Berg, Tim Oliver; Henzel, Steffen

Conference Paper

Point and Density Forecasts for the Euro Area Using Many Predictors: Are Large BVARs Really Superior?

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2013: Wettbewerbspolitik und Regulierung in einer globalen Wirtschaftsordnung - Session: Forecasting, No. E20-V3

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Point and Density Forecasts for the Euro Area Using Many Predictors: Are Large BVARs Really Superior?*

Tim Oliver Berg† Steffen R. Henzel‡

February 8, 2013

Abstract

Forecast models with large cross-sections are often subject to overparameterization leading to unstable parameter estimates and hence inaccurate forecasts. Recent articles suggest that a large Bayesian vector autoregression (BVAR) with sufficient prior information dominates competing approaches. In this paper we evaluate the forecast performance of large BVAR in comparison to its most natural competitors, i.e. averaging of small-scale BVARs and factor augmented BVARs with and without shrinkage. We derive point and density forecasts for euro area real GDP growth and HICP inflation conditional on an information set which is appropriate for all approaches and find no consistent outperformance of the large BVAR. While it produces good point forecasts, the performance is poor when density forecasts are used to evaluate predictive ability. Moreover, the ranking of the different approaches depends *inter alia* on the target variable, the forecast horizon, the state of the business cycle, and on the size of the dataset. Overall, we find that a factor augmented BVAR with shrinkage is competitive in all setups.

Keywords: Bayesian Vector Autoregression, Forecasting, Model Validation, Large Cross-Section, Euro Area

JEL-Codes: C11, C52, C53, E37

*We thank Kai Carstensen, Peter Zorn, and the participants of the Ifo Macro Seminar for helpful comments and suggestions.

†Contact: Tim Oliver Berg, Ifo Institute - Leibniz Institute for Economic Research at the University of Munich e.V., Poschingerstr. 5, 81679 Munich, Germany, Tel.: +49 (0) 89 9224 1689, Fax: +49 (0) 89 985369, email: berg@ifo.de

‡Contact: Steffen R. Henzel, Ifo Institute - Leibniz Institute for Economic Research at the University of Munich e.V., Poschingerstr. 5, 81679 Munich, Germany, Tel.: +49 (0) 89 9224 1652, Fax: +49 (0) 89 985369, email: henzel@ifo.de

1 Introduction

When forecasting economic outcomes, a large set of indicators is wishful in order to avoid model misspecification. However, forecast models with large cross-sections are often subject to over-parameterization leading to unstable parameter estimates and hence inaccurate forecasts. Factor models have been used traditionally to achieve dimension reduction and improve forecast performance.[1] In a recent article Bańbura, Giannone, and Reichlin (2010a) argue, however, that a vector autoregression (VAR) can forecast better even when the number of dependent variables is large. Such a situation usually entails the matrix inversion problem since the number of parameters may easily exceed the number of observations. Bańbura et al. (2010a) propose Bayesian methods to shrink the overparameterized VAR towards a parsimonious random walk benchmark. They impose additional information in form of a Minnesota-type prior and show that large Bayesian VARs (BVARs) can improve on factor models and small VAR systems with only a handful of variables. However, it is unclear what lies behind the good performance of the large BVAR. While the superiority of the large BVAR over smaller systems may be traceable to the fact that it processes a larger information set, the dominance over factor approaches may root in differences pertaining to the modeling and estimation approach.

In this paper we build on the results in Bańbura et al. (2010a) and evaluate the large BVAR by comparing it to its most natural competitors. Therefore, our model comparison relies on different variants of the BVAR model, i.e. averaging of small-scale BVARs, factor augmented BVARs with and without shrinkage, and large BVARs. To the best of our knowledge, averaging of BVARs has not been considered as a competing approach in the related literature before (see e.g. Bańbura et al., 2010a; Giannone, Lenza, and Primiceri, 2012; D'Agostino, Gambetti, and Giannone, 2013; Koop, 2013, among others). In order to ensure comparability across variants, we condition our model validation exercise on a given amount of information. The competing approaches are evaluated according to their out-of-sample forecast performance. Specifically, we forecast the quarterly change in the euro area real gross domestic product (GDP) and the harmonized index of consumer prices (HICP).

We innovate upon previous literature along various dimensions. First, we believe that a potential drawback of the analyses in Bańbura et al. (2010a) and related studies is that these authors compare forecast models of fairly different size. For instance Bańbura et al. (2010a) consider BVARs with 3, 7, 20, and 131 variables as well as a Bayesian factor augmented VAR

[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).

(BFAVAR) with also 131 indicators. The models hence produce forecasts conditional on different information sets, making it difficult to identify whether the relative outperformance of the large BVARs over the smaller systems is due to the mechanics of these models or simply the result of the richer information set. In this paper we aim at revealing possible differences among the competing approaches. For each variant of the BVAR we ensure that forecasts are produced conditional on *all* the information that is contained in our dataset. The model variants differ, however, with respect to the way information is condensed as well as in the complexity and computing time needed to produce a forecast. In order to evaluate whether adding indicator variables is useful at all, we also compare all variants to a random walk benchmark.

Second, most studies restrict their attention to the comparison of point forecasts and rank competing models on the basis of root mean squared errors (RMSE) only (for an exception see Koop, 2013; Giannone et al., 2012). While discriminating among models using RMSE is appropriate if the loss function of the forecaster depends solely on the forecast error, we argue that neglecting the uncertainty surrounding the forecasts might be highly misleading. For instance, it is now well documented that in particular monetary policymakers increasingly draw their attention to the uncertainty that is associated with business cycle and price developments. The density forecasts of the Bank of England's Monetary Policy Committee for U.K. inflation or those of the *Sveriges Riksbank* for Swedish inflation are prominent examples (see, e.g. Mitchell and Hall, 2005; Boero, Smith, and Wallis, 2011; Knüppel and Schultefrankenfeld, 2012, among others). In this paper we hence also rely on density forecasts to rank the competing variants. In this context, the predictive density is a standard tool to compare different models (see, e.g., Geweke and Amisano, 2010; Giannone et al., 2012; D'Agostino et al., 2013; Koop, 2013, among others).

Third, we also consider likelihood ratio tests as in Amisano and Giacomini (2007) to compare competing BVAR variants. These frequentist tests are based on weighted predictive densities and may also be applied to Bayesian predictive distributions. By focusing on different regions of the distribution, we are able to check how sensitive our results are with respect to extreme outliers and may compare model performance during normal times versus tail events. Given that our sample period covers several euro area recessions and booms as well as severe crises, we believe that it is interesting to inspect the underlying mechanics of the variants.

Fourth, unlike most existing studies, we turn our attention to aggregate euro area data. Our dataset comprises 44 quarterly macroeconomic and financial indicators spanning the years 1975 to 2009. 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 very specific case of the U.S. translate to other forecast situations. In our view, we consider a set of indicator variables 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 the variants we consider. Even for factor models it has been shown that approximately 40 series are sufficient to yield satisfactory forecast accuracy (see Bai and Ng, 2002; Boivin and Ng, 2006).

The results can be summarized as follows. All approaches substantially outperform the random walk benchmark for HICP inflation, suggesting that the dataset contains valuable information. Among these variants, the large BVAR delivers almost the best point forecast; for the first quarter horizon the BFAVAR with shrinkage is marginally better. Similarly, for GDP growth the large BVAR outperforms the other approaches in the short term while the BFAVAR with shrinkage is competitive. Overall, the large BVAR appears to be a good choice when the user is interested in point forecasts. However, when we use density forecasts to evaluate the performance of the BVAR variants it turns out that the ranking is almost reversed, and the large BVAR performs worst. However, the BFAVAR with shrinkage now provides the best forecast at short horizons whereas the BVAR averaging outperforms at the longer horizon. The likelihood ratio test reveals that all variants perform significantly better than the large BVAR whereas the remaining variants have a similar forecast performance. Notably, such a result does not solely depend on extreme or rare events. Even when we concentrate on normal times only, the BFAVAR with shrinkage outperforms the large BVAR in terms of predictive density. However, the large BVAR performs exceptionally poor during more turbulent times. Finally, we analyze the sensitivity of our results when we reduce the cross-sectional dimension of the dataset to 22 variables. While point forecasts remain virtually unaffected, we find that the ranking is affected when we consider density forecasts. Using the medium-scale dataset it turns out that the relative performance of the large BVAR improves. This is probably due to the fact that we impose a less restrictive prior on the coefficients of the large BVAR. Nevertheless, a BFAVAR with shrinkage remains among the best performing variants.

The remainder of this paper is organized as follows. In Section 2 we describe our dataset. Section 3 develops the BVAR model and all the variants we use to produce out-of-sample forecasts for euro area real GDP growth and HICP inflation. In Section 4 we explain our forecast experiment and present the main results. The results for the medium-size dataset are discussed in Section 5. Section 6 concludes. In the Appendix we provide additional information on our dataset and tables for the medium-size dataset.

2 Dataset

The dataset comprises 44 quarterly euro area macroeconomic and financial time series covering the period 1975:1 to 2009:4. 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 real GDP and the overall HICP as the variables 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. We take the natural logarithm of most series, except of those that are already expressed in rates, such as unemployment or interest rates. In most cases the series are obtained from the 10th (and most recent) update of the Area-wide Model (AWM) database which is maintained by the European Central Bank (ECB). The AWM database is a unique and rich source for aggregate euro area data. The historical series are backdated by the ECB staff using individual country information in a coherent manner. Moreover, the AWM database is the preferred source for researchers and policymakers alike interested in topics relevant for the euro area. In addition, survey data, monetary aggregates, and a share price index are downloaded from Datastream. 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 real GDP and the HICP. Besides the fact that forecasting euro area data is interesting in its own right, note that for countries other than the U.S. 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. In addition, the size of our cross-section is appropriate with respect to all the variants of the BVAR that we consider. Although factor approaches are designed for even larger cross-sections, Bai and Ng (2002) as well as Boivin and Ng (2006) show that approximately 40 series are sufficient to yield satisfactory forecast accuracy.

We have to stress at this point, however, that the euro area did not exist before January 1999 and vintage series are hence not available. 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 to produce out-of-sample forecasts for euro area real GDP growth and HICP inflation in the next section.

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, real GDP and the HICP; 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 $k = 7788$). 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 many 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. 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. The VAR is hence centered around the prior mean $y_{i,t} = c_i + y_{i,t-1} + u_{i,t}$ and imposing the prior amounts to shrinking the diagonal elements of B_1 towards one and the remaining coefficients in $B_1, ..., B_p$ 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 (1997) which allows for correlation among residual terms. The evidence in Bańbura et al. (2010a) 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.[2] We follow Bańbura et al. (2010a) and implement the prior by constructing the following set of artificial observations:

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

where diag$(\cdot)$ denotes a diagonal matrix. The hyperparameters δ_i are all set equal to 1, reflecting the prior belief that all variables are characterized by high persistence. 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 an initial sample running from 1975:1 to 1984:4. The lag order is the same as in the VAR. 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. λ is hence the key parameter in the BVAR and its calibration for each model variant is explained in detail in Section 3.2.

In order to further improve the forecast accuracy of BVARs, the literature proposes to consider additional prior information in form of a "sum-of-coefficients" prior (see, e.g., Robertson and Tallman, 1999; Bańbura et al., 2010a; 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:

$$y^{++} = \text{diag}(\mu_1, ..., \mu_n)/\tau\ , \qquad x^{++} = [(1\ 2 ... p) \otimes \text{diag}(\mu_1, ..., \mu_n)/\tau \quad 0_{n\times 1}]\ ,$$

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

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 initial sample period 1975:1 to 1984:4.

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

$$y^* = x^* B + u^*,$$

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:

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

where $\hat{\Sigma}$ and $\hat{\beta}$ are the covariance matrix and the coefficients from an OLS regression of y^* on x^*, respectively. In principle, the one-step-ahead predictive density $p\left(y_{T+1} | \beta, \Sigma, y\right)$ would also be available in closed form. When forecasting more than one period ahead, however, an analytical expression for the posterior predictive density does not exist since forecasts are non-linear combinations of model parameters. In this case we use a Gibbs sampler and we sequentially draw a covariance matrix (given the data), coefficients (given the covariance matrix and the data), residual terms (given the covariance matrix and the data), and produce out-of-sample forecasts up to horizon H. We repeat this cycle 500 times and obtain the posterior predictive density at all horizons by smoothing the empirical distribution of forecasts using a normal kernel function.[3]

3.2 Model Variants

We consider the following variants that are all nested in the BVAR. The variants differ, however, with respect to the way information is condensed. One important difference between the variants lies in the degree of shrinkage λ. As emphasized by De Mol, Giannone, and Reichlin (2008), a requirement for the hyperparameter λ is that the degree of shrinkage increases with the cross-sectional dimension. To determine λ we rest on the assumption that a three-variable VAR system is parsimonious and hence does not suffer from overparameterization. The Bayesian shrinkage procedure follows Bańbura et al. (2010a) by choosing λ such that the average in-sample fit for real GDP and the HICP of all variants is the same during the initial sample period

[3]See for instance D'Agostino et al. (2013) for a similar procedure.

from 1975:1 to 1984:4. That is, each model is shrunk to the size of a parsimonious VAR. We obtain the desired magnitude of fit by performing a search over a fine grid for λ. Notably, this procedure maintains the comparability across variants, and it ensures that we shrink more when the size of the model increases.

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

BVAR Averaging In the second variant we estimate a variety of three-variable BVARs each including real GDP and the HICP and one indicator variable at a time. 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; Henzel and Mayr, 2013, among others). This is motivated by portfolio diversification or hedging arguments, guaranteeing insurance against large forecast errors. Moreover, by segmenting the set of indicators and estimating a battery of parsimonious models, pooling is a way to condense information and avoid parameter proliferation. Note that BVAR averaging is straightforward and economizes computing time as the small-scale VAR models are easily estimated. We obtain predictions for real GDP and the HICP by averaging forecasts across models using equal weights. In particular, we follow Wallis (2005) and simulate a posterior distribution of mean forecasts. In each cycle we construct forecasts for each of the 42 models and record the arithmetic mean. We repeat this exercise 500 times and obtain an empirical distribution of mean forecasts for both real GDP and the HICP at all horizons. Since we assume that three-variable VAR models are not subject to overparameterization, we impose an uninformative prior, i.e. we set $\lambda = \infty$.

BFAVAR-1F It is often argued that factor augmented regression models are successful in achieving dimension reduction and 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 real GDP and the HICP and one factor as the third variant. 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 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 a BVAR by the first factor. Since the factor is extracted from differenced series, a random walk prior would not be appropriate in this case and we thus impose a white noise prior instead, i.e. we set $\delta_i = 0$. As the BVAR system consists of only three variables, we assume that overparameterization is not an issue here and perform the estimation without shrinkage and set $\lambda = \infty$. Similar to BVAR averaging, the method economizes on computing time as principal components are readily computed and the estimation involves only small-scale VARs.

BFAVAR-3F 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 apply shrinkage to further reduce the dimension of the system. Hence, this variant combines the advantages of the factor approach with Bayesian shrinkage. We set the shrinkage parameter λ such that the average in-sample fit of the BFAVAR for real GDP and the HICP in the initial sample period 1975:1 to 1984:4 is the same as that of the BVAR averaging. Note that the BVAR augmented with three factors is computationally more demanding than the BFAVAR-1F since simulation of the posterior involves repeated inversion of large matrices.

Large BVAR In the last variant we estimate the BVAR on all the 44 series at the same time. The literature refers to this variant as a large BVAR (see, e.g., Bańbura et al., 2010a; Giannone et al., 2012; Koop, 2013, among others). In order to deal with the dense parameterization of the model and to maintain comparability across specifications, we again apply Baysian shrinkage and choose λ such that the average in-sample fit for real GDP and the HICP during the period 1975:1 to 1984:4 is the same as that of the BFAVAR and the BVAR averaging. Notably, the large BVAR is by far the most computationally demanding and time consuming variant analyzed in this paper.

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 then present results for the predictive mean and the entire predictive density of real GDP growth and HICP inflation.

4.1 Forecast Experiment

We use each of the five BVAR variants to produce out-of-sample forecasts for real GDP and the HICP for four quarters. We start with the initial sample period from 1975:1 to 1984:4 and generate the posterior predictive density of each variant for the horizon 1985:1 to 1985:4. This procedure is iterated forward until 2008:4, producing forecasts for 2009:1 to 2009:4, always using the most recent 10 years of data and yields a sequence of 97 density forecasts for each variant. The evaluation period thus runs from 1985:1 to 2009:4 and coincides with the "Great Moderation" period of low and stable volatility.

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).

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 10 years seems to account well for both concerns.

For both real GDP and the HICP the evaluation target is the quarter-on-quarter growth rate, i.e. $\Delta y_{i,T+h} = y_{i,T+h} - y_{i,T+h-1}$, where h denotes the forecast horizon and T the last data point used in estimation. We hence evaluate out-of-sample forecast accuracy based on real GDP growth and HICP inflation rather than their respective level forecasts since these are the variables forecasters, policymakers, and researchers alike are typically interested in.

4.2 Predictive Mean

To begin with, we report the shrinkage hyperparameter λ in the last row of Table 1. These numbers are the result of the calibration procedure which maintains the average in-sample fit for real GDP and the HICP fixed across variants. As we emphasize above, for the BVAR averaging and the BFAVAR with one factor we assume that the overparameterization problem does not occur. Hence, the shrinkage parameter is set to $\lambda = \infty$, reflecting an uninformative prior. In case of the BFAVAR with three factors the prior becomes tighter and λ is 0.645. For the large BVAR with all the 44 variables we have to impose a lot more shrinkage, and λ reduces to 0.059.

In the following, we evaluate the different BVAR variants by the accuracy of the point forecast. Let the superscript m denote the mean of the predictive density or the predictive mean. We measure out-of-sample forecast accuracy for model variant λ with respect to variable i in terms of RMSE:

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

where $H = 4$ is the maximal forecast horizon, and $T_0 = 40$ and $T_1 = 140$ are the start and end 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 discriminate among model specifications (see Weiss, 1996). In Table 1 we report the RMSE for the different model variants.

Table 1: Root Mean Squared Errors

	Random Walk		BVAR Averaging		BFAVAR-1F		BFAVAR-3F		Large BVAR	
h	GDP	HICP	GDP	HICP	GDP	HICP	GDP	HICP	GDP	HICP
1	0.63	0.55	0.69	0.30	0.60	0.31	0.57	0.26	0.50	0.27
2	0.64	0.57	0.74	0.36	0.70	0.34	0.70	0.32	0.59	0.30
3	0.64	0.58	0.80	0.40	0.75	0.38	0.76	0.38	0.66	0.32
4	0.65	0.58	0.81	0.46	0.76	0.44	0.83	0.44	0.70	0.36
λ	0		∞		∞		0.645		0.059	

Notes: this table shows the RMSE for different BVAR variants. These variants include the random walk, the BVAR averaging, the BFAVAR with 1 and 3 factor(s), and the large BVAR with 44 variables. The hyperparameter λ determines the degree of Bayesian shrinkage and is set such that the average in-sample fit for real GDP and the HICP of all variants is the same during the initial sample period 1975:1 to 1984:4. The evaluation period runs from 1985:1 to 2009:4 and the forecast horizon h is in quarters. The variables we forecast are quarterly real GDP growth and HICP inflation. The lower the RMSE is the better is the forecast accuracy of a BVAR variant.

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 all horizons. When using a large BVAR for instance the forecaster could cut the RMSE by about half at $h = 1$ or one third at $h = 4$. Substantial improvements compared to the random walk forecast are also possible with the other variants.[4] We therefore conclude that the set of indicators contains valuable information about the prospective path of HICP inflation.

Second, we obtain rather small differences in the forecast accuracy of these BVAR variants. In some instances, the BVAR averaging appears to underperform the other variants. Overall, the RMSE ranges between 0.26 and 0.31 at $h = 1$ as well as 0.36 and 0.46 at $h = 4$, showing a marginal outperformance of the BFAVAR-3F with shrinkage at $h = 1$ and the large BVAR at $h = 2$ to 4. In Table B.1 in the Appendix we provide the results of 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 has a limited impact on the forecast accuracy of the predictive mean for HICP inflation.

Third, for real GDP growth the large BVAR delivers the best forecast performance at the short horizon $h = 1$ and 2. By contrast, the random walk produces the best forecast for real GDP growth at longer horizons $h = 3$ and 4, suggesting that the set of indicators contains little to no information about changes in real GDP one year ahead. The relatively poor forecast performance of all the variants with respect to quarterly GDP growth in the longer run is mainly due to the fact that the series shows little to no persistence and is hence largely driven by unpredictable shocks.[5]

Fourth, the differences in forecast accuracy with respect to real GDP growth are again rather small, particularly as the differences in predictive ability between these variants are insignificant in most setups (compare Table B.1 in the Appendix). The RMSE varies between 0.50 and 0.69 at $h = 1$ as well as 0.65 and 0.81 at $h = 4$. Overall, there appears to be a slight dominance of the large BVAR but the BFAVAR with shrinkage is also competitive.

4.3 Predictive Density

In the previous subsection we have compared different model variants on the basis of RMSE, which is appropriate if the forecaster is concerned only about the accuracy of the predictive mean but indifferent to the uncertainty that is surrounding it. In this subsection we relax

[4] Table B.1 in the Appendix reveals that these differences in forecast ability are significant according to a Giacomini and White (2006) test of conditional predictive ability.

[5] Note that we do not claim here that current quarter GDP growth is unpredictable. However, the VAR model class does not make use of valuable within-quarter information such as monthly industrial production or business survey data. For an extensive discussion on methods using mixed frequency or unsynchronized data ("bridging models" or "jagged edge data") to improve the forecast of current quarter GDP growth ("nowcasting"), we refer to *inter alia* Giannone, Reichlin, and Small (2008), Angelini, Bańbura, and Rünstler (2010), Bańbura, Giannone, and Reichlin (2010b), as well as Angelini, Camba-Mendez, Giannone, Reichlin, and Rünstler (2011).

this strong assumption and consider the entire density of the forecasts. In particular, we rank BVAR variants based on the log predictive density, 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 log predictive density 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). A perfect BVAR variant would hence generate a forecast that is equal to the actual outcome with a probability of 100%. The log predictive density is thus zero in the limit.

In Table 2 we document the average log predictive density of each variant λ with respect to variable i at horizon h evaluated at the realized value $\Delta y_{i,T+h}$:

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

The most important conclusion that we draw from Table 2 is that the ranking of the BVAR variants dramatically changes when the focus shifts from the predictive mean to the entire density of the forecasts. While the random walk still produces the worst forecast for HICP inflation at all horizons, the performance of the large BVAR substantially deteriorates when forecast uncertainty is taken into account. Despite its quite accurate mean forecast, the large BVAR apparently attaches a too low probability to events that actually occur. This seems to be traceable to the tightness of the prior restrictions which seems to result in a predictive density that is too concentrated around the predictive mean rather than the actual realization. By contrast, the BFAVAR with three factors does not lose much of its forecast performance and is now delivering the best forecast at the short horizon $h = 1$ and 2. Notably, the BVAR averaging now dominates the other approaches at the longer horizons $h = 3$ and 4. The predictive ability of the BVAR averaging and the BFAVAR with three factors with respect to HICP inflation is, however, similar.

For real GDP growth the ranking of the BVAR variants is completely reversed compared to the ranking based on RMSE. Both the random walk and the large BVAR now display a poor forecast performance at all horizons. As in the case of HICP inflation, the tight prior restrictions imposed seem to generate a predictive density that is too concentrated around the predictive mean. In contrast, the BVAR averaging, which produces rather inaccurate mean forecasts, now performs exceptionally well and appears to dominate the other approaches at longer horizons $h = 2$ to 4. Note that the BFAVAR with three factors remains competitive even when we consider the entire density to evaluate the performance. It generates the best forecast at $h = 1$.

Table 2: Average Log Predictive Densities

	Random Walk		BVAR Averaging		BFAVAR-1F		BFAVAR-3F		Large BVAR	
h	GDP	HICP	GDP	HICP	GDP	HICP	GDP	HICP	GDP	HICP
1	-2.86	-2.82	-1.82	-2.00	-1.94	-2.22	-1.79	-1.82	-2.24	-2.49
2	-3.29	-3.07	-1.81	-2.00	-2.07	-2.04	-1.91	-1.93	-2.69	-2.55
3	-3.45	-3.14	-1.72	-1.86	-1.95	-1.88	-1.95	-1.99	-3.08	-2.55
4	-3.71	-3.20	-1.71	-1.89	-1.95	-2.03	-2.16	-2.15	-3.25	-2.71
λ	0		∞		∞		0.645		0.059	

Notes: this table shows the average log predictive densities for different BVAR variants. The higher the average log predictive density is the better is the predictive ability of a variant. See also notes to Table 1.

In sum we find that an accurate mean forecast does not necessarily imply a good predictive ability. Model selection based on RMSE only may hence be misleading if the user is also concerned about forecast uncertainty.

4.4 Weighted Likelihood Ratio Tests

Given the rankings obtained in the previous section, we now investigate whether the predictive densities in Table 2 are significantly different from each other. We follow Amisano and Giacomini (2007) and construct for two competing density forecasts λ^1 and λ^0 a weighted average of likelihood ratios:

$$\mathrm{WLR}_{i,h}^{\lambda^1,\lambda^0} = \frac{1}{T_1 - T_0 - H + 1} \sum_{T=T_0+h}^{T_1-H+h} \mathrm{w}\left(\Delta y_{i,T+h}^{st}\right) \left(\log p\left(\Delta y_{i,T+h} | y, \lambda^1\right) - \log p\left(\Delta y_{i,T+h} | y, \lambda^0\right)\right),$$

where $\Delta y_{i,T+h}^{st}$ is the realized value, standardized using an estimate of the unconditional mean and standard deviation of $\Delta y_{i,t}$, and the weight function $\mathrm{w}(\cdot)$ is chosen to select a desired region of the distribution of $\Delta y_{i,t}$.

In particular, we consider three different choices for the weight function. First, we choose $\mathrm{w}(\cdot) = 1$ and construct an unweighted average of likelihood ratios. Second, we set $\mathrm{w}(\cdot) = \phi(\cdot)$, with ϕ denoting the standard normal probability density function, which allows us to focus on the center of the distribution. Realized values that are near to the unconditional mean of $\Delta y_{i,t}$ receive a higher weight than those at the tails of the distribution. Third, we select $\mathrm{w}(\cdot) = 1 - \phi(\cdot)/\phi(0)$, meaning that we attach a higher weight to values at the tails of the distribution than to those close to the unconditional mean.

Table 3: Unweighted Likelihood Ratio Tests

Benchmark	h	Random Walk GDP	Random Walk HICP	BVAR Averaging GDP	BVAR Averaging HICP	BFAVAR-1F GDP	BFAVAR-1F HICP	BFAVAR-3F GDP	BFAVAR-3F HICP
BVAR Averaging	1	-1.04	**-0.82**						
BFAVAR-1F		-0.92	-0.60	0.12	0.22				
BFAVAR-3F		**-1.07**	**-1.00**	-0.03	-0.18	-0.15	**-0.40**		
Large BVAR		-0.62	-0.33	0.42	**0.49**	0.30	0.28	**0.46**	**0.67**
BVAR Averaging	2	-1.49	**-1.08**						
BFAVAR-1F		-1.23	**-1.03**	0.26	0.05				
BFAVAR-3F		**-1.38**	**-1.15**	0.11	-0.07	-0.16	-0.11		
Large BVAR		-0.60	**-0.52**	**0.88**	**0.56**	**0.62**	**0.51**	**0.78**	**0.63**
BVAR Averaging	3	-1.73	**-1.28**						
BFAVAR-1F		-1.49	**-1.26**	0.23	0.02				
BFAVAR-3F		**-1.50**	**-1.15**	0.23	0.13	-0.01	0.11		
Large BVAR		-0.37	**-0.59**	**1.36**	**0.69**	**1.13**	**0.67**	**1.13**	**0.56**
BVAR Averaging	4	-2.00	**-1.31**						
BFAVAR-1F		-1.76	**-1.18**	0.24	0.13				
BFAVAR-3F		**-1.55**	**-1.06**	0.46	0.25	0.21	0.12		
Large BVAR		-0.46	-0.49	**1.55**	**0.81**	**1.31**	**0.68**	**1.09**	**0.56**

Notes: the entries are unweighted averages of likelihood ratios as in Amisano and Giacomini (2007). A positive (negative) value means that the BVAR variant under consideration outperforms (underperforms) its benchmark variant in terms of predictive ability. Bold entries denote significance at the 5 percent level. See also notes to Table 1.

For a uniform weight function, the test is a conventional likelihood ratio test often used for model selection. Comparing weighted averages of likelihood ratios provides some interesting additional information. For instance, we may evaluate to what extent extreme outliers influence our results, which seems important since the sample period includes two severe crises: the breakdown of the European Exchange Rate Mechanism in 1992/93 and the global financial crisis in 2008/09. Moreover, we may compare model performance during normal times vs. tail events. The latter covering for example recessions or deflationary episodes.

The test is based on the statistic

$$t = \frac{\mathrm{WLR}_{i,h}^{\lambda^1,\lambda^0}}{\hat{\sigma}/\sqrt{T_1 - T_0 - H + 1}},$$

where $\hat{\sigma}$ is an estimate of the standard deviation of $\mathrm{WLR}_{i,h}^{\lambda^1,\lambda^0}$. The null hypothesis of equal performance of λ^1 and λ^0 is rejected at the 5 percent level whenever $|t| > 1.96$. In case of

Table 4: Center-Weighted Likelihood Ratio Tests

		Random Walk		BVAR Averaging		BFAVAR-1F		BFAVAR-3F	
Benchmark	h	GDP	HICP	GDP	HICP	GDP	HICP	GDP	HICP
BVAR Averaging	1	0.05	0.07						
BFAVAR-1F		**0.08**	**0.14**	0.03	0.07				
BFAVAR-3F		0.06	−0.00	−0.04	**-0.07**	**-0.07**	**-0.04**		
Large BVAR		0.03	**0.10**	−0.02	0.04	−0.05	0.28	0.02	**0.10**
BVAR Averaging	2	0.04	0.06						
BFAVAR-1F		0.05	0.07	0.01	0.06				
BFAVAR-3F		0.03	−0.00	**-0.04**	−0.06	**-0.05**	−0.07		
Large BVAR		**0.06**	**0.10**	0.02	0.04	0.01	0.03	**0.06**	**0.10**
BVAR Averaging	3	0.02	−0.00						
BFAVAR-1F		0.03	0.02	0.01	0.02				
BFAVAR-3F		−0.01	0.01	−0.03	0.01	−0.04	−0.02		
Large BVAR		**0.08**	0.06	0.06	**0.06**	0.05	0.04	**0.10**	**0.05**
BVAR Averaging	4	−0.01	0.03						
BFAVAR-1F		0.00	0.07	0.01	0.04				
BFAVAR-3F		−0.02	0.05	−0.01	0.02	−0.02	−0.02		
Large BVAR		**0.08**	**0.10**	**0.09**	**0.07**	**0.08**	0.03	**0.10**	**0.05**

Notes: the entries are center-weighted averages of likelihood ratios as in Amisano and Giacomini (2007). A positive (negative) value means that the BVAR variant under consideration outperforms (underperforms) its benchmark variant in terms of predictive ability. Bold entries denote significance at the 5 percent level. See also notes to Table 1.

rejection, one would choose λ^1 (λ^0) whenever $\mathrm{WLR}_{i,h}^{\lambda^1,\lambda^0}$ is positive (negative).

We perform pairwise weighted likelihood ratio tests using our sequence of 97 density forecasts and show their results in Tables 3 to 5. The entries in the tables are the values of $\mathrm{WLR}_{i,h}^{\lambda^1,\lambda^0}$. Whenever the value is positive (negative) the BVAR variant under consideration outperforms (underperforms) its benchmark variant in terms of predictive ability. Bold entries denote significance at the 5 percent level.

In Table 3 we document the results for the unweighted case. Consistent with the findings of the previous subsection, the random walk is outperformed by all other variants at all horizons for all variables. For inflation the likelihood ratio is also significantly different from zero in most cases, while for GDP growth only the BFAVAR with three factors does significantly better than the random walk. Most notably, the large BVAR displays a poor predictive ability for both GDP growth and inflation. The large BVAR is typically outperformed by BVAR averaging and both

Table 5: Tails-Weighted Likelihood Ratio Tests

		Random Walk		BVAR Averaging		BFAVAR-1F		BFAVAR-3F	
Benchmark	h	GDP	HICP	GDP	HICP	GDP	HICP	GDP	HICP
BVAR Averaging	1	**-1.16**	**-0.98**						
BFAVAR-1F		**-1.11**	**-0.95**	0.06	0.03				
BFAVAR-3F		**-1.09**	**-0.99**	0.08	−0.01	0.02	−0.04		
Large BVAR		−0.68	**-0.58**	**0.48**	**0.40**	**0.42**	**0.37**	**0.40**	**0.41**
BVAR Averaging	2	**-1.60**	**-1.23**						
BFAVAR-1F		**-1.36**	**-1.19**	0.24	0.03				
BFAVAR-3F		−1.39	**-1.14**	0.21	−0.06	−0.03	0.05		
Large BVAR		−0.75	**-0.76**	**0.85**	**0.47**	**0.61**	**0.44**	**0.64**	**0.38**
BVAR Averaging	3	−1.78	**-1.28**						
BFAVAR-1F		**-1.56**	**-1.32**	0.22	−0.04				
BFAVAR-3F		**-1.46**	**-1.16**	0.31	0.01	0.10	0.16		
Large BVAR		−0.57	**-0.73**	**1.21**	**0.54**	**0.99**	**0.59**	**0.89**	**0.43**
BVAR Averaging	4	−1.99	**-1.38**						
BFAVAR-1F		**-1.77**	**-1.35**	0.22	0.03				
BFAVAR-3F		**-1.51**	**-1.17**	0.48	0.02	0.26	**0.17**		
Large BVAR		−0.66	**-0.76**	1.34	**0.64**	**1.11**	**0.61**	**0.85**	**0.43**

Notes: the entries are tails-weighted averages of likelihood ratios as in Amisano and Giacomini (2007). A positive (negative) value means that the BVAR variant under consideration outperforms (underperforms) its benchmark variant in terms of predictive ability. Bold entries denote significance at the 5 percent level. See also notes to Table 1.

variants of the BFAVAR. Among the latter three, the predictive ability does not differ from each other significantly in all but one cases.

The outcome of the center-weighted test is reported in Table 4. Two striking differences compared to the unweighted case are readily apparent. First, the average likelihood ratios are much smaller when we focus on the center of the distribution of $\Delta y_{i,t}$. Second, we obtain many insignificant values in the first two columns of Table 4 which suggests that the random walk is hard to beat when we are in normal times. However, it should be emphasized that the BFAVAR with three factors tends to outperform most of the other approaches. In particular it has a significantly better forecast ability than the large BVAR in all but one cases (GDP at $h = 1$). We thus conclude that the ranking obtained in Table 3 is not solely driven by extreme or rare events.

Table 5 collects the results of the tails-weighted test procedure. It turns out that the poor (unweighted) forecast performance of the random walk seems to be largely driven by tail events which suggests that information provided by indicator variables is valuable particularly during turbulent times. Most notably, the large BVAR is significantly outperformed by BVAR averaging and both BFAVARs in all but one constellations (GDP at $h = 4$). It appears that more extreme events that actually occur from time to time render the forecast performance of the large BVAR exceptionally bad. Finally, the differences between the remaining three BVAR variants (BVAR averaging, BFAVAR-1F, BFAVAR-3F) in Table 5 are only minor and in most cases insignificant.

5 Medium-Size Dataset

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, alternative loss functions, and different regions of the distribution of the target variable. However, the relative performance may also depend on the size of the dataset since each variant may be affected differently by the amount of cross-sectional information. In this section we investigate whether the results of our forecast experiment are sensitive to changes in the size of the dataset. We diminish the cross-section of the dataset and repeat the entire forecast experiment with a medium-size dataset consisting of 22 variables. The analysis can be motivated as follows. First of all, Bańbura et al. (2010a) argue that VARs with about 20 indicator variables produce reasonable forecasts and that adding more information improves the forecast performance only marginally. And second, in practice it is common not to use all the series available but to extract a subset of indicators - either based on past experience or on the basis of sound economic arguments. The latter consideration typically leads to information sets like the one we consider in this section. We provide details on the exact composition of the medium-size dataset in Appendix A. With respect to both the number and type of indicator variables, the medium-size dataset is similar to those typically considered in the related literature (see, e.g., Bańbura et al., 2010a; Giannone et al., 2012; Koop, 2013, among others).

We report the results for the medium-size dataset in Tables C.1 to C.5 in Appendix C. To begin with, Table C.1 shows that we have to shrink less in case of the BFAVAR with three factors and the large BVAR since the size of our cross-section decreased. As a result, λ increases from 0.645 to 0.880 (BFAVAR-3F) and from 0.059 to 0.081 (large BVAR). Notably, we obtain point forecasts and hence RMSE that are similar to those obtained with the larger dataset, supporting the notion that using more indicator variables does not always lead to a better outcome. For HICP inflation the large BVAR is now slightly better than the BFAVAR with three factors also at $h = 1$, suggesting that the reduction of variables does not harm the performance of the large BVAR. Moreover, the ranking obtained for the GDP forecasts does not change.

Table C.2 documents that the average log predictive density of the large BVAR improves when using the medium-size dataset, while the predictive ability of the BFAVARs and the BVAR averaging is virtually unaffected by the size of the cross-section. This finding squares with our interpretation of the results in Section 4.4. It appears that we have to shrink too much in case of the large BVAR when the size of the cross-section is large, resulting in a predictive density that is too concentrated around the predictive mean and hence attaches a too low probability to events that actually occur. Similarly, the likelihood ratio tests in Table C.4 show that the large BVAR is at least not significantly worse than other variants when the medium-size dataset is used. The improved relative performance of the large BVAR is also reflected in the center-weighted test results. It turns out that the GDP forecasts of the large BVAR now significantly outperform the BFAVAR with one factor. Finally, even at the tails of the distribution the differences between the variants become largely insignificant.

6 Summary and Conclusion

Recently, a number of studies have argued that large VARs combined with Bayesian shrinkage (large BVARs) seem to outperform other modeling approaches which had been proposed to overcome the overparameterization problem. In this paper we evaluate different variants of the general BVAR model, i.e. averaging of small-scale BVARs, factor augmented BVARs with and without shrinkage (BFAVARs), and large BVARs. These variants of the BVAR model differ in the way information is condensed as well as in the complexity and computing time needed for estimation. To evaluate how these variants process information contained in a large dataset, we condition our analysis on a given amount of information. The proposed variants are evaluated according to their out-of-sample forecast performance. In particular, we analyze whether the relative superiority of the large BVAR is maintained in our setup. To this end, we predict euro area real GDP growth and HICP inflation using a dataset which we believe is, first, adequate for all modeling approaches and, second, comprehensive. That is, we consider a set of 42 indicator variables that most forecasters would probably label a "typical" dataset because it contains all variables that are on the wish list of a practical forecaster. Note that for most countries long time series for hundreds of indicator variables are not available. Hence, the size of our dataset strikes a balance between the maximum availability of the cross-sectional and the time series dimension of the predictors.

To begin with, we evaluate the point forecasts and find that all approaches substantially outperform a naïve forecast for HICP inflation. For quarterly GDP growth the naïve scheme beats only the BVAR averaging. It turns out that the large BVAR delivers almost the best point forecast. For HICP inflation the factor augmented VAR with Bayesian shrinkage does marginally better. Hence, the large BVAR appears to be a good choice when the user is interested in point forecasts. However, many policy makers are interested in density forecasts which take into account the uncertainty surrounding a prediction. Consequently, we also study whether the loss function has an impact on the ranking of the different modeling strategies. That is, we use density forecasts to evaluate the performance of the BVAR variants. It turns out that the large BVAR performs worst in the sense that it provides the lowest predictive density. When we conduct a likelihood ratio test, the other approaches perform significantly better with slight advantages for the BFAVAR with shrinkage at short horizons and the BVAR averaging at longer horizons.

Moreover, we also provide a breakdown of the results conditional on the region of the distribution of the target variable. At the center of the distribution, the differences between the approaches are small. However, the large BVAR does not significantly outperform any of the other approaches whereas it is dominated by the BFAVAR with three factors and shrinkage. At the tails of the distribution, the large BVAR delivers forecasts which are significantly worse than those of any other BVAR variant. We hence emphasize that the large BVAR attaches more weight to prior information than the other variants. Overall, we conclude that the prior we have to impose on the large BVAR is probably too restrictive when it comes to density forecasting. As a result, the large BVAR attaches too little probability to extreme events. The BFAVAR with shrinkage appears to circumvent this problem. This is probably due to the fact that we have to shrink less because the number of coefficients is comparably small when the information is condensed in only a few factors.

Finally, we analyze the sensitivity of our results when we reduce the cross-sectional dimension of the dataset to 22 variables. We find that the ranking of the different BVAR variants is affected by the size of the dataset when we consider density forecasts. It turns out that the relative performance of the large BVAR improves when we rely on the medium-scale dataset. This result is probably due to the fact that we have to impose a less restrictive prior on the coefficients of the large BVAR. Nevertheless, a BFAVAR with shrinkage remains among the best performing variants.

We believe that there is no consistent outperformance of the large BVAR. Particularly, we have to be careful when we are interested in density forecasts. As the tightness of prior information increases with the number of predictors, there seems to be an upper limit to the number of predictors where the benefits from the large BVAR are retained. Moreover, the ranking of the different variants depends *inter alia* on the target variable, the forecast horizon, the size of the dataset, and – to a lesser extent – on the state of the business cycle. Overall, we find that a factor augmented BVAR with three factors is competitive in all setups. Hence, it appears to be advisable to combine a factor augmented VAR with Bayesian shrinkage. From a practical point of view it should also be noted that we found the large BVAR being computationally much more demanding and time-consuming than the BVAR averaging or the factor augmented BVARs.

References

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.

Angelini, E., M. Bańbura, and G. Rünstler (2010). Estimating and Forecasting the Euro Area Monthly National Accounts From a Dynamic Factor Model. *Journal of Business Cycle Measurement and Analysis 2010*(1), 1–22.

Angelini, E., G. Camba-Mendez, D. Giannone, L. Reichlin, and G. Rünstler (2011). Short-Term Forecasts of Euro Area GDP Growth. *Econometrics Journal 14*(1), 25–44.

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 (2010a). Large Bayesian Vector Autoregressions. *Journal of Applied Econometrics 25*(1), 71–92.

Bańbura, M., D. Giannone, and L. Reichlin (2010b). Nowcasting. CEPR Discussion Paper No. 7883.

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.

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.

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.

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

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

Giannone, D., M. Lenza, and G. E. Primiceri (2012). Prior Selection for Vector Autoregressions. NBER Working Paper 18467.

Giannone, D., L. Reichlin, and D. Small (2008). Nowcasting: The Real-Time Informational Content of Macroeconomic Data. *Journal of Monetary Economics 55*(4), 665–676.

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 (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*, forthcoming.

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.

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

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 Paper No. 11467.

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.

Wallis, K. F. (2005). Combining Density and Interval Forecasts: A Modest Proposal. *Oxford Bulletin of Economics and Statistics 67*(s1), 983–994.

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

A Description of the Dataset

The dataset comprises 44 quarterly euro area macroeconomic and financial time series covering the period 1975:1 to 2009:4. 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; DS = Datastream), and series transformation. The last column indicates if a series is included in the medium-size dataset.

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

B Testing for Equal Predictive Ability of Point Forecasts

Table B.1: Test of Equal Predictive Ability

		Random Walk		BVAR Averaging		BFAVAR-1F		BFAVAR-3F	
Benchmark	h	GDP	HICP	GDP	HICP	GDP	HICP	GDP	HICP
BVAR Averaging	1	8.10	**-45.25**						
BFAVAR-1F		-4.87	**-44.24**	-12.00	1.84				
BFAVAR-3F		-9.83	**-52.81**	**-16.58**	-13.81	-5.20	-15.37		
Large BVAR		-21.47	**-51.46**	**-27.36**	-11.35	-17.45	-12.95	-12.92	2.85
BVAR Averaging	2	14.97	**-36.75**						
BFAVAR-1F		8.95	**-39.70**	-5.23	-4.67				
BFAVAR-3F		9.43	**-42.74**	-4.81	-9.48	0.44	-5.04		
Large BVAR		-8.47	**-46.48**	-20.39	**-15.38**	-15.99	-11.24	-16.36	-6.52
BVAR Averaging	3	24.40	-30.15						
BFAVAR-1F		16.51	**-34.35**	-6.34	**-6.02**				
BFAVAR-3F		17.82	**-34.29**	-5.29	**-5.93**	1.13	0.09		
Large BVAR		3.37	**-43.87**	-16.91	-19.65	-11.28	-14.50	-12.27	-14.58
BVAR Averaging	4	25.80	-20.62						
BFAVAR-1F		17.18	**-23.62**	-6.85	**-3.78**				
BFAVAR-3F		28.30	-23.93	1.99	-4.16	9.49	-0.40		
Large BVAR		8.56	-37.27	-13.70	-20.98	**-7.36**	**-17.87**	-15.38	-17.54

Notes: the entries are differences in RMSE (as percentage of the respective BVAR variant). A positive (negative) value means that the BVAR variant under consideration outperforms (underperforms) its benchmark variant in terms of RMSE. Differences in forecast ability are tested with a Giacomini and White (2006) test of equal conditional predictive ability. The test is based on squared forecast error loss. Bold entries denote that the forecast ability of both approaches is significantly different at the 5 percent level.

C Additional Tables for Medium-Size Dataset

In this section we document the results for the forecast experiment which is based on a subset of 22 indicator variables.

Table C.1: Root Mean Squared Errors

	Random Walk		BVAR Averaging		BFAVAR-1F		BFAVAR-3F		Large BVAR	
h	GDP	HICP	GDP	HICP	GDP	HICP	GDP	HICP	GDP	HICP
1	0.63	0.55	0.69	0.30	0.59	0.30	0.53	0.29	0.49	0.26
2	0.64	0.57	0.74	0.36	0.71	0.33	0.66	0.34	0.58	0.29
3	0.65	0.58	0.80	0.41	0.77	0.36	0.71	0.38	0.66	0.32
4	0.65	0.58	0.81	0.46	0.76	0.43	0.77	0.43	0.70	0.35
λ	0		∞		∞		0.880		0.081	

Notes: this table shows the RMSE for different BVAR variants. See also notes to Table 1.

Table C.2: Average Log Predictive Densities

	Random Walk		BVAR Averaging		BFAVAR-1F		BFAVAR-3F		Large BVAR	
h	GDP	HICP	GDP	HICP	GDP	HICP	GDP	HICP	GDP	HICP
1	−2.30	−2.29	−1.83	−1.98	−1.92	−2.11	−1.73	−2.02	−1.87	−2.03
2	−2.64	−2.42	−1.82	−1.99	−2.18	−1.97	−1.93	−1.99	−2.13	−2.04
3	−2.75	−2.59	−1.71	−1.87	−2.08	−1.88	−2.21	−2.03	−2.39	−2.05
4	−2.99	−2.59	−1.68	−1.90	−2.04	−1.95	−2.54	−2.09	−2.50	−2.16
λ	0		∞		∞		0.880		0.081	

Notes: this table shows the average log predictive densities for different BVAR variants. See also notes to Tables 1 and 2.

Table C.3: Unweighted Likelihood Ratio Tests

		Random Walk		BVAR Averaging		BFAVAR-1F		BFAVAR-3F	
Benchmark	h	GDP	HICP	GDP	HICP	GDP	HICP	GDP	HICP
BVAR Averaging	1	−0.46	−0.31						
BFAVAR-1F		−0.38	−0.18	0.09	0.13				
BFAVAR-3F		−0.57	−0.27	−0.11	0.04	−0.19	−0.09		
Large BVAR		−0.43	−0.25	0.03	0.05	−0.05	−0.08	0.14	0.01
BVAR Averaging	2	−0.82	−0.44						
BFAVAR-1F		−0.46	**-0.45**	0.36	−0.02				
BFAVAR-3F		−0.71	**-0.43**	0.11	0.01	−0.25	0.02		
Large BVAR		−0.51	**-0.38**	0.32	0.05	−0.05	0.07	0.21	0.05
BVAR Averaging	3	−1.03	**-0.72**						
BFAVAR-1F		−0.67	**-0.71**	0.37	0.01				
BFAVAR-3F		−0.54	**-0.55**	0.49	0.16	0.13	0.15		
Large BVAR		−0.36	**-0.54**	0.67	0.18	0.31	0.17	0.18	0.02
BVAR Averaging	4	−1.31	**-0.69**						
BFAVAR-1F		−0.94	**-0.64**	0.37	0.05				
BFAVAR-3F		−0.44	**-0.50**	0.87	0.19	0.50	0.14		
Large BVAR		−0.49	−0.43	1.55	0.83	0.46	0.21	−0.04	0.07

Notes: the entries are unweighted averages of likelihood ratios as in Amisano and Giacomini (2007). A positive (negative) value means that the BVAR variant under consideration outperforms (underperforms) its benchmark variant in terms of predictive ability. Bold entries denote significance at the 5 percent level. See also notes to Table 1.

Table C.4: Center-Weighted Likelihood Ratio Tests

		Random Walk		BVAR Averaging		BFAVAR-1F		BFAVAR-3F	
Benchmark	h	GDP	HICP	GDP	HICP	GDP	HICP	GDP	HICP
BVAR Averaging	1	**0.09**	**0.11**						
BFAVAR-1F		**0.11**	**0.15**	0.02	0.05				
BFAVAR-3F		0.04	**0.11**	**-0.05**	0.01	**-0.07**	−0.05		
Large BVAR		0.02	**0.07**	0.03	0.05	**-0.09**	−0.09	−0.02	−0.04
BVAR Averaging	2	**0.08**	0.10						
BFAVAR-1F		**0.09**	**0.09**	0.01	−0.01				
BFAVAR-3F		0.03	**0.07**	**-0.05**	−0.03	**-0.06**	−0.01		
Large BVAR		0.03	**0.06**	0.32	0.05	**-0.06**	−0.03	0.00	−0.01
BVAR Averaging	3	**0.06**	0.04						
BFAVAR-1F		**0.07**	0.06	0.01	0.02				
BFAVAR-3F		0.03	**0.08**	−0.02	0.04	−0.04	0.02		
Large BVAR		**0.05**	**0.04**	0.67	0.18	−0.02	−0.02	0.02	**-0.05**
BVAR Averaging	4	0.03	**0.07**						
BFAVAR-1F		0.04	**0.09**	0.02	0.02				
BFAVAR-3F		**0.06**	**0.10**	0.03	0.03	0.01	0.01		
Large BVAR		**0.05**	**0.07**	1.55	0.83	0.00	−0.02	−0.01	−0.03

Notes: the entries are unweighted averages of likelihood ratios as in Amisano and Giacomini (2007). A positive (negative) value means that the BVAR variant under consideration outperforms (underperforms) its benchmark variant in terms of predictive ability. Bold entries denote significance at the 5 percent level. See also notes to Table 1.

Table C.5: Tails-Weighted Likelihood Ratio Tests

		Random Walk		BVAR Averaging		BFAVAR-1F		BFAVAR-3F	
Benchmark	h	GDP	HICP	GDP	HICP	GDP	HICP	GDP	HICP
BVAR Averaging	1	-0.70	**-0.57**						
BFAVAR-1F		**-0.65**	**-0.56**	0.05	0.01				
BFAVAR-3F		-0.67	**-0.54**	0.02	0.03	-0.02	0.02		
Large BVAR		-0.48	**-0.43**	0.22	0.15	0.17	0.14	0.20	0.12
BVAR Averaging	2	-1.03	**-0.68**						
BFAVAR-1F		**-0.68**	**-0.67**	0.35	0.02				
BFAVAR-3F		-0.79	**-0.61**	0.24	0.07	-0.11	0.05		
Large BVAR		-0.59	**-0.53**	**0.44**	0.15	0.09	0.13	0.20	0.08
BVAR Averaging	3	-1.18	**-0.82**						
BFAVAR-1F		**-0.84**	**-0.85**	0.34	-0.03				
BFAVAR-3F		**-0.62**	**-0.76**	0.56	0.07	0.22	0.10		
Large BVAR		-0.48	**-0.62**	0.70	0.20	0.36	0.23	0.14	0.13
BVAR Averaging	4	-1.39	**-0.86**						
BFAVAR-1F		-1.06	**-0.86**	0.33	0.00				
BFAVAR-3F		-0.59	**-0.75**	0.80	0.11	0.47	0.11		
Large BVAR		**-0.61**	**-0.60**	0.78	0.26	0.45	**0.26**	-0.02	0.15

Notes: the entries are unweighted averages of likelihood ratios as in Amisano and Giacomini (2007). A positive (negative) value means that the BVAR variant under consideration outperforms (underperforms) its benchmark variant in terms of predictive ability. Bold entries denote significance at the 5 percent level. See also notes to Table 1.