or a richer information set. While these authors focus on the potential benefits of the latter, we aim at revealing possible differences among the competing approaches with respect to the former. For each variant of the BVAR we ensure that forecasts are produced conditional on all the information that is contained in our dataset. In order to evaluate whether adding indicator variables is useful at all, we compare all variants to both the random walk variant and a univariate AR model.

Few studies exist so far which evaluate BVAR forecasts for aggregate euro area data (see, e.g., Giannone, Lenza, Momferatou, and Onorante, 2010, for a rare exception).[2] Our dataset comprises up to 44 quarterly macroeconomic and financial indicators for the euro area spanning the years 1975 to 2011. While applications for the U.S. often build on datasets containing more than one hundred variables, we believe that such large cross-sections are typically not available for most countries. This assumption should at least be true when the time series dimension is required to be large as well. Thus, it is not clear whether conclusions drawn from the specific case of the U.S. translate to other forecast situations. In our view, we consider a set of indicators that most forecasters would probably label a typical dataset. Moreover, we emphasize at this point that the size of our cross-section is also appropriate with respect to all variants we consider. Even for BFAVARs it has been shown that about 40 series are sufficient to yield satisfactory forecast accuracy (see Bai and Ng, 2002; Boivin and Ng, 2006). On the other hand, it has been documented that large BVARs achieve a good forecast performance with about 20-25 variables (see, e.g., Bańbura et al., 2010; Giannone et al., 2012; Koop, 2013). Our baseline results are thus derived from a subset of 22 variables which is similar to that considered in the related literature whereas all 44 variables are considered in a robustness check.

While discriminating among BVARs based on the accuracy of their point forecasts is appropriate if the loss function of the forecaster depends solely on the forecast error, such a proceeding neglects the uncertainty surrounding the forecasts. Policymakers nowadays closely monitor the uncertainty that is associated with business cycle and inflation developments. The density forecasts of the Bank of England's Monetary Policy Committee and the *Sveriges Riksbank* are prominent examples (see, e.g., Mitchell and Hall, 2005; Boero, Smith, and Wallis,

[2]Results for Germany are provided by Pirschel and Wolters (2014).

2011; Knüppel and Schultefrankenfeld, 2012, among others). We hence also evaluate the density forecasts and rank the different BVAR variants on the basis of their predictive likelihood, which is the standard tool to compare density forecasts in a Bayesian setting (see, e.g., Geweke and Amisano, 2010; Clark, 2011; Giannone et al., 2012; D'Agostino et al., 2013; Koop, 2013; Carriero et al., 2014, among others). In addition to this forecast competition, we also utilize calibration tests to assess the performance of the density forecasts in absolute terms. Given that recent work by Rossi and Sehkposyan (2013) suggests that density forecasts generated by BVARs are often not a reasonable description of actual uncertainty, it is important to investigate which specification choices help to achieve correct calibration.

The main findings are as follows. Regarding point forecasts we find that all BVARs outperform the random walk variant for HICP inflation at both horizons. The differences in forecast accuracy among these variants are small and often insignificant. For real GDP growth the large BVAR delivers the best forecast one step ahead but cannot improve on the random walk four steps ahead. Moreover, we find that neither setting the shrinkage parameter optimally nor estimating it in a hierarchical fashion, in general, helps to improve the forecast accuracy compared to the shrinkage procedure of Bańbura et al. (2010). A similar conclusion can be drawn for the BVAR averaging. Choosing the aggregation weights in an optimal way does not improve forecast accuracy compared to a naïve equal weighting scheme. Finally, we find that BFAVARs with 1 and 3 factors perform satisfactory across target variables and forecast horizons. With respect to the density forecasts we obtain evidence that an accurate point forecast does not necessarily imply a high predictive likelihood and *vice versa*. In particular, it turns out that the performance of the large BVARs deteriorates when the predictive density is considered, while that of the BVAR averaging improves. And we again find that choosing the degree of shrinkage optimally or estimating it in a hierarchical fashion does not change this result. Moreover, the calibration tests suggest that most BVARs produce density forecasts that appear to be a reasonable description of the true distribution for HICP inflation, but detect a poor calibration of almost all variants with respect to real GDP growth. Finally, we demonstrate that these findings are not overly sensitive to changes in the forecast setup.

The remainder of this paper is organized as follows. In Section 2 we describe our dataset. Section 3 develops the BVAR model and all variants we use to produce out-of-sample forecasts for euro area HICP inflation and real GDP growth. In Section 4 we explain our forecast experiment and present the main results. We evaluate the relative performance of the variants based on point and density forecasts. Moreover, we investigate how good the forecasts are in absolute terms. In Section 5 robustness checks are provided. Finally, Section 6 concludes.

2 Dataset

The full dataset comprises 44 quarterly euro area macroeconomic and financial time series covering the period 1975:1 to 2011:4. In case of aggregate euro area data both the cross-sectional and time series dimension approximately represent the maximal size available at the moment. The dataset includes the overall HICP and real GDP as the series of interest. The indicator variables cover the following seven categories: national accounts data, price indexes, international data, employment data, surveys, monetary aggregates, and financial data. For the baseline specification, we follow the common practice in the forecasting literature and achieve stationarity by differencing or log differencing (non-annualized) the data. We also consider the variables in levels in a robustness check, where we take the natural logarithm of most series, except of those that are already expressed in rates.

In most cases the series are obtained from the 12th update of the Area-wide Model (AWM) database[3] which is maintained by the European Central Bank (ECB). The historical series are backdated using individual country information in a coherent manner. The AWM database is the preferred source for researchers and policymakers alike interested in topics relevant for the euro area. In addition, we include survey data, monetary aggregates, and a share price index obtained from the Main Economic Indicators (MEI) database compiled by the Organisation for Economic Co-operation and Development (OECD). A detailed description of the dataset is provided in Appendix A.

The dataset thus includes all variables that a forecaster typically has on her wish list when forecasting euro area HICP inflation and real GDP growth. Besides the fact that forecasting euro area data is interesting in its own right, note that for countries other than the United States long time series for literally hundreds of indicator variables are not available. Hence, our dataset is also typical in the sense that it strikes a balance between the maximum availability of the cross-sectional and the time series dimension of the predictors.

Our baseline results are, however, derived from a subset of only 22 variables since the related literature found that BVARs with about 20-25 variables are in general superior to those including more variables (see, e.g., Bańbura et al., 2010; Giannone et al., 2012; Koop, 2013). In practice, a forecaster may choose a subset of indicators from all series available either based on past experience or on the basis of sound economic arguments. The latter consideration leads to information sets like the one we consider in the baseline specification. With respect to both the number and type of indicator variables, the baseline dataset is thus similar to those considered

[3]See also Fagan, Henry, and Mestre (2005). The vintage we use is as of September 2012.

in the related literature. We provide details on its exact composition in Appendix A. Note that we use the full dataset as a robustness check.

Finally, we have to stress at this point that consistent vintage data is not available since the euro area did not exist before January 1999. While the absence of real-time data is a potential drawback in studies where the objective is to conduct a realistic out-of-sample forecast experiment, we believe it is not in our case. The forecast experiment in this paper is understood as a model validation exercise designed to investigate which variant of the BVAR deals best with the problem of overparameterization that is inevitably inherent in models with large cross-sections. Evaluating out-of-sample forecasts is an appropriate and established procedure to do so since forecasts reflect all sources of error typically associated with the modeling of economic outcomes including parameter uncertainty and model misspecification.

3 Forecasting with a BVAR Model

In this section we develop the BVAR model and all the variants we use. Moreover, we discuss different specification choices for the hyperparameters and the lag length.

3.1 BVAR Model

We consider the following VAR model

$$y_t = c + B_1 y_{t-1} + ... + B_p y_{t-p} + u_t, \tag{1}$$

where y_t is a $n \times 1$ vector of variables including, among others, HICP inflation and real GDP growth; c is a $n \times 1$ vector of intercepts; B_i are $n \times n$ matrices of coefficients; $i = 1, ..., p$ denotes the lags included; u_t is a $n \times 1$ vector of normally distributed residual terms with zero mean and covariance matrix Σ; and data are available for $t = 1 - p, ..., T$. Let us denote $y = (y_1, ..., y_T)'$, $x_t = \left(y'_{t-1}, ..., y'_{t-p}, 1\right)'$, $x = (x_1, ..., x_T)'$, $B = (B_1, ..., B_p, c)'$, and $u = (u_1, ..., u_T)'$. The VAR in (1) can be rewritten as $y = xB + u$. Moreover, let $\beta = \text{vec}\,(B)$ with $\text{vec}\,(\cdot)$ being the column stacking operator and $k = n\,(1 + np)$. Then β is a $k \times 1$ vector containing all coefficients of the model.

In the forecast experiment we estimate the VAR on up to $n = 44$ variables including $p = 4$ lags of each variable (hence $k = 7,788$). Such a large dimensional system of multivariate regressions is, however, not estimable without imposing additional prior beliefs on the parameters. In addition, there is evidence that even VARs with only a handful of variables might benefit from imposing prior information (see, e.g., Robertson and Tallman, 1999, among others). We follow

common practice and use a variant of the Minnesota prior to deal with the dense parameterization of the model. The basic idea is that a random walk with drift is a reasonable description of the data generating process behind most macroeconomic and financial time series when considered in levels. In addition, the prior captures the belief that own lags are more informative than those of other variables and that more recent lags contain more information than more distant ones. For differenced data a white noise process is the appropriate prior. The VAR is hence centered around the prior mean $y_{i,t} = c_i + u_{i,t}$ (or $y_{i,t} = c_i + y_{i,t-1} + u_{i,t}$ for level data) and imposing the white noise prior amounts to shrinking the coefficients towards zero.

In contrast to the original Minnesota prior developed in Litterman (1980, 1986), we do not assume the residual covariance matrix Σ to be known and diagonal. Instead, we use a generalized version of the prior proposed in Kadiyala and Karlsson (1993, 1997) which allows for correlation among residual terms. The evidence in Bańbura et al. (2010) as well as Robertson and Tallman (1999) suggests that a generalized Minnesota prior produces accurate forecasts for major macroeconomic series such as GDP growth or inflation even though the $n(n+1)/2$ distinct elements of Σ have to be estimated on top of the k coefficients. In particular, we consider a conjugate Normal-Inverse-Wishart prior of the following form:

$$\Sigma \sim \text{IW}(\Psi, d) \quad \text{and} \quad \beta|\Sigma \sim \text{N}(b, \Sigma \otimes \Omega), \tag{2}$$

where $\otimes$ denotes the Kronecker product and the elements Ψ, d, b, and Ω are functions of hyperparameters. The conjugate prior implies a likelihood and posterior that come from the same family of distributions and hence makes Bayesian inference feasible for researchers even for n being large.[4] We follow Bańbura et al. (2010) and implement the prior by constructing the following set of artificial observations:

$$y^+ = \begin{bmatrix} \text{diag}(\delta_1\sigma_1, ..., \delta_n\sigma_n)/\lambda \\ 0_{n(p-1)\times n} \\ \text{diag}(\sigma_1, ..., \sigma_n) \\ 0_{1\times n} \end{bmatrix}, \quad x^+ = \begin{bmatrix} \text{diag}(1, 2, .., p) \otimes \text{diag}(\sigma_1, ..., \sigma_n)/\lambda & 0_{np\times 1} \\ 0_{n\times np} & 0_{n\times 1} \\ 0_{1\times np} & \epsilon \end{bmatrix}, \tag{3}$$

where $\text{diag}(\cdot)$ denotes a diagonal matrix. The hyperparameters δ_i are either set equal to 1, reflecting the prior belief that the variables are characterized by high persistence (level data), or

[4]Non-conjugate priors are an alternative to conjugate priors in systems with up to 20 variables but are not available for larger models (see Koop, 2013).

set to 0 for variables showing strong mean reversion (differenced data). The hyperparameters σ_i account for the different scale and variability of the series and are set equal to the standard deviation of a residual from a univariate autoregression for the variable $y_{i,t}$ on a training sample running from 1975:2 to 1990:1. The lag order is the same as in the VAR ($p = 4$). The hyperparameter ϵ is set to a very small number (10^{-4}), reflecting a diffuse prior for the intercept terms. Finally, the hyperparameter λ determines the degree of shrinkage and hence the tightness of the prior. As $\lambda \to \infty$ the prior becomes uninformative and posterior expectations coincide with the ordinary least squares (OLS) estimates. For $\lambda \to 0$ the posterior equals the prior and the information variables do not influence the estimation outcome. The hyperparameter λ is hence the key parameter in the BVAR and its calibration (or estimation) for each model variant is discussed in detail below.

To further improve the forecast accuracy of BVARs, the literature proposes to impose additional information in form of a "sum-of-coefficients" prior (see, e.g., Doan, Litterman, and Sims, 1984; Bańbura et al., 2010; Giannone et al., 2012, among others). This prior is implemented by generating n artificial observations and reflects the belief that a no-change forecast is a good forecast at the beginning of a sample period. In particular, we construct:

$$y^{++} = \text{diag}\left(\delta_1\mu_1, ..., \delta_n\mu_n\right)/\tau \ , \qquad x^{++} = \left[1_{1\times p} \otimes \text{diag}\left(\delta_1\mu_1, ..., \delta_n\mu_n\right)/\tau \quad 0_{n\times 1}\right] \ , \tag{4}$$

where $\tau = 10\lambda$ and the hyperparameters μ_i capture some prior belief about the average level of variable $y_{i,t}$. Consistent with the calibration of the σ_i's, we set μ_i equal to the average value of $y_{i,t}$ in the training sample period 1975:2 to 1990:1.

The artificial observations are added on top of the data matrices, which are then used for inference. The augmented regression model reads as

$$y^* = x^* B + u^*, \tag{5}$$

where $y^* = \left(y', y^{+'}, y^{++'}\right)'$, $x^* = \left(x', x^{+'}, x^{++'}\right)'$, and $u^* = \left(u', u^{+'}, u^{++'}\right)'$. Adding artificial observations solves the matrix inversion problem which arises in VARs with large cross-sections.

The posterior of the parameters can be computed in closed form as a function of the hyperparameters:

$$\Sigma|y \ \sim \ \text{IW}\left(\hat{\Sigma}, T+n+2\right) \quad \text{and} \quad \beta|\Sigma, y \sim \text{N}\left(\hat{\beta}, \Sigma \otimes \left(x^{*'}x^*\right)^{-1}\right), \tag{6}$$

where $\hat{\Sigma}$ and $\hat{\beta}$ are the covariance matrix and the coefficients from an OLS regression of y^* on x^*, respectively.

The one-step-ahead predictive density $p\left(y_{T+1}|\beta, \Sigma, y\right)$ is also available in closed form. Moreover, we fully exploit the conjugacy of the prior and make use of the fact that after integrating out Σ, draws for the one-step-ahead predictive density can be directly obtained from its marginal posterior which follows a multivariate t-distribution (see Koop, 2013). In particular, the predictive mean is given by

$$\mathrm{E}\left(y_{T+1}|y\right) = \left(x_{T+1}\hat{B}\right)', \tag{7}$$

while the predictive covariance matrix is

$$\mathrm{var}\left(y_{T+1}|y\right) = \frac{1}{T+n}\left(1 + x_{T+1}\left(x^{*\prime}x^*\right)^{-1}x'_{T+1}\right)\hat{\Sigma}. \tag{8}$$

Drawing from the marginal posterior is convenient and fast, even for systems with more than 20 variables. When forecasting more than one period ahead, however, an analytical expression for the predictive mean and its covariance matrix is not available. In this case we follow Koop (2013) and rely on the direct forecasting method. In particular, we consider

$$y_t = c_h + B_{1,h}y_{t-(h-1)-1} + ... + B_{p,h}y_{t-(h-1)-p} + u_{t,h}. \tag{9}$$

For $h > 1$ the problem is hence converted into one which involves only one-step-ahead forecasting, and draws for the predictive density can again be obtained from the marginal posterior.

3.2 Model Variants

We consider the following variants of the BVAR which provide us with different ways to deal with the overparameterization problem and condense the information in the dataset.

Random Walk As a benchmark we consider the random walk variant. Random walk forecasts are obtained by imposing a dogmatic prior (hence $\lambda \to 0$).[5] The posterior beliefs are thus not shaped by the indicator variables, and the random walk is a natural benchmark to investigate whether using these series is useful at all in forecasting. The predictive mean for HICP inflation and real GDP growth is therefore equal to the estimated constant. Note that this term may adapt over time, providing a naïve but competitive benchmark for any forecast model.

[5]In practice we set $\lambda = 10^{-4}$.

BVAR Averaging with Equal Weights Recent studies argue that averaging forecasts is a simple though successful method to handle a large dataset and improve out-of-sample forecast accuracy (see, e.g., Clark and McCracken, 2010; Aiolfi, Capistrán, and Timmermann, 2011; Barnett, Mumtaz, and Theodoridis, 2012; Henzel and Mayr, 2013; Wolters, 2014, among others). The idea is motivated by portfolio diversification or hedging arguments, guaranteeing insurance against large forecast errors. In this second variant we thus estimate $m = n - 2$ three-variable BVARs each including HICP inflation and real GDP growth as well as one indicator variable at a time. We then average the forecasts of the individual models. By segmenting the set of indicators and estimating a battery of parsimonious models, pooling is a way to condense information and avoid parameter proliferation. Therefore, we assume that no shrinkage is needed, and we impose an uninformative prior ($\lambda \to \infty$).[6] We obtain the predictive density for HICP inflation and real GDP growth by combining predictive densities of different models as follows:

$$p\left(y_{T+1}|y, w_{T-1}\right) = \sum_{i=1}^{m} w_{T-1,i}\, p_i\left(y_{T+1}|y\right), \tag{10}$$

where $w_{T-1} = (w_{T-1,1} ... w_{T-1,m})$ contains nonnegative weights which sum to one. Thus, the linear combination of predictive densities is again a predictive density (see Hall and Mitchell, 2007; Aastveit, Gerdrup, Jore, and Thorsrud, 2011; Geweke and Amisano, 2011, 2012, among others). In this variant we fix $w_{T-1,i}$ at $1/m$ for all T and i. Such a naïve equal weighting scheme has been found to perform astonishingly well by several authors (see, e.g., Clark and McCracken, 2010; Henzel and Mayr, 2013, among others).

BVAR Averaging with Optimal Weights The third variant is identical to the previous one with the exception that the weights are now chosen optimally and computed in real time. In particular, we follow Aastveit et al. (2011) as well as Geweke and Amisano (2011, 2012) and calculate in each period T that vector w^*_{T-1} which maximizes the historical (log) predictive likelihood:

$$w^*_{T-1} = \text{argmax} \sum_{s=1}^{T-1} \log \left[\sum_{i=1}^{m} w_{T-1,i}\, p_i\left(y_s|y\right) \right]. \tag{11}$$

The real-time optimal weights vector can be easily computed using a standard numerical solution algorithm.[7] It should be noted, however, that maximizing the historical predictive likelihood does not necessarily guarantee a better out-of-sample forecast accuracy compared to a

[6] In practice we set $\lambda = 10^4$.

[7] We use the Matlab routine "fmincon".

naïve equal weighting scheme since the approach introduces an additional estimation uncertainty regarding the weights. Moreover, historical weights might be a poor approximation of current weights, particularly when structural breaks are present in the data.

BFAVAR with 1 Factor It is often argued that factor augmented regression models are successful in achieving dimension reduction when forecasting macroeconomic time series (see, e.g. Stock and Watson, 2002, 2005, 2006, 2011; Forni et al., 2003; Barhoumi, Darné, and Ferrara, 2013, among others). The idea is that a bulk of the variation in the indicator variables may be explained by a rather small number of factors which are added to a model with the variables of interest. Here, we consider a BFAVAR including HICP inflation and real GDP growth and one factor. To estimate the parameters we follow Bernanke, Boivin, and Eliasz (2005) and use a two-step approach. In the first step we difference the indicator variables (if needed) to achieve stationarity and standardize them by subtracting the mean and dividing by the standard deviation. In principal, we obtain k factors by extracting the first k principal components from these standardized series. In the second step we augment the BVAR by the first factor and estimate the system using the Bayesian methods outlined above. As the BFAVAR system consists of only three variables, overparameterization is not a major problem and little shrinkage is needed.

BFAVAR with 3 Factors Since adding one factor might not be sufficient to capture the dynamics in our dataset, we augment the BVAR from above with $k = 3$ factors. Given that a VAR with five variables is already large and likely subject to overparameterization, we use Bayesian shrinkage to further reduce the dimension of the system. Note that both BFAVAR variants hence combine the advantages of the factor approach with that of Bayesian shrinkage.

Large BVAR In the last variant we estimate the BVAR on all the 22 series at the same time. The literature refers to this variant as a large BVAR (see, e.g., Bańbura et al., 2010; Giannone et al., 2012; Koop, 2013, among others). In order to deal with the dense parameterization of the model, the model is estimated using Bayesian shrinkage. Given the large dimensionality of the system, we shrink a lot in this case.

Finally, the BVAR variants are accompanied by univariate AR models for HICP inflation and real GDP growth. Such models are the industry standard and an additional competitive benchmark besides the random walk variant. We estimate the AR models using an uninformative prior, while the lag length p is selected via the Bayesian Information Criterion (BIC).

3.3 Choice for Hyperparameters

There are different ways to choose the degree of shrinkage λ which will have an impact on the forecast performance of each BVAR variant. As emphasized by De Mol, Giannone, and Reichlin (2008), a requirement for the hyperparameter λ is that the degree of shrinkage increases with the number of parameters, i.e. $\lambda \to 0$ if $n \to \infty$.

Bańbura, Giannone, and Reichlin (2010) One way to determine λ rests on the assumption that small VAR systems are parsimonious and hence do not suffer from overparameterization. The procedure follows Bańbura et al. (2010) by choosing λ such that the in-sample fit of a BVAR is the same as that of a parsimonious VAR. Specifically, we set the degree of shrinkage such that the average in-sample fit of each BVAR variant for HICP inflation and real GDP growth during the training sample period 1975:2 to 1990:1 is identical to the fit of the three-dimensional BVARs used for BVAR averaging. We obtain the desired magnitude of fit by performing a search over a fine grid for λ. This procedure ensures that we shrink more when the size of the model increases, and it maintains the comparability across variants. For that reason we also apply Bayesian shrinkage to the BFAVAR with one factor.

We should emphasize, however, that this procedure is not optimal in the sense that the choice for λ is not based on any optimality criterion. Moreover, the shrinkage parameter does not evolve over time but is kept fix at the value obtained during the training sample period. Carriero et al. (2014) and Giannone et al. (2012) have recently proposed different procedures to overcome these potential drawbacks. Both methods are straightforward to be implemented, and we consider them for both BFAVAR variants as well as the large BVAR. In this context we also discuss how the lag length p, which has been fixed so far at the frequency of the data, may be chosen optimally.

Carriero, Clark, and Marcellino (2014) One possibility is to choose the optimal degree of shrinkage λ_T^* and the optimal lag length p_T^* at each point in time by maximizing the (log) marginal likelihood of the model:

$$(\lambda_T^*, p_T^*) = \text{argmax} \log p\left(y\right), \tag{12}$$

while the marginal likelihood $p\left(y\right)$ can be obtained in closed form[8] by integrating out the uncertainty about the parameters of the model θ, i.e.

[8]We provide the analytical expression for the marginal likelihood in Appendix B.

$$p(y) = \int p(y|\theta)\, p(\theta)\, d\theta. \tag{13}$$

We optimize over the two-dimensional grid $\lambda \in (0.01, 0.02, ..., 5)$ and $p \in (1, 2, ..., 8)$ and choose in each period that pair which maximizes $p(y)$. Note that the optimal value for λ decreases when more lags are included. Moreover, we update the remaining hyperparameters in each period accordingly (rolling training sample).

Giannone, Lenza, and Primiceri (2012) Alternatively, λ can be treated as an additional unknown parameter and estimated in a hierarchical fashion. The procedure is even more flexible since it also relaxes some of the simplifying choices made for the remaining hyperparameters. Both the main diagonal of the prior covariance matrix Σ and the scaling parameter τ are also treated as unknown. This variant thus requires to add one more layer to the prior structure by placing a prior on the hyperparameters - a hyperprior. Let γ collect all the unknown hyperparameters of the model, i.e. $\gamma = (\lambda, \sigma_1, ..., \sigma_n, \tau)'$. It can be shown that the marginal posterior for γ, i.e. after integrating out the uncertainty about the model's parameters, is:

$$p(\gamma|y) \propto p(y|\gamma) \cdot p(\gamma), \tag{14}$$

where $p(y|\gamma)$ is the marginal likelihood of the model (conditional on γ) and $p(\gamma)$ is the hyperprior. The latter reflects how confident we are about the values for γ. We follow common practice and choose relatively diffuse hyperpriors. For λ and τ we use Gamma densities with mode equal to 0.2 and 1, and standard deviations of 0.4 and 1, respectively. The prior on the diagonal elements of Σ is an Inverse-Gamma with scale and shape equal to 0.005^2. Since the joint posterior distribution for the predictive density and γ is not available in closed form, we use a Metropolis-Hastings algorithm to simulate the distribution.[9] The sampler uses Markov chain Monte Carlo methods and generates γ from the marginal posterior with a Metropolis-Hastings update. After convergence of the sampler, the predictive density can be drawn from its marginal distribution conditional on γ. In contrast to the procedures proposed in Bańbura et al. (2010) and Carriero et al. (2014), this variant hence accounts for the uncertainty related to the choice of the hyperparameters such as λ. As a result, this procedure may have an impact on the shape of the predictive density. Finally, it should be noted that the lag length is again set equal to four and that parsimony is achieved only via optimal estimation of the hyperparameters.

[9]The Metropolis-Hastings algorithm is outlined in detail in Appendix C.

4 Model Validation

In this section we evaluate the performance of the different BVAR variants in terms of out-of-sample forecast accuracy. We first explain our forecast experiment and discuss the choice for the shrinkage parameter across model variants. We then present results for the predictive mean and the entire predictive density of HICP inflation and real GDP growth. Moreover, we assess how good the forecasts are in absolute terms.

4.1 Forecast Experiment

We use each BVAR variant to produce out-of-sample forecasts for HICP inflation and real GDP growth for two horizons ($h = 1$ and $h = 4$). For $h = 1$ we start with the sample period from 1976:1 to 1990:4 and generate the predictive density of each variant for 1991:1. This procedure is iterated forward until 2011:3, producing forecasts for 2011:4, always using the most recent 15 years of data and hence yields a sequence of 84 density forecasts for each variant. For $h = 4$ we start with the sample period from 1975:2 to 1990:1, also producing predictive densities for 1991:1 and iterate accordingly. The evaluation period thus runs from 1991:1 to 2011:4 for both horizons. Since our evaluation period is relatively long, we investigate the subsample stability of the results in a robustness check.

For the baseline specification we prefer a rolling-window forecast scheme to a recursive (or expanding) scheme, which uses all the past observations, since a rolling scheme may better handle parameter instabilities that are likely to be present in aggregate euro area data. Moreover, it is consistent with the conventional view that more recent observations are more informative than those at the very beginning of a sample period. In addition, the statistical tests we use to compare the forecast performance explicitly build on an asymptotically non-vanishing estimation uncertainty; an assumption which would, however, be violated in an expanding-window forecast scheme (see Giacomini and White, 2006; Amisano and Giacomini, 2007). In a robustness check, however, we evaluate how sensitive our results are with respect to this setup and repeat the forecast experiment using a recursive scheme.

The choice for the size of the estimation window is motivated by two competing influences. On the one hand, the window should not be too small since otherwise a meaningful estimation of the BVAR would not be possible. On the other hand, the window should not be too large since otherwise the sequence of density forecasts would be too short for inference. In particular, the likelihood ratio tests require a relatively large number of observations. An estimation window of 15 years seems to account well for both concerns.

4.2 Degree of Shrinkage

We document the choice for the shrinkage parameter λ in the first ($h = 1$) and second ($h = 4$) column of Table 1. Since we estimate a different model for each forecast horizon we also obtain a different λ for each horizon. The numbers in the first panel are the result of the calibration procedure as proposed by Bańbura et al. (2010) which equates the average in-sample fit for HICP inflation and real GDP growth across variants (except for the random walk benchmark for which $\lambda = 0$). As we have emphasized above, we assume that the overparameterization problem does not occur for the BVAR averaging, and we use an uninformative prior ($\lambda = \infty$) in this case. For the BFAVAR with one factor, it turns out that parameter proliferation is not really an issue, and we shrink only a little (0.328 for $h = 1$ and 0.519 for $h = 4$). In case of the BFAVAR with three factors the prior becomes tighter and λ is 0.105 and 0.151, respectively. For the large BVAR with all 22 variables we have to impose a lot more shrinkage, and λ reduces to 0.052 at $h = 1$ and 0.051 at $h = 4$. In the second panel λ is obtained by maximizing the marginal likelihood as in Carriero et al. (2014). Since the optimal choice for the tightness parameter varies over time (it does so only a little actually), we report the average choice for λ across periods. Given that the lag length p is determined simultaneously, the numbers are not directly comparable to those in the first panel. It turns out, however, that the optimal choice for p is in most periods 1 or 2, which requires less shrinkage than a model with $p = 4$. In the third panel we document the average posterior mean for λ across periods when a hyperprior is used as in Giannone et al. (2012). The numbers are similar to those in the second panel although the lag length is set to $p = 4$. However, in the latter case the remaining hyperparameters (Σ, τ) are also estimated in real time, and numbers are hence hard to compare. In particular, the elements of Σ show substantial variation over time.

4.3 Predictive Mean

We first evaluate the different BVAR variants and the AR model by the accuracy of their point forecasts. Let the superscript m denote the predictive mean. We measure out-of-sample forecast accuracy for model j with respect to variable i at horizon h in terms of RMSE:

$$\text{RMSE}_{i,h}^{j} = \sqrt{\frac{1}{T_1 - T_0 - H + 1} \sum_{T=T_0+H}^{T_1} \left(y_{i,T|T-h}^{m,j} - y_{i,T} \right)^2}, \tag{15}$$

where $H = 4$ is the maximal forecast horizon, and $T_0 + H = 64$ and $T_1 = 147$ are the start (1991:1) and end (2011:4) of the evaluation period, respectively. For a forecaster with a quadratic loss function the predictive mean is the optimal forecast and the RMSE is an appropriate measure to

Table 1: Root Mean Squared Error

	λ		HICP Inflation		GDP Growth	
	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$
M01: Random Walk	0		0.39	0.43	0.62	0.63
M02: BVAR Averaging Equal Weights	∞		0.26	0.33	0.57	0.74
M03: BVAR Averaging Optimal Weights	∞		0.26	0.33	0.52	0.75
M04: BFAVAR 1 Factor	0.328	0.519	0.25	0.29	0.50	0.69
M05: BFAVAR 3 Factors	0.105	0.151	0.26	0.28	0.48	0.65
M06: Large BVAR	0.052	0.051	0.26	0.30	0.45	0.63
M07: BFAVAR 1 Factor, p and λ optimal	0.356	0.267	0.26	0.29	0.52	0.66
M08: BFAVAR 3 Factors, p and λ optimal	0.316	0.258	0.26	0.29	0.48	0.67
M09: Large BVAR, p and λ optimal	0.218	0.171	0.26	0.28	0.44	0.65
M10: BFAVAR 1 Factor, hyperpriors	0.367	0.272	0.26	0.30	0.50	0.66
M11: BFAVAR 3 Factors, hyperpriors	0.314	0.271	0.25	0.30	0.46	0.68
M12: Large BVAR, hyperpriors	0.264	0.203	0.25	0.29	0.44	0.67
M13: AR(p), p optimal	n/a		0.27	0.31	0.56	0.64

Notes: this table shows the RMSE for different BVAR variants and an AR(p) model for which p is chosen by the Bayesian Information Criterion (BIC). The evaluation period runs from 1991:1 to 2011:4 and the forecast horizon h is in quarters. For $h = 4$ we use a direct forecast. The lower the RMSE is the better is the forecast accuracy of a model. The BVAR variants include the random walk, the BVAR averaging with equal weights and optimal weights obtained by maximizing the historical predictive density, the BFAVAR with 1 and 3 factor(s), and the large BVAR with 22 variables. The hyperparameter λ is set in the first panel such that the average in-sample fit for HICP inflation and GDP growth of all BVAR variants is the same during the training sample 1975:2 to 1990:1, while $p = 4$. In the second panel p and λ are chosen by maximizing the marginal likelihood. We report the average choice for λ across periods. In the third panel we impose a prior on the hyperparameters and estimate the variants in a hierarchical fashion, while $p = 4$. We report the average posterior mean for λ across periods.

discriminate among models (see Weiss, 1996). In Table 1 we report the RMSE for the different BVAR variants as well as the additional AR benchmark model.

With respect to the forecast accuracy of the different BVAR variants the following results emerge from Table 1. First, the random walk benchmark produces the worst forecast for HICP inflation at $h = 1$ and $h = 4$. When using a large BVAR for instance the forecaster could cut the RMSE by about one third at both horizons, independent of how the shrinkage parameter or the lag length are chosen. Substantial improvements compared to the random walk forecast are also possible with all other variants.[10] This finding may either be the result of the richer dynamics

[10]Table D.1 in Appendix D reveals that these differences in forecast ability are significant at a 5 percent

of the other BVAR variants compared to the random walk benchmark or due to the additional information contained in the set of indicators. Given that the AR model is also inferior to almost all BVAR variants (though only slightly), we believe that the indicator variables indeed contain valuable information regarding the prospective path of HICP inflation.

Second, we obtain rather small differences in forecast accuracy for HICP inflation among the BVAR variants. Only for $h = 4$, the BVAR averaging, both with equal and optimal weights, appears to underperform the other variants. Overall, the RMSE ranges between 0.25 and 0.26 at $h = 1$ as well as 0.28 and 0.33 at $h = 4$. In Table D.1 in Appendix D we provide a Giacomini and White (2006) test of conditional predictive ability of the point forecast. It turns out that the differences among the variants are insignificant in the majority of cases. Apparently, the way information is condensed within the BVAR setting has a limited impact on the forecast accuracy of the predictive mean for HICP inflation. Moreover, the forecast accuracy does hardly change when the lag length, the hyperparameters, or the aggregation weights are chosen optimally.

Third, for real GDP growth, all models tend to outperform the random walk benchmark at $h = 1$. In contrast to HICP inflation, the random walk (jointly with one large BVAR variant) delivers the best forecast for real GDP growth at $h = 4$, closely tracked by the AR model, suggesting that the set of indicators has limited information about quarterly changes in real GDP one year ahead. This relatively poor forecast performance of most BVAR variants with respect to real GDP growth in the longer run is probably due to the fact that the series shows little to no persistence and unpredictable shocks may thus account for large parts of the dynamics.

Fourth, the differences among the BVAR variants are somewhat larger than for HICP inflation. The RMSE varies between 0.44 and 0.57 at $h = 1$ as well as 0.63 and 0.75 at $h = 4$. The large BVAR with p and λ optimized and that with hyperpriors deliver the best forecast performance at $h = 1$. Both variants perform significantly better than the BVAR averaging (compare Table D.1 in Appendix D). In most cases, the differences in predictive ability between the BVAR variants are, however, insignificant. Overall, there appears to be a slight dominance of the three large BVAR variants but the versions of the BFAVAR with 3 factors are also competitive.

4.4 Predictive Density

In the previous subsection we have compared different models in terms of RMSE, which is appropriate if the forecaster is concerned only about the accuracy of the predictive mean but indif-

level according to a Giacomini and White (2006) test of conditional predictive ability.

ferent to the uncertainty that is surrounding it. In this subsection we relax this strong assumption and consider the entire density of the forecasts. In particular, we rank BVAR variants based on the predictive likelihood, which is a convenient way of comparing models in a Bayesian setting and has become the standard in the related literature (see, e.g., Geweke and Amisano, 2010; Giannone et al., 2012; D'Agostino et al., 2013; Koop, 2013, among others). The larger the predictive likelihood is the higher is the probability that a variant generates a forecast that equals (or is close to) the realized value of the variable given the history of the data (y) and the parameters (β, Σ). The perfect BVAR variant would hence be characterized by a forecast which attaches a probability of 100% to the actual outcome.

In Table 2 we document the average of the log predictive likelihood of model j with respect to variable i at horizon h evaluated at the realization $y_{i,T}$:

$$\mathrm{PL}_{i,h}^{j} = \frac{1}{T_1 - T_0 - H + 1} \sum_{T=T_0+H}^{T_1} \log p\left(y_{i,T} | y, j, h\right). \tag{16}$$

Table 2 reveals that the random walk produces the worst forecast for HICP inflation at both horizons, which is in line with our previous results for the point forecasts. However, we find that the ranking of the BVAR variants changes when the focus shifts from the predictive mean to the entire density of the forecasts. Particularly, the performance of the large BVAR deteriorates. In most cases, it is now not much better than the AR model, if at all. Despite its quite accurate mean forecast, the large BVAR apparently attaches a too low probability to events that actually occur. Even when the lag length and the shrinkage parameter are chosen optimally, or a prior on the hyperparameters is imposed, the large BVAR variants underperform the competing approaches. By contrast, the BFAVAR with three factors still performs well when forecast uncertainty is taken into account. When compared to the other variants it is now delivering the best forecast at both horizons, while the BFAVAR with only one factor is also competitive. For both BFAVARs we find that choosing the lag length or the shrinkage parameter in an optimal way does, in general, not help to improve predictive ability. Notably, the BVAR averaging, both with equal and optimal weights, which underperforms the competing approaches in terms of RMSE, now ranks between the BFAVARs and the large BVARs. Optimizing the aggregation weights with respect to in-sample forecast accuracy does not improve out-of-sample predictive ability.

For both horizons of real GDP growth, the random walk and the large BVAR jointly display the poorest forecast performance. As in the case of HICP inflation, the ranking among the BVAR variants is reversed when compared to the ranking based on RMSE. Again, the predictive density of the large BVAR attaches a too low probability to events that actually occur, and taking

Table 2: Predictive Likelihood

	λ		HICP Inflation		GDP Growth	
	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$
M01: Random Walk	0		-1.87	-1.94	-1.96	-2.32
M02: BVAR Averaging Equal Weights	∞		-1.69	-1.65	-1.58	-1.99
M03: BVAR Averaging Optimal Weights	∞		-1.71	-1.66	-1.55	-2.01
M04: BFAVAR 1 Factor	0.328	0.519	-1.59	-1.61	-1.55	-2.02
M05: BFAVAR 3 Factors	0.105	0.151	-1.53	-1.60	-1.53	-2.06
M06: Large BVAR	0.052	0.051	-1.68	-1.78	-1.67	-2.34
M07: BFAVAR 1 Factor, p and λ optimal	0.356	0.267	-1.62	-1.63	-1.60	-2.08
M08: BFAVAR 3 Factors, p and λ optimal	0.316	0.258	-1.65	-1.64	-1.59	-2.10
M09: Large BVAR, p and λ optimal	0.218	0.171	-1.87	-1.77	-1.72	-2.51
M10: BFAVAR 1 Factor, hyperpriors	0.367	0.272	-1.61	-1.58	-1.52	-1.94
M11: BFAVAR 3 Factors, hyperpriors	0.314	0.271	-1.58	-1.65	-1.53	-2.07
M12: Large BVAR, hyperpriors	0.264	0.203	-1.70	-1.78	-1.60	-2.30
M13: AR(p), p optimal	n/a		-1.80	-1.72	-1.61	-1.98

Notes: this table shows the average log predictive likelihood for different BVAR variants and an AR(p) model. The higher the average log predictive likelihood is the better is the predictive ability of a model. See also notes to Table 1.

the uncertainty regarding the prior choice into account does little to nothing to overcome this problem. In contrast, both BVAR averaging variants, which produced rather inaccurate mean forecasts, now perform rather well and appear to dominate many of the other approaches at $h = 4$. The best predictive ability at both horizons, however, displays the BFAVAR with one factor when a hyperprior is imposed. However, the additional gain of estimating the hyperparameters of the BFAVAR compared to the shrinkage procedure used in Bańbura et al. (2010) appears to be small.

In the following, we investigate whether the predictive likelihoods reported in Table 2 are significantly different from each other. We follow Amisano and Giacomini (2007) and construct for two competing density forecasts j and k an average of likelihood ratios:

$$\mathrm{LR}_{i,h}^{j,k} = \frac{1}{T_1 - T_0 - H + 1} \sum_{T=T_0+H}^{T_1} \left(\log p\left(y_{i,T}|y,j,h\right) - \log p\left(y_{i,T}|y,k,h\right)\right). \tag{17}$$

The test is a conventional likelihood ratio test which is often used for model selection (see, e.g., Giannone et al., 2012, among others) and based on the statistic

$$t = \frac{\mathrm{LR}_{i,h}^{j,k}}{\hat{\sigma}/\sqrt{T_1 - T_0 - H + 1}}, \tag{18}$$

where $\hat{\sigma}$ is an estimate of the standard deviation of $\mathrm{LR}_{i,h}^{j,k}$. The null hypothesis of equal performance of j and k is rejected at the 5 percent level whenever $|t| > 1.96$. In case of rejection, one would choose j (k) whenever $\mathrm{LR}_{i,h}^{j,k}$ is positive (negative). We perform pairwise likelihood ratio tests using our sequence of 84 density forecasts and summarize their results for HICP inflation and real GDP growth in Tables 3 and 4, respectively. The tables read as follows. A $+$ ($-$) means that the model under consideration (column) significantly outperforms (underperforms) its benchmark model (row) in terms of predictive ability, while a $\cdot$ denotes insignificance.

Table 3 reveals that, when forecasting HICP inflation, the random walk as well as the AR model perform poorly, and none of the BVAR variants is significantly outperformed by these benchmark models. Making use of the set of indicator variables is apparently beneficial in this circumstance. For $h = 1$ we find that both BFAVARs, when combined with a shrinkage procedure as proposed by Bańbura et al. (2010), significantly outperform most of their competitors. Optimizing the lag length or the tightness of the prior does not improve forecast ability further. Likewise, there is no additional gain obtained from estimating the hyperparameters as suggested by Giannone et al. (2012). The same conclusion can be drawn for the choice of the aggregation weights when averaging BVAR forecasts since the differences among both BVAR averaging variants are insignificant. Finally, the large BVAR with p and λ optimal as in Carriero et al. (2014) is dominated at the short horizon by nearly all its competitors, including the large BVAR with the shrinkage procedure as in Bańbura et al. (2010). For $h = 4$ similar results obtain when compared to $h = 1$. Overall, the BFAVARs tend to significantly outperform the other approaches whereas the large BVAR variants never dominate the competing variants.

For real GDP growth the picture is different. Table 4 shows that none of the BVAR variants is able to significantly outperform the AR or the random walk. Moreover, the differences among almost all BVAR variants are insignificant. As for HICP inflation, the large BVAR with p and λ optimized appears to deliver a poor forecast for real GDP growth and is significantly worse than the BFAVAR combined with the shrinkage approach of Bańbura et al. (2010). At $h = 4$ real GDP growth is apparently difficult to forecast since the AR is significantly better than most BVAR variants, and we thus conclude that the indicators contain little to no information about real GDP growth four quarters ahead. But we find that both large BVAR variants with optimally chosen hyperparameters are significantly outperformed by all BFAVAR variants, suggesting that neither choosing the lag length and the degree of shrinkage in an optimal fashion nor the estimation of the hyperparameters helps to improve forecast ability of the large BVAR.

In sum, we find that an accurate mean forecast does not necessarily imply a high predictive likelihood and *vice versa*. Model selection based on RMSE only may thus be misleading if the forecaster is also concerned about forecast uncertainty. Overall, the BFAVARs – which combine the advantages of factor analysis and Bayesian shrinkage – deliver a good forecast performance in all setups. Moreover, optimizing the lag length and the shrinkage parameter, or imposing a prior on the hyperparameters does not necessarily help to improve predictive ability. Likewise, we cannot reduce the forecast error by optimizing aggregation weights when averaging BVAR forecasts.

4.5 Probability Integral Transform

While the relative performance of Bayesian forecast models is commonly compared on the basis of predictive likelihoods, in this section we use the probability integral transform (PIT) to assess how good the predictive densities are in absolute terms. In contrast to predictive likelihoods, which are a local measure since they depend only on the predictive density evaluated at the realization, the PIT provides a non-local assessment of the predictive ability of a forecast model (see Geweke and Amisano, 2010). Given that recent work by Rossi and Sehkposyan (2013) suggests that predictive densities generated by BVAR models are often not a reasonable description of actual uncertainty, we investigate which specification choices are helpful to achieve correct calibration. Moreover, we may obtain some insights about the reasons behind the unsatisfying predictive ability of the large BVARs in terms of density forecasting.

We first decompose each predictive density into intervals, each covering 10 percent of the probability mass. We then assign each realization to the corresponding interval. If a predictive density would be a reasonable description of the the true distribution of the data, each probability band should contain 10 percent of the realizations (about 8 in our case). This procedure has been developed by Diebold, Gunther, and Tay (1998) and can be formalized as follows.[11] For a proposed predictive density p the PIT is the corresponding cumulative density function (CDF) evaluated at the realized value $y_{i,T}$:

$$z_{i,T}^{h} = \int_{-\infty}^{y_{i,T}} p\left(u|y,h\right) du = P\left(y_{i,T}|y,h\right), \quad \text{for } T = T_0 + H, ..., T_1. \tag{19}$$

If the proposed predictive density is consistent with the true predictive density, the density of $\left\{z_{i,T}^{h}\right\}_{T=T_0+H}^{T_1}$ is an independent and identically distributed U $(0,1)$ and its CDF is the $45°$ line.

[11]See also Diebold, Hahn, and Tay (1999).

Table 3: Likelihood Ratio Tests - HICP Inflation

Benchmark	h	M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12
M02		$\cdot$											
M03		$\cdot$	$\cdot$										
M04		$-$	$\cdot$	$-$									
M05		$-$	$-$	$-$	$\cdot$								
M06		$-$	$\cdot$	$\cdot$	$\cdot$	$+$							
M07	1	$-$	$\cdot$	$\cdot$	$\cdot$	$+$	$\cdot$						
M08		$\cdot$	$\cdot$	$\cdot$	$\cdot$	$+$	$\cdot$	$\cdot$					
M09		$\cdot$	$+$	$\cdot$	$+$	$+$	$+$	$+$	$+$				
M10		$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$-$			
M11		$-$	$\cdot$	$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$-$	$-$	$\cdot$		
M12		$\cdot$	$\cdot$	$\cdot$	$+$	$+$	$\cdot$	$\cdot$	$\cdot$	$-$	$\cdot$	$+$	
M13		$\cdot$	$\cdot$	$\cdot$	$+$	$+$	$\cdot$	$+$	$\cdot$	$\cdot$	$+$	$+$	$\cdot$
M02		$-$											
M03		$-$	$\cdot$										
M04		$-$	$\cdot$	$\cdot$									
M05		$-$	$\cdot$	$\cdot$	$\cdot$								
M06		$-$	$\cdot$	$\cdot$	$+$	$+$							
M07	4	$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$-$						
M08		$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$-$	$\cdot$					
M09		$-$	$\cdot$	$\cdot$	$+$	$+$	$\cdot$	$+$	$+$				
M10		$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$-$	$\cdot$	$\cdot$	$-$			
M11		$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$-$	$\cdot$	$\cdot$	$-$	$\cdot$		
M12		$-$	$\cdot$	$\cdot$	$+$	$+$	$\cdot$	$+$	$+$	$\cdot$	$+$	$+$	
M13		$-$	$\cdot$	$\cdot$	$+$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$+$	$\cdot$	$\cdot$

Notes: this table shows the results of pairwise likelihood ratio tests. A $+$ $(-)$ means that the model under consideration (column) significantly outperforms (underperforms) its benchmark (row) in terms of predictive ability. A $\cdot$ denotes insignificance. The significance level is 5 percent. The models are: M01 = Random Walk; M02 = BVAR Averaging Equal Weights; M03 = BVAR Averaging Optimal Weights; M04 = BFAVAR 1 Factor; M05 = BFAVAR 3 Factors; M06 = Large BVAR; M07 = BFAVAR 1 Factor, p and λ optimal; M08 = BFAVAR 3 Factors, p and λ optimal; M09 = Large BVAR, p and λ optimal; M10 = BFAVAR 1 Factor, hyperpriors; M11 = BFAVAR 3 Factors, hyperpriors; M12 = Large BVAR, hyperpriors; M13 = AR(p), p optimal. See also notes to Table 1.

Table 4: Likelihood Ratio Tests - GDP Growth

Benchmark	h	M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12
M02		·											
M03		·	·										
M04		·	·	·									
M05		·	·	·	·								
M06		·	·	·	·	+							
M07	1	·	·	·	·	·	·						
M08		·	·	·	·	·	·	·					
M09		·	·	·	+	+	·	·	·				
M10		·	·	·	·	·	·	·	·	−			
M11		·	·	·	·	·	·	·	·	−	·		
M12		·	·	·	·	·	·	·	·	·	·	·	
M13		·	·	·	·	·	·	·	·	·	·	·	·
M02		·											
M03		·	·										
M04		·	·	·									
M05		·	·	·	·								
M06		·	·	·	·	·							
M07	4	·	·	·	·	·	−						
M08		·	·	·	·	·	·	·					
M09		+	·	·	·	+	+	+	+				
M10		·	·	·	·	·	·	·	−	−			
M11		·	·	·	·	·	·	·	·	−	·		
M12		·	·	·	·	+	·	+	+	·	+	+	
M13		−	·	·	·	−	−	−	−	−	·	·	−

Notes: see notes to Tables 1 and 3.

We provide a graphical assessment of the PITs for HICP inflation and real GDP growth in Figures 1 and 2, respectively. Following Rossi and Sehkposyan (2013) we also use an Anderson-Darling (AD) and a Kolmogorov-Smirnov (KS) test to formally check for the uniformity of the distribution of the PITs. Both tests build on the absolute difference between the CDF of a PIT and the CDF of a uniform distribution, i.e. the $45°$ line. We document their outcome in Tables D.3 and D.4 in Appendix D.

In case of HICP inflation we obtain support for the conclusions drawn in the previous subsections. The performance of the random walk variant is again rather poor, and Figure 1 shows that too little weight is given to the lower tail of the distribution. In contrast, most BVAR variants as well as the AR model produce predictive densities that appear to be a reasonable description of the true distribution of the data. However, we reject uniformity of the PITs for the large BVAR with shrinkage as proposed by Bańbura et al. (2010). Such a result is in line with the unsatisfying predictive ability of this variant. The plot of the PITs suggests that the large BVAR attaches a too low probability to very low or negative inflation rates, and, at the same time, too much weight is given to the upper tails of the distribution of inflation. However, we cannot reject uniformity for the large BVAR at a 5 percent level when we take the uncertainty surrounding the choice for the hyperparameters into account. While using hyperpriors apparently helps to generate a more realistic density forecast, our results in Section 4 show that this is not sufficient to outperform the other variants. The BFAVAR with 1 factor combined with shrinkage as in Bańbura et al. (2010) appears to be well calibrated at both horizons, whereas optimizing the lag length and the degree of shrinkage, or imposing a hyperprior, leads to a rejection of correct calibration at $h = 4$. Moreover, we again find that all BFAVAR variants with 3 factors have a good predictive ability for HICP inflation at both horizons.

Given the evidence of the previous subsections, it is not surprising that our findings are different for real GDP growth. With only few exceptions, uniformity is rejected by both tests at the 5 percent level. A notable exception is the BFAVAR with 3 factors ($h = 1$) where the optimal choice of the hyperparameters helps to achieve a realistic density forecast. Similarly, the large BVAR appears to benefit slightly from optimally chosen hyperparameters. A visual assessment of the plots in Figure 2 suggests that most BVAR variants apparently generate too little mass in the lower tail of the distribution since too many realizations appear to fall into the bins covering the first 10-20% of the probability mass. That is, they underestimate the occurrence of low or even negative real GDP growth rates.

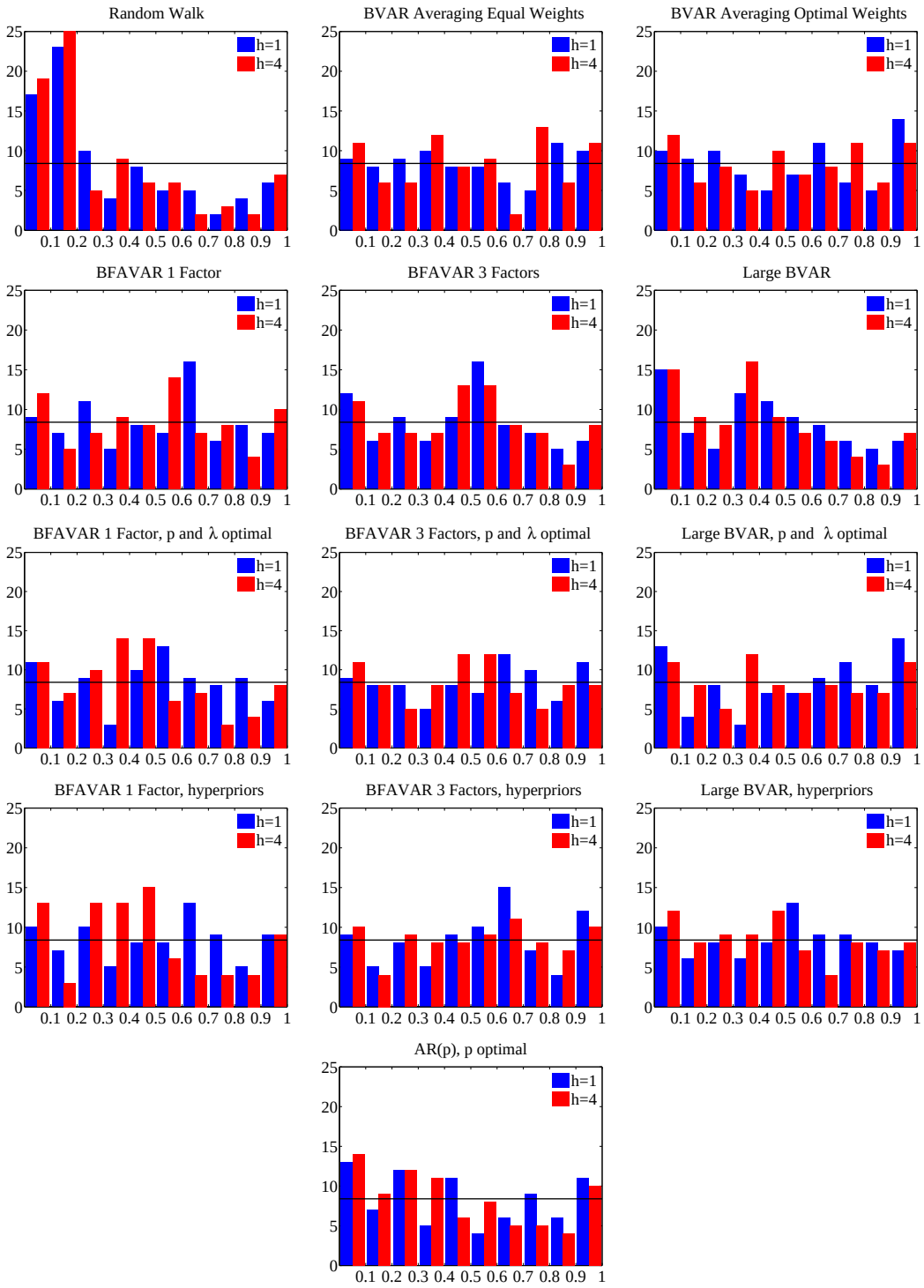

Figure 1: Probability Integral Transform - HICP Inflation. Notes: this figure shows the PITs for different BVAR variants and an AR model to check for the correct calibration of the predictive density. The horizontal line denotes uniformity. x-axis: probability band; y-axis: number of realizations.

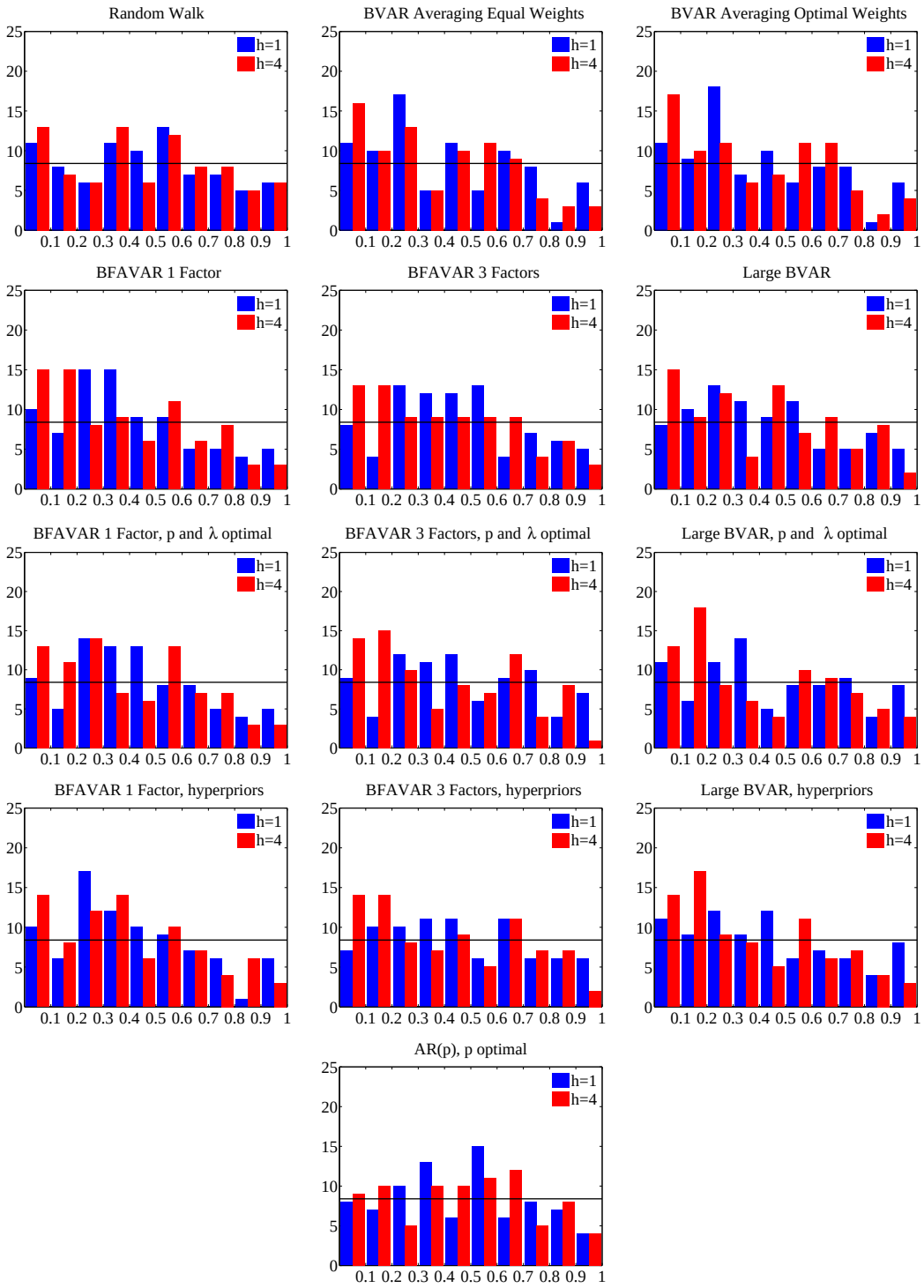

Figure 2: Probability Integral Transform - GDP Growth. Notes: this figure shows the PITs for different BVAR variants and an AR model to check for the correct calibration of the predictive density. The horizontal line denotes uniformity. x-axis: probability band; y-axis: number of realizations.

5 Robustness

In the previous section we have shown how the ranking of the BVAR variants changes when the focus is on different target variables, different forecast horizons, and alternative loss functions. Moreover, we explored which specification choices may help to achieve correct calibration. In this section we investigate the sensitivity of our findings with respect to the following features of our forecast experiment. First, we consider the full dataset. Second, we employ a recursive (or expanding) rather than a rolling estimation scheme. Third, we explore the subsample stability of our results. And finally, we let the variables enter the models in levels rather than first differences. The corresponding tables are all provided in Appendix D.

Full Dataset So far we have used a subset of 22 variables of our dataset, thereby neglecting indicators that are potentially useful for forecasting. In the first robustness check we extend the information set and repeat the forecast experiment with the full cross-section consisting of 44 time series.

While the accuracy of the point forecasts hardly changes across variants when the size of the cross-section increases, each variant appears to be affected differently when we consider the density forecasts. For instance, we find that the predictive ability of the BFAVARs and the BVAR averaging improves. The forecast performance of the BFAVARs improves since a larger cross-section potentially helps to better identify the common factors while the size of the system to be estimated does not increase. The BVAR averaging probably benefits from the fact that more (density) forecasts enter the pooled prediction, which makes it less vulnerable to outliers. By contrast, we find that the performance of the large BVARs worsens further when we enlarge the set of indicator variables. The existing literature on large BVARs is clear about the fact that models with about 20-25 variables are in general superior to larger systems. Likewise, our results suggest that the large BVAR variant cannot benefit from a larger information set.

Moreover, it appears that the additional gains of the larger information set are more pronounced for HICP inflation than for real GDP growth, which is not surprising since we have already learned that the information variables seem to be more relevant for the former than for the latter. Overall, using a larger information set thus underpins our baseline results.

Recursive Estimation Scheme Our baseline results are obtained from a rolling estimation scheme. Such a proceeding is advisable when the data generating process is subject to structural breaks since the estimated parameters are allowed to change quickly over time. When the data generating process is rather stable over time, however, discarding distant observations may introduce unnecessary high estimation uncertainty. In the second robustness check we thus use

a recursive (or expanding) window to estimate the models.

It turns out that the ranking of the variants based on the RMSE is not affected by such a forecast setup. Turning to the predictive likelihood, we find that all approaches benefit substantially when we use an expanding window. Furthermore, we obtain some evidence that the predictive ability of the large BVARs improves more than that of the other variants when more observations enter the estimation. The large BVAR with shrinkage as in Bańbura et al. (2010) now shows the highest predictive likelihood for HICP inflation at both horizons. The improvement in predictive ability is, however, restricted to this large BVAR variant. The variant with λ and p optimized and that with hyperpriors still underperform most of their competitors. When real GDP growth is considered, the ranking among the BVAR variants is hardly affected. The large BVAR now tends to be slightly better than the BVAR averaging for $h = 1$, while the BFAVARs with three factors still have the highest predictive ability. For $h = 4$ the predictive ability of the large BVARs is still unsatisfactory.

Subsample Stability In the third robustness check we shorten the evaluation period and concentrate on the sample 1999:1 to 2011:4. We choose that period to evaluate the subsample stability of our results for the following reasons. First, this sample excludes the period of inflation convergence which may have changed the properties of the inflation process. Second, we begin with the creation of the euro area in January 1999 which mimics the situation encountered by researchers forecasting the euro area at that time. And third, this sample represents the period in which a common monetary policy was in effect. Note that shortening the evaluation period comes at the cost of less precise estimates for RMSE and the predictive likelihood.

All in all, we find that our results are robust to the choice of the evaluation period. The relative performance of the BVAR variants hardly changes when the evaluation period is shortened. The only exception is the random walk variant that now shows a better performance in terms of RMSE than before when we consider HICP inflation. Given that the predictive mean of the random walk adapts only slowly over time, it does not come as a surprise that excluding the period of inflation convergence reduces the forecast error of that benchmark model.

Level Specification In the last robustness check the variables enter the models in levels rather than in first differences. We remain within the same estimation framework as before with the exception that we impose the random walk prior for all variables. A closed form solution for the marginal posterior of the predictive density for HICP inflation and real GDP growth is, however, not available in this case. For that reason we evaluate the levels directly.

Overall, the ranking remains relatively stable when we consider the variables in levels. For HICP inflation and $h = 1$, our results suggest that the differences among the BVAR variants

evaluated by the RMSE remain small and the relative performance of the BVAR variants is not affected by this specification choice. The BVAR averaging seems to benefit from a specification in levels at $h = 4$. For real GDP growth, the differences in RMSE remain small as well. One exception should be noted. While the large BVAR with hyperpriors as in Giannone et al. (2012) remains among the best performing variants, the large BVARs with shrinkage as in Bańbura et al. (2010) or Carriero et al. (2014) appear to perform worse when we consider the variables in levels. Turning to the density forecasts, we still find that all large BVAR variants are outperformed by the other approaches while the BFAVARs deliver the best forecasts. Choosing p and λ optimal or using hyperpriors does not, in general, help to improve the forecast performance. In fact, the large BVAR with p and λ optimal now shows the lowest predictive ability among the different BVAR variants for all variables and horizons.

6 Conclusion

Recent articles suggest that VARs with Bayesian shrinkage produce accurate point and density forecasts even when the number of variables becomes large. However, there are numerous ways to implement the BVAR approach and these specification choices may affect predictive ability. In this paper we evaluate different variants of the BVAR within the same framework. In particular, we consider BVAR averaging with equal and optimal weights, BFAVARs with 1 and 3 factors, as well as large BVARs. Moreover, we use the three predominant approaches of the related literature to determine the hyperparameters of the models, in particular the tightness of the prior. We thus evaluate all major specification choices discussed in the prevalent literature. To be able to investigate how efficiently these BVAR variants process information contained in a large dataset, we condition our analysis on a given amount of information. The variants are then evaluated according to their out-of-sample forecast performance.

We first evaluate the predictive mean and find that all variants outperform the random walk variant for HICP inflation at $h = 1$ and $h = 4$, while the differences in forecast accuracy between these variants are small and often insignificant. For real GDP growth the large BVAR variants deliver the best forecast at $h = 1$ but cannot improve on the random walk at $h = 4$. Moreover, we find that neither setting the shrinkage hyperparameter optimally nor estimating it in a hierarchical fashion helps to improve the forecast accuracy compared to a simple shrinkage procedure of Bańbura et al. (2010). Likewise, choosing the aggregation weights for the BVAR averaging in an optimal way does not, in general, improve forecast accuracy compared to a naïve equal weighting scheme. Finally, we find that both BFAVARs perform satisfactory across target variables and forecast horizons.

When we evaluate the entire predictive density, we obtain evidence that an accurate mean forecast does not necessarily imply a high predictive likelihood and *vice versa*. In particular, it turns out that the performance of the large BVARs deteriorates when the predictive density is considered, while that of the BVAR averaging improves. And we again find that choosing the degree of shrinkage optimally or estimating it in a hierarchical fashion does not, in general, lead to better predictive ability. Moreover, we use calibration tests in order to assess how good the predictive densities are in absolute terms rather than competing against each other. While all BVARs except for the large BVAR variant produce predictive densities that appear to be a reasonable description of the true distribution for HICP inflation, the calibration tests detect a poor calibration of all variants with respect to real GDP growth. It appears that the large BVAR assigns too little weight to very low inflation rates, while it gives too much weight to the upper tails of the distribution.

All in all, our findings are robust to changes in the forecast setup. However, robustness checks also suggest that especially the performance of the large BVARs seems to depend on the features of the forecast experiment such as the size of the cross-section, the number of observations used for estimation or whether we use variables in levels or differences. While most BVAR variants appear to benefit from a larger information set, the large BVARs apparently lose forecast ability when we increase the size of the dataset. It appears that there is an upper limit to the number of predictors where the benefits from the large BVAR are retained. Such a result suggests that a preselection device could help to determine the (subset of) variables that should enter the information set of the large BVAR.

We have shown how the ranking of the different variants depends on the forecast situation at hand. Given the evidence of this paper, we believe that there is no consistent outperformance of a particular BVAR variant. Particularly, the loss function of the forecaster seems to be of great importance for the ranking. Overall, we find that a factor augmented BVAR with three factors is competitive in all setups. Hence, it appears to be advisable to combine the factor approach which helps to eliminate noisy signals from the different indicator variables with Bayesian shrinkage to further reduce estimation uncertainty. For future research it might also be fruitful to assess the performance of more flexible models. For instance, the introduction of stochastic volatility to explicitly account for changes in the shock process may help to achieve correct calibration. Moreover, modeling time varying parameters may improve forecast accuracy in an unstable environment. The work of Clark (2011), Barnett et al. (2012), and Carriero, Clark, and Marcellino (2012, 2013) goes in this direction.

References

Aastveit, K. A., K. R. Gerdrup, A. S. Jore, and L. A. Thorsrud (2011). Nowcasting GDP in Real-Time: A Density Combination Approach. Working Paper 2011/11, Norges Bank.

Aiolfi, M., C. Capistrán, and A. Timmermann (2011). Forecast Combinations. In M. P. Clements and D. F. Hendry (Eds.), *Oxford Handbook of Economic Forecasting*. Oxford: Oxford University Press.

Amisano, G. and R. Giacomini (2007). Comparing Density Forecasts via Weighted Likelihood Ratio Tests. *Journal of Business & Economic Statistics 25*(2), 177–190.

Bai, J. and S. Ng (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica 70*(1), 191–221.

Bańbura, M., D. Giannone, and L. Reichlin (2010). Large Bayesian Vector Autoregressions. *Journal of Applied Econometrics 25*(1), 71–92.

Barhoumi, K., O. Darné, and L. Ferrara (2013). Testing the Number of Factors: An Empirical Assessment for a Forecasting Purpose. *Oxford Bulletin of Economics and Statistics 75*(1), 64–79.

Barnett, A., H. Mumtaz, and K. Theodoridis (2012). Forecasting UK GDP Growth, Inflation and Interest Rates under Structural Change: A Comparison of Models with Time-Varying Parameters. Bank of England working papers 450, Bank of England.

Berg, T. O. (2014). Technology News and the U.S. Economy: Time Variation and Structural Changes. forthcoming in *Scottish Journal of Political Economy*.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005). Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach. *Quarterly Journal of Economics 120*(1), 387–422.

Boero, G., J. Smith, and K. F. Wallis (2011). Scoring Rules and Survey Density Forecasts. *International Journal of Forecasting 27*(2), 379–393.

Boivin, J. and S. Ng (2006). Are More Data Always Better for Factor Analysis? *Journal of Econometrics 132*(1), 169–194.

Carriero, A., T. E. Clark, and M. Marcellino (2012). Common Drifting Volatility in Large Bayesian VARs. CEPR Discussion Papers 8894, C.E.P.R. Discussion Papers.

Carriero, A., T. E. Clark, and M. Marcellino (2013). Real-Time Nowcasting with a Bayesian Mixed Frequency Model with Stochastic Volatility. CEPR Discussion Papers 9312, C.E.P.R. Discussion Papers.

Carriero, A., T. E. Clark, and M. Marcellino (2014). Bayesian VARs: Specification Choices and Forecast Accuracy. forthcoming in *Journal of Applied Econometrics*.

Carriero, A., G. Kapetanios, and M. Marcellino (2009). Forecasting Exchange Rates with a Large Bayesian VAR. *International Journal of Forecasting 25*(2), 400–417.

Clark, T. E. (2011). Real-Time Density Forecasts From Bayesian Vector Autoregressions With Stochastic Volatility. *Journal of Business and Economic Statistics 29*(3), 327–341.

Clark, T. E. and M. W. McCracken (2010). Averaging Forecasts from VARs with Uncertain Instabilities. *Journal of Applied Econometrics 25*(1), 5–29.

D'Agostino, A., L. Gambetti, and D. Giannone (2013). Macroeconomic Forecasting and Structural Change. *Journal of Applied Econometrics 28*(1), 82–101.

De Mol, C., D. Giannone, and L. Reichlin (2008). Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components? *Journal of Econometrics 146*(2), 318–328.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998). Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review 39*(4), 863–883.

Diebold, F. X., J. Hahn, and A. S. Tay (1999). Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High-Frequency Returns on Foreign Exchange. *Review of Economics and Statistics 81*(4), 661–673.

Doan, T., R. Litterman, and C. Sims (1984). Forecasting and Conditional Projection Using Realistic Prior Distributions. *Econometric Reviews 3*(1), 1–100.

Fagan, G., J. Henry, and R. Mestre (2005). An Area-Wide Model for the Euro Area. *Economic Modelling 22*(1), 39–59.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2003). Do Financial Variables Help Forecasting Inflation and Real Activity in the Euro Area? *Journal of Monetary Economics 50*(6), 1243–1255.

Geweke, J. (1992). Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics*, pp. 169–194. Oxford University Press.

Geweke, J. and G. Amisano (2010). Comparing and Evaluating Bayesian Predictive Distributions of Asset Returns. *International Journal of Forecasting 26*(2), 216–230.

Geweke, J. and G. Amisano (2011). Optimal Prediction Pools. *Journal of Econometrics 164*(1), 130–141.

Geweke, J. and G. Amisano (2012). Prediction with Misspecified Models. *American Economic Review: Papers & Proceedings 102*(3), 482–486.

Giacomini, R. and H. White (2006). Tests of Conditional Predicitve Ability. *Econometrica 74*(6), 1545–1578.

Giannone, D., M. Lenza, D. Momferatou, and L. Onorante (2010). Short-Term Inflation Projections: A Bayesian Vector Autoregressive Approach. CEPR Discussion Papers 7746, C.E.P.R. Discussion Papers.

Giannone, D., M. Lenza, and G. E. Primiceri (2012). Prior Selection for Vector Autoregressions. NBER Working Papers 18467, National Bureau of Economic Research, Inc.

Hall, S. G. and J. Mitchell (2007). Combining Density Forecasts. *International Journal of Forecasting 23*(1), 1–13.

Henzel, S. R. and J. Mayr (2013). The Mechanics of VAR Forecast Pooling - A DSGE Model based Monte Carlo Study. *North American Journal of Economics and Finance 24*(1), 1–24.

Kadiyala, K. R. and S. Karlsson (1993). Forecasting with Generalized Bayesian Vector Autoregressions. *Journal of Forecasting 12*(4), 365–378.

Kadiyala, K. R. and S. Karlsson (1997). Numerical Methods for Estimation and Inference in Bayesian VAR-Models. *Journal of Applied Econometrics 12*(2), 99–132.

Knüppel, M. and G. Schultefrankenfeld (2012). How Informative Are Central Bank Assessments of Macroeconomic Risks? *International Journal of Central Banking 8*(3), 87–139.

Koop, G. (2013). Forecasting with Medium and Large Bayesian VARS. *Journal of Applied Econometrics 28*(2), 177–203.

Litterman, R. B. (1980). A Bayesian Procedure for Forecasting with Vector Autoregressions. Manuscript, Massachusetts Institute of Technology, Department of Economics.

Litterman, R. B. (1986). Forecasting With Bayesian Vector Autoregressions - Five Years of Experience. *Journal of Business & Economic Statistics 4*(1), 25–38.

Mitchell, J. and S. G. Hall (2005). Evaluating, Comparing and Combining Density Forecasts Using the KLIC with an Application to the Bank of England and NIESR 'Fan' Charts of Inflation. *Oxford Bulletin of Economics and Statistics 67*(s1), 995–1033.

Pirschel, I. and M. H. Wolters (2014). Forecasting German Key Macroeconomic Variables Using Large Dataset Methods. Manuscript, University of Kiel.

Primiceri, G. E. (2005). Time Varying Structural Vector Autoregressions and Monetary Policy. *Review of Economic Studies 72*(3), 821–852.

Robertson, J. C. and E. W. Tallman (1999). Vector Autoregressions: Forecasting and Reality. *Economic Review 84*(1), 4–18.

Rossi, B. and T. Sehkposyan (2013). Evaluating Predictive Densities of U.S. Output Growth and Inflation in a Large Macroeconomic Data Set. Working Papers 689, Barcelona Graduate School of Economics.

Stock, J. H. and M. W. Watson (2002). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics 20*(2), 147–161.

Stock, J. H. and M. W. Watson (2005). Implications of Dynamic Factor Models for VAR Analysis. NBER Working Papers 11467, National Bureau of Economic Research, Inc.

Stock, J. H. and M. W. Watson (2006). Forecasting Using Many Predictors. In G. Elliott, C. Granger, and A. Timmerman (Eds.), *Handbook of Economic Forecasting*. Amsterdam: North Holland.

Stock, J. H. and M. W. Watson (2011). Dynamic Factor Models. In M. P. Clements and D. F. Hendry (Eds.), *Oxford Handbook of Forecasting*. Oxford: Oxford University Press.

Weiss, A. A. (1996). Estimating Time Series Models Using the Relevant Cost Function. *Journal of Applied Econometrics 11*(5), 539–560.

Wolters, M. H. (2014). Evaluating Point and Density Forecasts of DSGE Models. forthcoming in *Journal of Applied Economtrics*.

A Description of the Dataset

The dataset comprises 44 quarterly euro area macroeconomic and financial time series covering the period 1975:1 to 2011:4. The series are described in detail below. The format is as follows: series number, series mnemonic used by the original data source, series label, series category, the original data source (AWM = Area-wide Model database; OECD = Organisation for Economic Co-operation and Development), and series transformation (1 = Log-Difference; 2 = Difference). The last column indicates if a series is included in the benchmark specification.

No.	Mnemonic	Label	Category	Source	Transform.	Benchmark
01	YER	Real GDP	National Accounts	AWM	1	x
02	PCR	Real Private Consumption	National Accounts	AWM	1	x
03	GCR	Real Government Consumption	National Accounts	AWM	1	
04	ITR	Real Gross Investment	National Accounts	AWM	1	x
05	XTR	Real Exports of Goods and Services	National Accounts	AWM	1	
06	MTR	Real Imports of Goods and Services	National Accounts	AWM	1	
07	YFN	GDP at Factor Costs	National Accounts	AWM	1	
08	WIN	Compensation to Employees	National Accounts	AWM	1	
09	GON	Gross Operating Surplus	National Accounts	AWM	1	
10	TIN	Indirect Taxes (net of subsidies)	National Accounts	AWM	1	
11	YIN	GDP Income Side	National Accounts	AWM	1	
12	NFNYEN	Net Factor Income from Abroad/GDP	National Accounts	AWM	2	
13	SAX	Household's Savings Ratio	National Accounts	AWM	2	
14	HICP	Overall HICP	Price Indexes	AWM	1	x
15	YED	GDP Deflator	Price Indexes	AWM	1	x
16	PCD	Private Consumption Deflator	Price Indexes	AWM	1	x
17	GCD	Government Consumption Deflator	Price Indexes	AWM	1	
18	ITD	Gross Investment Deflator	Price Indexes	AWM	1	x
19	XTD	Exports of Goods and Services Deflator	Price Indexes	AWM	1	
20	MTD	Imports of Goods and Services Deflator	Price Indexes	AWM	1	
21	YFD	GDP at Factor Costs Deflator	Price Indexes	AWM	1	
22	YWR	Real World GDP	International	AWM	1	x
23	YWRX	Real World Demand	International	AWM	1	
24	YWD	World GDP Deflator	International	AWM	1	x
25	COMPR	Commodity Prices	International	AWM	1	
26	PCOMU	Non-Oil Commodity Prices	International	AWM	1	x
27	POILU	Oil Prices	International	AWM	1	x
28	LFN	Labor Force	Employment	AWM	1	
29	LNN	Total Employment	Employment	AWM	1	x
30	LEN	Employees	Employment	AWM	1	
31	UNN	Number of Unemployed	Employment	AWM	1	
32	URX	Unemployment Rate	Employment	AWM	2	x
33	LPROD	Labor Productivity	Employment	AWM	1	
34	ULC	Unit Labor Costs	Employment	AWM	1	
35	WRN	Wages	Employment	AWM	1	
36	EKOL2002Q	Composite Leading Indicator	Surveys	OECD	2	x
37	EKOCS002Q	Consumer Confidence Indicator	Surveys	OECD	2	x
38	EKQMA027B	M1 Money Stock	Monetary Aggregates	OECD	1	x
39	EKQMA013B	M3 Money Stock	Monetary Aggregates	OECD	1	x
40	STN	Short-Term Interest Rate	Financial	AWM	2	x
41	LTN	Long-Term Interest Rate	Financial	AWM	2	x
42	EMSHRPRCF	Share Price Index	Financial	OECD	1	x
43	EEN	Nominal Effective Exchange Rate	Financial	AWM	1	x
44	EXR	Euro per U.S.D. Exchange Rate	Financial	AWM	1	x

B Marginal Likelihood

According to Carriero et al. (2014) the analytical expression for the marginal likelihood is:

$$p\left(y\right)=k^{-1}\times\left|\Psi+\left(y-xb\right)'\left(I+x\Omega x'\right)^{-1}\left(y-xb\right)\right|^{-\frac{T+d}{2}}, \tag{B.1}$$

where

$$k=\pi^{\frac{TN}{2}}\times\left|\left(I+x\Omega x'\right)^{-1}\right|^{-\frac{N}{2}}\times\left|\Psi\right|^{-\frac{d}{2}}\times\frac{\Gamma_N\left(\frac{d}{2}\right)}{\Gamma_N\left(\frac{T+d}{2}\right)}, \tag{B.2}$$

and with $\Gamma_N\left(\cdot\right)$ denoting the N-variate gamma function. See Giannone et al. (2012) for a similar expression.

C Markov Chain Monte Carlo Estimation

This appendix outlines the Metropolis-Hastings algorithm we use to simulate the joint posterior of the predictive density and the hyperparameters γ. Moreover, we assess the convergence of the Markov chain with inefficiency factors.

C.1 Metropolis-Hastings Algorithm

The steps are:

Step 1: Choose a starting point γ^0. We use the posterior mode, which is obtained by numerical optimization.[12] Then run a loop over the following steps.

Step 2: Draw a proposal γ^* from a jumping distribution $\mathrm{J}\left(\gamma^*|\gamma^{j-1}\right)=\mathrm{N}\left(\gamma^{j-1},c\cdot\Sigma_m\right)$, where Σ_m is the inverse of the Hessian computed at the posterior mode, and c is a scaling constant chosen to obtain an acceptance ratio of about 20 percent.

Step 3: Compute the acceptance ratio:

$$r=\frac{p\left(\gamma^*|y\right)}{p\left(\gamma^{j-1}|y\right)} \tag{C.1}$$

Step 4: Randomly draw ν from $\mathrm{U}\left(0,1\right)$.

Step 5: Accept or discard the proposal γ^* according to the following rule, and update, if necessary, the jumping distribution:

[12] We use the Matlab routine "fmincon".

$$\gamma^j = \begin{cases} \gamma^* & : \text{ if } \quad \nu \leq r \\ \gamma^{j-1} & : \text{ otherwise} \end{cases} \tag{C.2}$$

In total, we perform 20,000 iterations of the sampler, discarding the first 10,000 for convergence. Conditional on the retained draws, the predictive density can be obtained from its marginal posterior. Below we present inefficiency factors, showing that the chain has converged.

C.2 Convergence

In order to check for the convergence of the Markov chain, we follow Primiceri (2005) and calculate inefficiency factors (IFs) for all the hyperparameters of each BVAR variant (in total, we have 420 for the BFAVAR with 1 factor, 588 for the BFAVAR with 3 factors, and 2,016 for the large BVAR, respectively). The IF is defined as $1 + 2\sum_{s=1}^{\infty} \rho_s$, where ρ_s is the estimated autocorrelation of the chain at lag s. Since independence sampling produces an IF that is equal to one and dependence sampling typically produces an IF greater than one, the IF quantifies the relative efficiency loss in the computation of posterior draws from dependent versus independent samples. In practice, values around 20 are regarded as efficient (see, e.g., Primiceri, 2005, among others), meaning that the econometrician needs to draw 20 times as many draws as from uncorrelated samples.

Following Berg (2014), we calculate the IFs as the inverse of the relative numerical efficiency measure (RNE) of Geweke (1992):

$$\text{RNE} = (2\pi)^{-1} \frac{1}{\text{S}(0)} \int_{-\pi}^{\pi} \text{S}(\omega)\, d\omega, \tag{C.3}$$

where $\text{S}(\omega)$ denotes the spectral density of the sequence of draws at frequency ω. We estimate the spectral densities by smoothing the periodograms in the frequency domain using a 4 percent tapered window as in Primiceri (2005).

In Figure C.1 we document the IFs for the three BVAR variants at both forecast horizons. The numbers all range between 5 and 25, providing some evidence that the posterior draws for the hyperparameters come from a converged distribution. Even the large BVAR with more than 2,000 hyperparameters shows IFs that are only slightly higher than those for the BFAVARs. All in all, we regard these numbers as satisfactory.

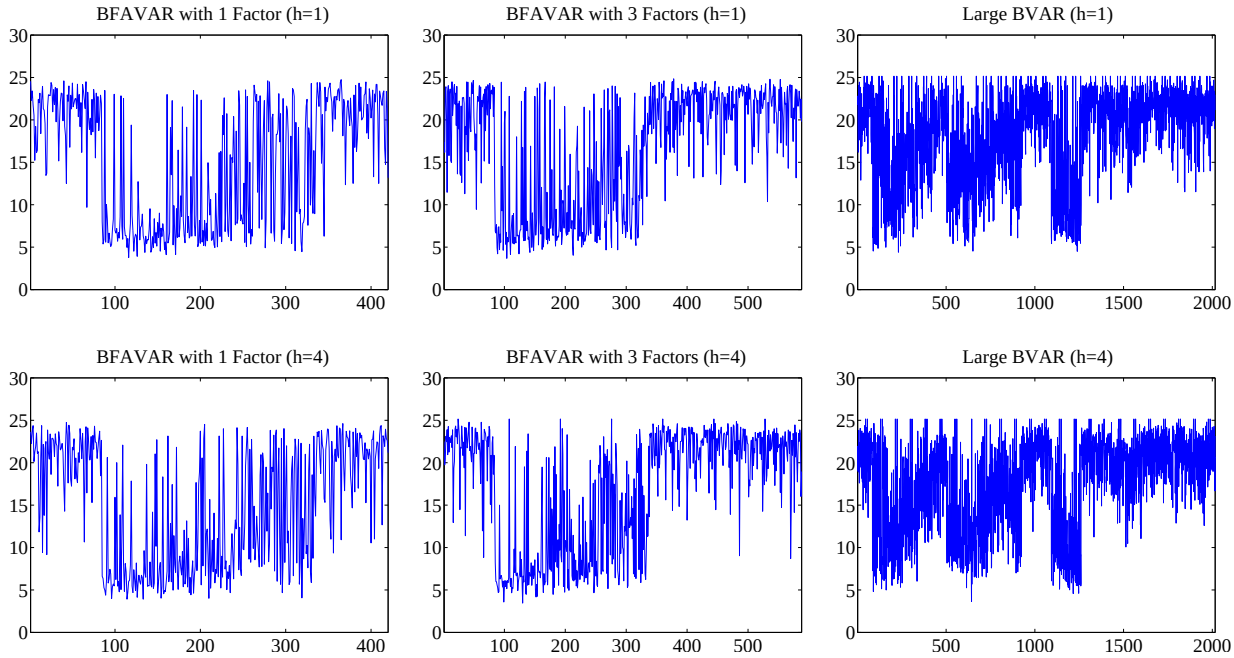

Figure C.1: Inefficiency Factors. Notes: this figure shows the IFs for all the hyperparameters of the BVARs to check for the convergence of the Markov chain. x-axis: hyperparameter; y-axis: IF.

D Additional Tables

In this section we provide additional material for our paper to which we refer to in the main text. In particular, we document the results of a Giacomini and White (2006) test of equal conditional predictive ability of the point forecast for HICP inflation and real GDP growth. Moreoever, we show the outcome of an Anderson-Darling (AD) as well as a Kolmogorov-Smirnov (KS) test of uniformity of the probability integral transform (PIT) for the different BVAR variants and the AR model. And finally, we provide the corresponding tables to our robustness checks. We report the root mean sqared error (RMSE) and the predictive likelihood for the BVAR variants and the AR model when (a) the full sample including all 44 variables is considered; (b) a recursive rather than a rolling estimation scheme is used; (c) the evaluation period is shortend to 1999:1 to 2011:4; (d) the variables enter the models in levels instead of first differences. For further details we refer to the main text.

Table D.1: Test of Equal Predictive Ability - HICP Inflation

Benchmark	h	M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12
M02		$-$											
M03		$-$	$\cdot$										
M04		$-$	$\cdot$	$\cdot$									
M05		$-$	$\cdot$	$\cdot$	$\cdot$								
M06		$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$							
M07	1	$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$						
M08		$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$					
M09		$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$				
M10		$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$			
M11		$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$		
M12		$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	
M13		$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
M02		$-$											
M03		$-$	$\cdot$										
M04		$-$	$\cdot$	$\cdot$									
M05		$-$	$\cdot$	$\cdot$	$\cdot$								
M06		$-$	$\cdot$	$\cdot$	$\cdot$	$+$							
M07	4	$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$						
M08		$-$	$\cdot$	$\cdot$	$\cdot$	$+$	$\cdot$	$-$					
M09		$-$	$\cdot$	$\cdot$	$-$	$+$	$\cdot$	$-$	$\cdot$				
M10		$-$	$\cdot$	$\cdot$	$+$	$\cdot$	$\cdot$	$\cdot$	$+$	$\cdot$			
M11		$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$-$	$+$	$\cdot$	$\cdot$	$\cdot$		
M12		$-$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$-$	$\cdot$	
M13		$-$	$\cdot$	$\cdot$	$+$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$+$	$\cdot$	$+$	$\cdot$

Notes: this table shows the results of a Giacomini and White (2006) test of equal conditional predictive ability. The test is based on squared forecast error loss. A $+$ $(-)$ means that the model under consideration (column) significantly (at a 5 percent level) outperforms (underperforms) its benchmark (row) in terms of RMSE. A $\cdot$ denotes insignificance. The models are: M01 = Random Walk; M02 = BVAR Averaging Equal Weights; M03 = BVAR Averaging Optimal Weights; M04 = BFAVAR 1 Factor; M05 = BFAVAR 3 Factors; M06 = Large BVAR; M07 = BFAVAR 1 Factor, p and λ optimal; M08 = BFAVAR 3 Factors, p and λ optimal; M09 = Large BVAR, p and λ optimal; M10 = BFAVAR 1 Factor, hyperpriors; M11 = BFAVAR 3 Factors, hyperpriors; M12 = Large BVAR, hyperpriors; M13 = AR(p), p optimal. See also notes to Table 1.

Table D.2: Test of Equal Predictive Ability - GDP Growth

Benchmark	h	M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12
M02		·											
M03		·	–										
M04		·	·	·									
M05		·	·	·	·								
M06		·	·	·	·	·							
M07	1	·	·	·	·	·	·						
M08		·	·	·	·	·	·	·					
M09		·	–	–	·	·	·	·	–				
M10		·	·	·	·	·	·	·	·	·			
M11		·	·	·	·	·	·	·	·	·	·		
M12		·	–	–	·	·	·	·	·	·	·	·	
M13		·	·	·	·	·	·	·	·	·	·	·	·
M02		·											
M03		·	·										
M04		·	·	·									
M05		·	·	·	·								
M06		·	·	·	·	·							
M07	4	·	·	·	·	·	·						
M08		·	·	·	·	·	·	·					
M09		·	·	·	·	·	·	·	·				
M10		·	·	·	·	·	·	·	·	·			
M11		+	·	·	·	·	·	·	·	·	·		
M12		·	·	·	·	·	·	·	·	·	·	·	
M13		·	·	·	·	·	·	·	·	·	·	–	·

Notes: See notes to Tables 1 and D.1.

Table D.3: Anderson-Darling (AD) Test of Uniformity

	λ		HICP Inflation		GDP Growth	
	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$
M01: Random Walk	0		0.00	0.00	0.02	0.01
M02: BVAR Averaging Equal Weights	∞		0.55	0.63	0.01	0.00
M03: BVAR Averaging Optimal Weights	∞		0.49	0.60	0.01	0.00
M04: BFAVAR 1 Factor	0.328	0.519	0.64	0.54	0.00	0.00
M05: BFAVAR 3 Factors	0.105	0.151	0.21	0.19	0.06	0.00
M06: Large BVAR	0.052	0.051	0.03	0.00	0.02	0.00
M07: BFAVAR 1 Factor, p and λ optimal	0.356	0.267	0.54	0.02	0.01	0.00
M08: BFAVAR 3 Factors, p and λ optimal	0.316	0.258	0.38	0.45	0.27	0.00
M09: Large BVAR, p and λ optimal	0.218	0.171	0.03	0.39	0.11	0.00
M10: BFAVAR 1 Factor, hyperpriors	0.367	0.272	0.71	0.03	0.00	0.00
M11: BFAVAR 3 Factors, hyperpriors	0.314	0.271	0.37	0.52	0.18	0.00
M12: Large BVAR, hyperpriors	0.264	0.203	0.32	0.09	0.04	0.00
M13: AR(p), p optimal	n/a		0.19	0.01	0.12	0.04

Notes: this table shows the p-values of an AD-test of uniformity of the PITs. See also notes to Table 1.

Table D.4: Kolmogorov-Smirnov (KS) Test of Uniformity

	λ		HICP Inflation		GDP Growth	
	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$
M01: Random Walk	0		0.00	0.00	0.18	0.25
M02: BVAR Averaging Equal Weights	∞		0.57	0.86	0.02	0.00
M03: BVAR Averaging Optimal Weights	∞		0.56	0.90	0.01	0.01
M04: BFAVAR 1 Factor	0.328	0.519	0.63	0.69	0.00	0.00
M05: BFAVAR 3 Factors	0.105	0.151	0.18	0.21	0.05	0.03
M06: Large BVAR	0.052	0.051	0.06	0.00	0.05	0.02
M07: BFAVAR 1 Factor, p and λ optimal	0.356	0.267	0.90	0.01	0.00	0.01
M08: BFAVAR 3 Factors, p and λ optimal	0.316	0.258	0.42	0.57	0.63	0.01
M09: Large BVAR, p and λ optimal	0.218	0.171	0.16	0.95	0.15	0.01
M10: BFAVAR 1 Factor, hyperpriors	0.367	0.272	0.97	0.01	0.00	0.00
M11: BFAVAR 3 Factors, hyperpriors	0.314	0.271	0.21	0.77	0.24	0.01
M12: Large BVAR, hyperpriors	0.264	0.203	0.75	0.26	0.09	0.00
M13: AR(p), p optimal	n/a		0.19	0.02	0.15	0.24

Notes: this table shows the p-values of a KS-test of uniformity of the PITs. See also notes to Table 1.

Table D.5: Root Mean Squared Error - Full Dataset

	λ		HICP Inflation		GDP Growth	
	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$
M01: Random Walk	0		0.39	0.43	0.62	0.63
M02: BVAR Averaging Equal Weights	∞		0.26	0.33	0.59	0.75
M03: BVAR Averaging Optimal Weights	∞		0.27	0.31	0.52	0.72
M04: BFAVAR 1 Factor	0.147	0.368	0.24	0.27	0.54	0.66
M05: BFAVAR 3 Factors	0.094	0.170	0.24	0.28	0.52	0.64
M06: Large BVAR	0.040	0.039	0.25	0.29	0.47	0.64
M07: BFAVAR 1 Factor, p and λ optimal	0.356	0.291	0.25	0.28	0.57	0.66
M08: BFAVAR 3 Factors, p and λ optimal	0.371	0.273	0.25	0.28	0.52	0.65
M09: Large BVAR, p and λ optimal	0.175	0.139	0.26	0.28	0.46	0.66
M10: BFAVAR 1 Factor, hyperpriors	0.357	0.268	0.25	0.27	0.56	0.67
M11: BFAVAR 3 Factors, hyperpriors	0.318	0.250	0.24	0.29	0.53	0.68
M12: Large BVAR, hyperpriors	0.080	0.043	0.25	0.29	0.49	0.66
M13: AR(p), p optimal	n/a		0.27	0.31	0.56	0.64

Notes: this table shows the RMSE when the full dataset is used. See also notes to Table 1.

Table D.6: Predictive Likelihood - Full Dataset

	λ		HICP Inflation		GDP Growth	
	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$
M01: Random Walk	0		-2.14	-2.24	-2.26	-2.72
M02: BVAR Averaging Equal Weights	∞		-1.69	-1.65	-1.60	-1.97
M03: BVAR Averaging Optimal Weights	∞		-1.73	-1.63	-1.55	-1.97
M04: BFAVAR 1 Factor	0.147	0.368	-1.49	-1.57	-1.59	-1.99
M05: BFAVAR 3 Factors	0.094	0.170	-1.48	-1.59	-1.56	-2.07
M06: Large BVAR	0.040	0.039	-1.87	-2.00	-1.90	-2.79
M07: BFAVAR 1 Factor, p and λ optimal	0.356	0.291	-1.59	-1.60	-1.68	-2.08
M08: BFAVAR 3 Factors, p and λ optimal	0.371	0.273	-1.66	-1.61	-1.63	-2.10
M09: Large BVAR, p and λ optimal	0.175	0.139	-2.18	-1.98	-1.99	-3.06
M10: BFAVAR 1 Factor, hyperpriors	0.357	0.268	-1.51	-1.57	-1.62	-1.92
M11: BFAVAR 3 Factors, hyperpriors	0.318	0.250	-1.53	-1.58	-1.59	-2.15
M12: Large BVAR, hyperpriors	0.080	0.043	-1.80	-1.94	-1.96	-3.07
M13: AR(p), p optimal	n/a		-1.80	-1.72	-1.61	-1.98

Notes: this table shows the average log predictive likelihood when the full dataset is used. See also notes to Tables 1 and 2.

Table D.7: Root Mean Squared Error - Recursive Estimation Scheme

	λ		HICP Inflation		GDP Growth	
	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$
M01: Random Walk	0		0.66	0.72	0.63	0.63
M02: BVAR Averaging Equal Weights	∞		0.27	0.38	0.56	0.66
M03: BVAR Averaging Optimal Weights	∞		0.27	0.35	0.50	0.65
M04: BFAVAR 1 Factor	0.328	0.519	0.27	0.37	0.53	0.67
M05: BFAVAR 3 Factors	0.105	0.151	0.26	0.32	0.48	0.64
M06: Large BVAR	0.052	0.051	0.26	0.31	0.43	0.64
M07: BFAVAR 1 Factor, p and λ optimal	0.362	0.295	0.29	0.36	0.53	0.66
M08: BFAVAR 3 Factors, p and λ optimal	0.324	0.261	0.27	0.34	0.46	0.64
M09: Large BVAR, p and λ optimal	0.151	0.119	0.28	0.31	0.43	0.63
M10: BFAVAR 1 Factor, hyperpriors	0.298	0.241	0.28	0.35	0.53	0.66
M11: BFAVAR 3 Factors, hyperpriors	0.275	0.265	0.28	0.33	0.44	0.66
M12: Large BVAR, hyperpriors	0.266	0.183	0.27	0.32	0.43	0.64
M13: AR(p), p optimal	n/a		0.28	0.37	0.56	0.64

Notes: this table shows the RMSE when a recursive estimation scheme is used. See also notes to Table 1.

Table D.8: Predictive Likelihood - Recursive Estimation Scheme

	λ		HICP Inflation		GDP Growth	
	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$
M01: Random Walk	0		-1.43	-1.50	-1.71	-1.80
M02: BVAR Averaging Equal Weights	∞		-1.53	-1.44	-1.45	-1.74
M03: BVAR Averaging Optimal Weights	∞		-1.53	-1.41	-1.45	-1.76
M04: BFAVAR 1 Factor	0.328	0.519	-1.55	-1.42	-1.43	-1.78
M05: BFAVAR 3 Factors	0.105	0.151	-1.37	-1.31	-1.37	-1.75
M06: Large BVAR	0.052	0.051	-1.33	-1.25	-1.41	-1.86
M07: BFAVAR 1 Factor, p and λ optimal	0.362	0.295	-1.59	-1.35	-1.43	-1.74
M08: BFAVAR 3 Factors, p and λ optimal	0.324	0.261	-1.56	-1.38	-1.38	-1.73
M09: Large BVAR, p and λ optimal	0.151	0.119	-1.66	-1.35	-1.43	-1.87
M10: BFAVAR 1 Factor, hyperpriors	0.298	0.241	-1.52	-1.35	-1.42	-1.72
M11: BFAVAR 3 Factors, hyperpriors	0.275	0.265	-1.57	-1.39	-1.31	-1.78
M12: Large BVAR, hyperpriors	0.266	0.183	-1.60	-1.40	-1.42	-1.94
M13: AR(p), p optimal	n/a		-1.61	-1.44	-1.46	-1.70

Notes: this table shows the average log predictive likelihood when a recursive estimation scheme is used. See also notes to Tables 1 and 2.

Table D.9: Root Mean Squared Error - Evaluation period 1999:1 to 2011:4

	λ		HICP Inflation		GDP Growth	
	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$
M01: Random Walk	0		0.30	0.31	0.68	0.70
M02: BVAR Averaging Equal Weights	∞		0.29	0.36	0.57	0.79
M03: BVAR Averaging Optimal Weights	∞		0.29	0.36	0.49	0.80
M04: BFAVAR 1 Factor	0.328	0.519	0.27	0.31	0.46	0.72
M05: BFAVAR 3 Factors	0.105	0.151	0.27	0.31	0.49	0.69
M06: Large BVAR	0.052	0.051	0.28	0.30	0.46	0.67
M07: BFAVAR 1 Factor, p and λ optimal	0.356	0.267	0.28	0.31	0.51	0.71
M08: BFAVAR 3 Factors, p and λ optimal	0.316	0.258	0.28	0.31	0.49	0.71
M09: Large BVAR, p and λ optimal	0.218	0.171	0.29	0.30	0.44	0.69
M10: BFAVAR 1 Factor, hyperpriors	0.367	0.272	0.28	0.31	0.47	0.71
M11: BFAVAR 3 Factors, hyperpriors	0.314	0.271	0.28	0.31	0.47	0.73
M12: Large BVAR, hyperpriors	0.264	0.203	0.27	0.31	0.41	0.71
M13: AR(p), p optimal	n/a		0.30	0.32	0.60	0.70

Notes: this table shows the RMSE when the evaluation period is shortened. See also notes to Table 1.

Table D.10: Predictive Likelihood - Evaluation period 1999:1 to 2011:4

	λ		HICP Inflation		GDP Growth	
	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$
M01: Random Walk	0		-1.91	-2.00	-2.29	-2.83
M02: BVAR Averaging Equal Weights	∞		-1.89	-1.83	-1.60	-2.21
M03: BVAR Averaging Optimal Weights	∞		-1.90	-1.87	-1.53	-2.22
M04: BFAVAR 1 Factor	0.328	0.519	-1.74	-1.81	-1.62	-2.29
M05: BFAVAR 3 Factors	0.105	0.151	-1.68	-1.83	-1.65	-2.38
M06: Large BVAR	0.052	0.051	-1.87	-2.02	-1.84	-2.77
M07: BFAVAR 1 Factor, p and λ optimal	0.356	0.267	-1.75	-1.84	-1.73	-2.43
M08: BFAVAR 3 Factors, p and λ optimal	0.316	0.258	-1.80	-1.89	-1.76	-2.44
M09: Large BVAR, p and λ optimal	0.218	0.171	-2.08	-2.08	-1.86	-3.00
M10: BFAVAR 1 Factor, hyperpriors	0.367	0.272	-1.75	-1.76	-1.58	-2.22
M11: BFAVAR 3 Factors, hyperpriors	0.314	0.271	-1.71	-1.85	-1.67	-2.41
M12: Large BVAR, hyperpriors	0.264	0.203	-1.89	-2.06	-1.67	-2.70
M13: AR(p), p optimal	n/a		-2.03	-1.89	-1.80	-2.33

Notes: this table shows the average log predictive likelihood when the evaluation period is shortened. See also notes to Tables 1 and 2.

Table D.11: Root Mean Squared Error - Level Specification

	λ		HICP		GDP	
	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$
M01: Random Walk	0		0.39	1.56	0.62	1.94
M02: BVAR Averaging Equal Weights	∞		0.27	0.97	0.54	1.97
M03: BVAR Averaging Optimal Weights	∞		0.27	0.88	0.50	1.90
M04: BFAVAR 1 Factor	0.253	0.148	0.25	1.08	0.49	1.80
M05: BFAVAR 3 Factors	0.098	0.092	0.26	0.96	0.49	1.81
M06: Large BVAR	0.034	0.041	0.26	0.77	0.53	2.01
M07: BFAVAR 1 Factor, p and λ optimal	0.918	2.104	0.29	0.90	0.49	1.80
M08: BFAVAR 3 Factors, p and λ optimal	0.420	0.520	0.26	0.86	0.46	1.71
M09: Large BVAR, p and λ optimal	0.217	1.443	0.27	0.96	0.53	2.22
M10: BFAVAR 1 Factor, hyperpriors	0.408	0.515	0.25	0.76	0.51	1.94
M11: BFAVAR 3 Factors, hyperpriors	0.350	0.378	0.25	0.84	0.47	1.84
M12: Large BVAR, hyperpriors	0.227	0.579	0.25	0.95	0.45	1.92
M13: AR(p), p optimal	n/a		0.28	0.88	0.58	2.03

Notes: this table shows the RMSE when the models are estimated in levels. See also notes to Table 1.

Table D.12: Predictive Likelihood - Level Specification

	λ		HICP		GDP	
	$h=1$	$h=4$	$h=1$	$h=4$	$h=1$	$h=4$
M01: Random Walk	0		−1.87	−2.06	−1.96	−2.74
M02: BVAR Averaging Equal Weights	∞		−1.68	−1.94	−1.50	−2.38
M03: BVAR Averaging Optimal Weights	∞		−1.72	−1.91	−1.49	−2.42
M04: BFAVAR 1 Factor	0.253	0.148	−1.56	−1.67	−1.52	−2.39
M05: BFAVAR 3 Factors	0.098	0.092	−1.52	−1.66	−1.55	−2.35
M06: Large BVAR	0.034	0.041	−1.65	−1.82	−1.80	−2.97
M07: BFAVAR 1 Factor, p and λ optimal	0.918	2.104	−1.66	−1.83	−1.60	−2.42
M08: BFAVAR 3 Factors, p and λ optimal	0.420	0.520	−1.63	−1.84	−1.61	−2.44
M09: Large BVAR, p and λ optimal	0.217	1.443	−1.92	−3.27	−1.86	−4.68
M10: BFAVAR 1 Factor, hyperpriors	0.408	0.515	−1.53	−1.63	−1.53	−2.57
M11: BFAVAR 3 Factors, hyperpriors	0.350	0.378	−1.58	−1.88	−1.54	−2.54
M12: Large BVAR, hyperpriors	0.227	0.579	−1.72	−2.89	−1.64	−4.08
M13: AR(p), p optimal	n/a		−1.55	−2.00	−1.59	−2.36

Notes: this table shows the average log predictive likelihood when the models are estimated in levels. See also notes to Tables 1 and 2.