International Journal of Forecasting 31 (2015) 113–129

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Testing the value of probability forecasts for calibrated combining

Kajal Lahiri *, Huaming Peng, Yongchen Zhao

Department of Economics, University at Albany, SUNY, Albany, NY, 12222, USA

ARTICLE INFO

Keywords:
SPF recession forecasts
Welch-type tests
Autocorrelation and skewness corrections
Beta-transfromed pool

ABSTRACT

We combine the probability forecasts of a real GDP decline from the US Survey of Professional Forecasters, after trimming the forecasts that do not have "value", as measured by the Kuiper Skill Score and in the sense of Merton (1981). For this purpose, we use a simple test to evaluate the probability forecasts. The proposed test does not require the probabilities to be converted to binary forecasts before testing, and it accommodates serial correlation and skewness in the forecasts. We find that the number of forecasters making valuable forecasts decreases sharply as the horizon increases. The beta-transformed linear pool combination scheme, based on the valuable individual forecasts, is shown to outperform the simple average for all horizons on a number of performance measures, including calibration and sharpness. The test helps to identify the good forecasters ex ante, and therefore contributes to the accuracy of the combined forecasts.

1. Introduction

Currently, more than forty-five years of quarterly expert forecasts for a number of US macroeconomic variables are available in the US Survey of Professional Forecasters (SPF).[1] In particular, the SPF probabilities of real GDP declines for the current quarter and the next four quarters have attracted special interest from economists and the mass media alike. The New York Times columnist David Leonhardt (September 1, 2002) called the one-quarter-ahead GDP decline probability the "Anxious Index". Drawing on methodologies developed in statistics, the atmospheric sciences, and psychology, economists have studied the quality and characteristics of these subjective probabilities, and have reached a certain broad consensus.[2] First, a simple average of the individual probabilities seems to encompass all of the information embedded in the individual forecasts (Clements & Harvey, 2011); and second, these average forecast probabilities do not seem to have any predictive power beyond the second quarter (Lahiri & Wang, 2013). Many researchers have noted the limitations of these average probabilities in forecasting economic downturns. Stock and Watson (2003) point out that the SPF could not foresee the 2001 recession: the signal came with a lag in 2001:Q4, when the negative growth period had already passed. Rudebusch and Williams (2009) show that the SPF participants do not seem to use the information in the yield spread (i.e., the difference between long-term and short-term interest rates) in forecasting recessions, despite it having been well-known since at least the 1980s

* Corresponding author.
E-mail addresses: klahiri@albany.edu (K. Lahiri), hpeng@albany.edu (H. Peng), hzhao@albany.edu (Y. Zhao).

[1] The European Central Bank, the Banco de Mexico and the Reserve Bank of India have also been collecting similar forecasts under the same name. Thus, any methodological improvement in the use of the US data will be useful for a number of countries as well.

[2] See for instance Braun and Yaniv (1992), Clements (2008, 2009, 2011), Clements and Harvey (2011), Engelberg, Manski, and Williams (2010), Galbraith and van Norden (2011, 2012), Graham (1996), and Lahiri and Wang (2006, 2013).

http://dx.doi.org/10.1016/j.ijforecast.2014.03.005
0169-2070/

that the yield spread is useful in forecasting real GDP growth and recessions. However, Lahiri, Monokroussos, and Zhao (2013) find that, when averaged over the relatively better forecasters, the combined SPF forecasts do incorporate the information from the yield spread, as well as that from a myriad of other economic indicators. Galbraith and van Norden (2012) use the pioneering "fan charts" of the Bank of England to calculate the probability that the annual rates of inflation and output growth will exceed given thresholds. They find a serious departure of these aggregate forecasts from perfect calibration and reasonable sharpness.

The main motivation of this paper comes from a recent finding by Ranjan and Gneiting (2010) that the linear opinion pool is sub-optimal in terms of calibration and sharpness. Following their suggestion, we use the nonlinearly recalibrated beta-transformed linear pool (BLP) to determine whether the performance of aggregated SPF forecasts can be improved. As the fallout from the recent recession of 2007–2009 has painfully reminded us again, even a small improvement in our capability to foresee recessions is of enormous benefit to any modern society. Our second motivation is derived from the novel methodological approach to forecast combination by Aiolfi and Timmermann (2006). Following their approach, and due to the huge amount of missing data in the SPF, we first sort all of the forecasters into four clusters based on their past performances. Then, we pool the forecasts within each cluster, and follow this with an application of the BLP methodology to the cluster-level aggregates, in order to obtain an improved combined forecast. However, before constructing the clusters, we trim the forecasters whose forecasts are found to have no value in the sense of the Kuiper Skill Score (KSS), which is a sample analog of the concept of "value" in Merton (1981). Stekler (1994) initiated a very important line of research in this area of forecast evaluation, see also Schnader and Stekler (1990). For assessing the "value" of the forecasts, we propose to use a Welch-type test[3] that explicitly addresses the potential issue of serial correlation and skewness.

The plan of the paper is as follows. In Section 2, we describe the nature and characteristics of the SPF probability forecasts, which naturally leads on to the methodological approach undertaken in this paper. In Section 3, we describe the test for zero forecast value and discuss its robustness to serial correlation and skewness. Section 4 reports our empirical results, including the test results, the beta-transformed linear pool methodology, the combination of the valuable forecasts, and a detailed evaluation of the combined forecasts in- and out-of-sample. Section 5 concludes.

[3] The Welch-type t test is a modified Student's t-test which is designed to test the equality of two means with possibly unequal variances. That is, $t_\upsilon = \frac{\widehat{\mu}_1 - \widehat{\mu}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$, where $\widehat{\mu}_i, s_i^2$ and N_i are the sample mean, sample variance and sample size respectively for $i = 1, 2$, and the degrees of freedom υ is given approximately by $\upsilon \approx \left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2 / \left(\frac{s_1^4}{N_1^2(N_1-1)} + \frac{s_2^4}{N_2^2(N_2-1)}\right)$.

2. Probability forecasts in SPF

We use the probability forecasts of real GDP declines from the US Survey of Professional Forecasters (SPF). The SPF is the oldest quarterly survey of macroeconomic forecasts in the United States.[4] Since 1990Q2, the survey has been being administered by the Federal Reserve Bank of Philadelphia. Survey respondents are asked to supply point and density forecasts for a wide range of variables covering output, prices, and employment situations. The survey is used widely by researchers. Examples of recent work using data from the survey include the studies by Capistran and Timmermann (2009), Clements (2011), and Lahiri and Wang (2013). We examine the probability forecasts of real GDP declines for the current quarter (denoted $h = 0$)[5] and the following four quarters ($h = 1, \ldots, 4$). The equally-weighted average of individual forecasts, i.e., the equally-weighted linear opinion pool (ELP), is used as a benchmark when evaluating the performance of the combined forecasts we construct. Our data spans the 174 quarters from 1968Q4 to 2012Q1. A total of 32,379 forecasts are available from 426 forecasters for 5 horizons.

Frequent data revisions and definitional changes affect real GDP values after their initial release. In practice, such changes are rarely predictable before they are announced. The choice of actual values between an earlier data vintage and the latest vintage depends crucially on the objective function of a forecasting client. We construct the binary outcome series (the actual values) y_t using the first vintage in the Real-Time Data Set for Macroeconomists available from the Federal Reserve Bank of Philadelphia.[6] For a discussion of the real time data issues involved in evaluating SPF forecasts, see Stark (2010). The same approach is taken by, among others, Clements and Harvey (2010). Our binary outcome series contains 24 quarters of real GDP declines, which is about 14% of the sample. Since a real GDP decline is a relatively uncommon event, special considerations are needed in forecast evaluation. Examples of recent work focusing on evaluating probability forecasts include the studies by Galbraith and van Norden (2012) and Lahiri and Wang (2013).

The SPF contains a large amount of missing data, see Capistran and Timmermann (2009) and Lahiri, Peng, and Zhao (2012). In our case, more than 91% of the data is missing.[7] To make the situation more complicated, after the Federal Reserve Bank took over the survey, most of the old forecasters stopped forecasting and many new forecasters joined the sample. As a result, for most of the

[4] The survey was introduced in 1968 by the American Statistical Association and the National Bureau of Economic Research. For more information and background about the survey, see Croushore (1993). The survey itself can be accessed from http://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/.

[5] The "current quarter" refers to the quarter in which the survey is conducted.

[6] A robustness check was conducted using the latest vintage, and our main conclusions remained the same. Results from the robustness check are available from the authors.

[7] A fully balanced panel with all forecasters who have ever participated and all quarters would have 426 forecasters × 174 quarters × 5 horizons = 370,620 forecasts; but we have only 32,379 forecasts.