International Journal of Forecasting 30 (2014) 662–682

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Evaluating predictive densities of US output growth and inflation in a large macroeconomic data set

Barbara Rossi [a,*], Tatevik Sekhposyan [b]

[a] ICREA-UPF, Barcelona GSE, and CREI, Carrer Ramon Trias Fargas 25-27, Mercè Rodoreda bldg., 08005 Barcelona, Spain

[b] Bank of Canada, 234 Wellington Street, Ottawa, ON, K1A 0G9, Canada

ARTICLE INFO

Keywords:
Predictive density evaluation
Structural change
Output growth forecasts
Inflation forecasts

ABSTRACT

We evaluate conditional predictive densities for US output growth and inflation using a number of commonly-used forecasting models that rely on large numbers of macroeconomic predictors. More specifically, we evaluate how well conditional predictive densities based on the commonly-used normality assumption fit actual realizations out-of-sample. Our focus on predictive densities acknowledges the possibility that, although some predictors can cause point forecasts to either improve or deteriorate, they might have the opposite effect on higher moments. We find that normality is rejected for most models in some dimension according to at least one of the tests we use. Interestingly, however, combinations of predictive densities appear to be approximated correctly by a normal density: the simple, equal average when predicting output growth, and the Bayesian model average when predicting inflation.

1. Introduction

Forecasts are traditionally used to evaluate models' performances. In most cases, forecasts are judged as good or otherwise based mainly on the models' (median or mean) point forecasts. For example, Stock and Watson (2003) conducted an extensive evaluation of a large data set of predictors of US output growth and inflation, focusing on point forecasts; while Banerjee and Marcellino (2006), Banerjee, Marcellino, and Masten (2005) and Marcellino, Stock, and Watson (2003) conducted similarly broad analyses for the Euro area. Furthermore, Rossi and Sekhposyan (2010) investigated the stability of point forecasts of output growth and inflation using the same data set. However, it is becoming more and more important to determine the correct specification of the uncertainty around models' point forecasts. For example, central banks are increasingly concerned about the uncertainty around their point forecasts of inflation or unemployment targets, and in particular, how well models perform in forecasting a range of future values of important macroeconomic variables.

In this paper, we consider models that have been used extensively in the literature for forecasting output growth and inflation (and seemingly doing a good job, according to their point forecasts), and investigate whether their predictive densities are calibrated correctly by the commonly-used normal approximation (see Stock & Watson, 2002). We use the Probability Integral Transform (PIT) technique, which was originally introduced by Rosenblatt (1952), and more recently has been proposed by Diebold, Gunther, and Tay (1998) for evaluating the correct specifications of predictive densities. Corradi and Swanson (2006) provide a comprehensive recent overview of tests for predictive density evaluation, and Garratt, Lee, Pesaran, and Shin (2003) and Granger and Pesaran (2000) complement the discussion further. The differences between this paper and those in the previous literature are the fact that we operate in a data-rich environment using the extensive data set of Stock and Watson (2003), and the wide range of evaluation techniques we use.

The empirical results of this paper are based on several model specifications. Regarding the models, we consider

* Corresponding author.
E-mail addresses: barbara.rossi@upf.edu (B. Rossi), tsekhposyan@bankofcanada.ca (T. Sekhposyan).

0169-2070/$ – see front matter © 2013 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ijforecast.2013.03.005

not only predictive densities based on autoregressive distributed lag (ADL) models with several predictors considered one at a time (as did Stock & Watson, 2003), but also forecast combinations. We include predictive density combinations with either equal weights or weights which are equal to the posterior probabilities of the models. In addition, we also consider several different estimation techniques: we combine models estimated by OLS with those from Bayesian shrinkage methods and a posterior simulator algorithm that samples models from the model space with the highest posterior probability. Finally, we use methods that pool the information in various series at the estimation stage, as opposed to combining them ex-post; i.e., factor models as well as Bayesian VARs.

We determine the correct specification of predictive densities using several tests. The tests we consider include tests of uniformity, serial correlation and identical distribution. Among the PIT-based tests of uniformity, we consider the histogram-based evaluation technique employed by Diebold et al. (1998) and Diebold, Tay, and Wallis (1999), as well as the Kolmogorov–Smirnov and Anderson–Darling tests. We also consider tests based on the inverse normal transformation of the PIT, including the Berkowitz (2001) and Doornik and Hansen (2008) tests. Regarding tests for independence, we consider the Ljung–Box test and a version of Berkowitz's (2001) test for the absence of serial correlation (in the PITs).[1] Finally, regarding tests of identical distribution, we consider Andrews' (1993) test of stability applied to the PITs.

Our main empirical findings can be summarized as follows. Overall, the performances of ADL models across the various tests depend crucially on the predictor included in the model. The most interesting result is that pooled predictive densities based on simple averaging as well as Bayesian Model Averaging (BMA) appear to be fairly well calibrated, particularly the simple model average for one-year-ahead output growth forecasts and the BMA for one-quarter-ahead inflation forecasts. Most of the other models that pool information at either the estimation or prediction stage report occasional failings in the correct specification of predictive densities, according to at least one of the tests we consider. Interestingly, the fact that a simple average of several parsimonious ADL models and the BMA has desirable properties in terms of forecasting is a point that has been emphasized many times in the literature in the context of point forecasts (see e.g. Stock & Watson, 2003; Timmermann, 2006; Wright, 2009). When testing the appropriateness of the normal distribution, we find that this also extends to density forecasts.

In more detail, based on both the Kolmogorov–Smirnov and Anderson–Darling tests, we find more pervasive evidence against uniformity for the predictive densities of inflation relative to output growth, at both short and medium horizons. Similar results hold when assessing the proper calibration of predictive densities in terms of independence: there is more evidence of serial correlation in the PITs of inflation relative to output growth, particularly in the second moment of the PITs. However, there is more evidence of correlation in the PITs of one-quarter-ahead density forecasts than in one-year-ahead ones. The tests also find some evidence of instabilities in the density forecasts over time, especially at the one-year-ahead horizon; in general, such instabilities are more pronounced for output growth than for inflation. Berkowitz's (2001) test confirms the results of no serial correlation in the first moments of the PITs, yet rejects uniformity in a wide set of models of output growth and inflation, particularly at short horizons. However, the normality of the simple average model for output growth and the BMA for inflation is not rejected. This is a result that holds in general based on variety of tests except the Doornik and Hansen's (2008) test. Doornik and Hansen's (2008) test rejects the proper calibration of simple average densities based on non-zero higher (third and fourth) moments of the PITs at the one-quarter-ahead horizon for output growth; it also rejects for the BMA model at the one-year-ahead horizon for inflation.

Overall, under the assumption of normality, the predictive densities of simple averaging and BMA models are among the best calibrated, in spite of the target variable which we consider. The occasional failings are associated mainly with the higher (greater than first) moments of the PITs when we use the simple average model to forecast inflation at the one-year-ahead forecast horizon, and with a lack of uniformity of the PITs at the one-quarter-ahead forecast horizon. Similarly, the BMA also performs fairly well for output growth, although it fails uniformity for the one-quarter-ahead forecast horizon, and stability for one-year-ahead.

An analysis which is similar in spirit to the one considered in this paper is that of Clements and Smith (2000). However, there are several differences between our work and theirs. First, they focus only on forecasting output growth and unemployment, and do not consider forecasts of inflation, which is another important variable for which we are interested in the predictive density. Furthermore, unlike our paper, they do not consider a large data set of macroeconomic predictors, nor a large selection of models, and focus instead on linear and non-linear univariate models and vector autoregressions with selected predictors. Finally, their paper (like most papers that evaluate density forecasts, starting from Diebold et al., 1998) focuses on testing the uniformity and uncorrelatedness of the PITs, whereas we also formally test the hypothesis of identical distributions over time.

Our paper is also related to the study by Clark (2011), who, however, focused on evaluating density forecasts from BVARs, whereas we also focus on the linear models and use a rich data set of predictors considered by Stock and Watson (2003). Importantly, unlike Clark (2011), our objective is not to improve the forecasting models (which Clark accomplishes by allowing for stochastic volatility): rather, we consider models that are used extensively in the literature and test whether their density forecasts, based on the commonly used normal approximation, are correctly specified.

[1] Note that our focus throughout this paper is on testing for serial correlation in the PITs (as opposed to serial correlation in the forecasts). Serial correlation in the PITs indicates that the pattern of rejection of the correct specification is not random over time, and may signal misspecification in the dynamics of the underlying models.