Economic Modelling 67 (2017) 294–299

Contents lists available at ScienceDirect

Economic Modelling

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

Dissecting models' forecasting performance☆

Boriss Siliverstovs

KOF Swiss Economic Institute, ETH Zurich, Switzerland

ARTICLE INFO

JEL classification:
C22
C53

Keywords:
Forecasting
Forecast evaluation
Change point detection
Bayesian estimation

ABSTRACT

The fact that the predictive performance of models used in forecasting stock returns, exchange rates, and macroeconomic variables is not stable and varies over time has been widely documented in the forecasting literature. Under these circumstances excessive reliance on forecast evaluation metrics that ignores this instability in forecasting accuracy, like squared errors averaged over the whole forecast evaluation sample, masks important information regarding the temporal evolution of relative forecasting performance of competing models. In this paper we suggest an approach based on the combination of the Cumulated Sum of Squared Forecast Error Differential (CSSFED) of Welch and Goyal (2008) and the Bayesian change point analysis of Barry and Hartigan (1993) that tracks the contribution of forecast errors to the aggregate measures of forecast accuracy observation by observation. In doing so, it allows one to track the evolution of the relative forecasting performance over time. We illustrate the suggested approach by using forecasts of the GDP growth rate in Switzerland.

1. Introduction

One of the stylised facts in the forecasting literature is that the predictive content of models used in out-of-sample forecasting varies over time. In other words, it is rather the rule than the exception that (ex post) one can identify periods when a more sophisticated model beats a naïve benchmark model in terms of forecast accuracy. There are also periods when both models produce similar forecast accuracy. One may also identify periods when the benchmark model produces more accurate out-of-sample predictions than the more sophisticated model. In such cases one observes a so-called reversal in the relative forecasting performance of the competing models, which if unnoticed may lead to erroneous reliance on the more sophisticated model producing less accurate forecasts.

The instability in the predictive ability was recorded for models forecasting stock returns (Welch and Goyal, 2008; Ang and Bekaert, 2007; Paye and Timmermann, 2006), exchange rate models (Rossi, 2013; Giacomini and Rossi, 2010; Rogoff and Stavrakeva, 2008; Schinasi and Swamy, 1989), and models developed to predict macroeconomic variables like GDP growth or inflation (Schrimpf and Wang, 2010; Stock and Watson, 2007; Giacomini and Rossi, 2006, *inter alia*).

As discussed in Giacomini and Rossi (2010), in the presence of instability in the predictive content of competing models, the usual tests addressing the null hypothesis of equal predictive ability (Clark and West, 2007; Giacomini and White, 2006; Clark and McCracken, 2001; West, 1996; Diebold and Mariano, 1995, *inter alia*) may be misleading. The reason is that these tests are based on comparison of the average forecasting performance over the whole forecast evaluation sample and therefore are not informative regarding time variation in the relative forecasting performance of the models. This means that these tests are not suitable for detecting situations when the initially best forecasting model turns out to be the worst one, i.e. a reversal in the relative forecasting performance takes place. This fact can have far reaching consequences. A failure to detect a reversal in the relative forecasting performance of the models, for instance, may lead to erroneous conclusions regarding the ranking of the models and their relative importance for policy making or investment decisions.

A further concern of ours is that a focus on the global forecasting performance will also give a biased view in situations when a few but large forecast errors are mainly accountable for the difference in the reported forecast accuracy measures between competing models. In this respect, it is important to be able to distinguish between situations when one model consistently produces more accurate forecasts than its competitor and situations when one model occasionally produces much more accurate forecasts than its competitor, but most of the time their relative predictive content is about the same.

☆ The author acknowledges helpful comments from participants at the 33rd CIRET Conference in Copenhagen (Denmark), the ITISE 2016 in Granada (Spain), the 2nd HenU/INFER Workshop on Applied Macroeconomics in Kaifeng (China), the ESOBE 2015 meeting in Gerzensee (Switzerland), and the Annual DIW Berlin Macroeconometric Workshop in Berlin (Germany) as well as from Niklas Ahlgren, Daniel Wochner, the journal's editor, and two anonymous referees. Computations and graphics were produced using the R language (http://cran.r-project.org/). The usual disclaimer applies.

E-mail address: siliverstovs@kof.ethz.ch.

http://dx.doi.org/10.1016/j.econmod.2017.01.008
Received 23 May 2016; Received in revised form 14 December 2016; Accepted 12 January 2017
Available online 23 January 2017

The concern about the possible loss of information by focusing on the global forecasting performance is addressed in Giacomini and Rossi (2010), where two statistical tests specifically focusing on the local forecasting performance are proposed: the *fluctuation* and the *one-time reversal* tests. The *fluctuation* test addresses the question of equal predictive ability while allowing for time variation in the relative forecasting performance. However, since the fluctuation test is applied over a rolling window of a fixed size, it essentially represents a version of tests comparing global forecasting performance, though applied on a more localised time scale. Therefore, it is prone to similar problems as its global counterparts. Moreover, in smaller sub-samples the effect of large forecast errors is even more exacerbated since the assessment window is only a fraction of the whole sample. The *one-time reversal* test is designed to estimate the timing when a single reversal in forecasting performance takes place. Its use is therefore limited in very unstable environments characterised by multiple reversals in models' predictive ability.

Our paper contributes to the literature in the following way. We suggest a procedure that intends to facilitate tracking how the relative forecasting performance of competing models evolves over time. Rather than ranking the models in terms of aggregate measures of their forecasting performance computed, for instance, as averages of squared forecast errors over the whole forecast evaluation period, we suggest to dissect the models' relative forecasting performance observation-wise, that is by scrutinising differences in model-specific forecast errors observation by observation. This observation-wise approach allows us to detect multiple changes and structural breaks in the relative forecasting performance of competing models as well as to single out periods when differences in the forecasting performance over-proportionally contribute to aggregate measures of forecast accuracy like Mean Squared Forecast Errors, for example. Our procedure is based on an assessment of the models' relative forecasting performance based on the Cumulated Sum of Squared Forecast Error Differential (CSSFED) suggested in Welch and Goyal (2008), in combination with the sample partition algorithm suggested in Barry and Hartigan (1993), to which we refer as BH henceforth.

The rest of the paper is structured as follows. In Section 2 the outline of the econometric methodology is presented. A description of the data is provided in Section 3. Section 4 illustrates the suggested approach using GDP forecasts for Switzerland. The final section concludes.

2. Econometric methodology

2.1. Cumulated sum of squared forecast error differential

Welch and Goyal (2008) introduce the Cumulated Sum of Squared Forecast Error Differential (CCSFED) as a helpful graphical tool allowing to monitor the evolution of the relative forecasting performance of equity premium regressions with respect to forecasts from a benchmark model based on the historical mean. This simple suggestion turned out to be a very powerful tool such that it is commonly used in the equity premium (Rapach et al., 2013) as well as in the commodity prices (Buncic and Moretto, 2015) forecasting literature. At the same time its use in the macroeconomic forecasting literature still remains very limited (see e.g. (Aastveit et al., 2016; Schrimpf and Wang, 2010)).

The CSSFED is defined as the cumulated sum of squared forecast error difference between a benchmark model and its competitor:

$$CSSFED_{(\tau_0,\tau_1)} = \sum_{t=\tau_0}^{\tau_1} [(e_{ARM,t})^2 - (e_{DFM,t})^2], \quad (1)$$

where $e_{1,t}$ and $e_{2,t}$ denote forecast errors from a benchmark and more sophisticated models, respectively, using the forecast evaluation period (τ_0, τ_1). Here the benchmark model is a univariate autoregression of order one, which is a rather common choice in the macroeconomic forecasting literature (see e.g. (Girardi et al., 2016; Rünstler et al., 2009)). We postpone a description of the more sophisticated model to Section 3.

Upward trending of the CSSFED reflects a tendency of the benchmark model to produce larger forecast errors than its competitor up to any given point in time. Downward trending indicates the opposite. A horizontal movement of the CSSFED implies that neither model dominates the other in terms of forecast accuracy. Positive and negative values of the CSSFED observed in the last period unequivocally indicate whether the MSFE of the benchmark model is higher or lower than that of the competing model. However, contrary to the MSFE which is a scalar variable, the CSSFED is a time series displaying the whole sample path of the relative forecasting performance. Plotting the CSSFED over time allows one to visually assess the contribution of the difference in squared forecast errors $[(e_{1,t})^2 - (e_{2,t})^2]$ to the CSSFED that we observe in *every* period in the forecast evaluation sample. In doing so, the CSSFED brings a new dimension to assessing the relative forecasting performance of competing models. In the macroeconomic forecasting literature it is still a common practice to rank competing models in terms of their Mean Square Forecast Errors (MSFE) computed over the whole forecast evaluation sample. However, as argued above this often masks important details concerning the relative forecasting performance of the models and how it evolves over time.

It is of a particular interest to determine whether the superior forecasting performance of one model relative the other model, as indicated by a lower relative MSFE, for example, is due to a continuous improvement in forecast accuracy accrued over time. This occurs when one model most of the time exhibits small but steady gains in forecast accuracy. Or the observed gains in forecasting accuracy is the result of a small number of influential observations driving the observed difference in the model-specific MSFEs. Typically, these influential observations are associated with extraordinary events like the Great Recession or, in a case specific to Switzerland, the Franc shock of January 15, 2015 when the Swiss National Bank unexpectedly lifted the exchange rate floor of 1.20 CHF/EURO introduced on September 6, 2011 (see Siliverstovs (2016), for an analysis of the effects of this event). Precisely this information is contained in graphical plots of the CSSFED. Small but steady gains in the forecasting accuracy of one model over the other results in a smooth trending behaviour in the CSSFED, whereas large gains in forecasting accuracy will be reflected in abrupt jumps in the plotted CSSFED. In the latter case, one will be able to distinguish between periods when both models exhibited similar predictive ability and periods when one model generated much more accurate forecasts than its competitor. Being able to discriminate between these two types of periods is important for a forecasting practitioner because it allows him or her to dissect models' relative forecasting performance observation by observation. For forecasting models employed in real time, the occurrence of such extraordinary events serves as a stress test of their forecasting ability. This is of particular importance during periods of economic or financial distress characterised by an increased level of uncertainty. In such times accurate forecasts are especially in high demand.

2.2. Change point detection algorithm

The main contribution of the paper is to suggest applying the change point detection algorithm of Barry and Hartigan (1993) (henceforth, BH) to the sequence of CSSFED. The advantage of this procedure is that the BH algorithm provides a probabilistic assessment of a change point at each time point in the forecasting sample.

The BH algorithm defines a partition $\tau = (U_1, U_2, \ldots, U_T)$, where an element U_t=1 indicates a boundary between two segments at t, i.e. a change point at $t + 1$. The algorithm is initialised by setting U_t=0 for all $t < T$ and U_T=1. Markov chain sampling is used in order to draw values of U_t from the conditional distribution of U_t given the data $\mathbf{X} = (x_1, x_2, \ldots, x_T)'$ and the current partition τ. As shown in Barry and