Dissecting Models' Forecasting Performance*

Boriss Siliverstovs**
KOF Swiss Economic Institute
ETH Zurich

November 11, 2015

Abstract

In this paper we suggest an approach to comparison of models' forecasting performance in unstable environments. Our approach is based on combination of the Cumulated Sum of Squared Forecast Error Differential (CSSFED) suggested earlier in Welch and Goyal (2008) and the Bayesian change point analysis based on Barry and Hartigan (1993). The latter methodology provides the formal statistical analysis of the CSSFED time series which turned out to be a powerful graphical tool for tracking how the relative forecasting performance of competing models evolves over time. We illustrate the suggested approach by using forecasts of the GDP growth rate in Switzerland.

Keywords: Forecasting, Forecast Evaluation, Change Point Detection, Bayesian Estimation
JEL code: C22, C53

*The author acknowledges helpful comments from participants at the ESOBE2015 meeting in Gerzensee, Switzerland. Computations and graphics were produced using the R language (http://cran.r-project.org/). The usual disclaimer applies.
**E-mail: siliverstovs@kof.ethz.ch

1 Introduction

A seminal contribution of Diebold and Mariano (1995) where a formal statistical procedure was proposed for testing the null hypothesis of equal predictive ability of competing models laid a cornerstone for the rapidly developing literature which compares models' relative forecasting performance both from theoretical as well as empirical angles (West, 1996; Clark and McCracken, 2001; Clark and West, 2007; Giacomini and White, 2006, *inter alia*). However, much of this literature focuses on comparing average model predictive ability over the whole forecasting sample. In case when there are instabilities in the forecasting performance of the competing models, for example, when initially the best forecasting model turns out to be eventually the worst one, such a focus on the average performance hides an important information. Thus, a failure to detect the reversal in the relative forecasting performance of the models, for instance, may lead to erroneous conclusions regarding their ranking and relative importance for policy making or investment decisions. Our further concern is that such a focus on the global forecasting performance will also give a biased view in situations when a few but large forecast errors are accountable for the difference in the reported forecast accuracy measures between the competing models. This effect is even more aggravated if comparison is made under a quadratic loss function, e.g. in terms of model-specific Mean Squared Forecast Error (MSFE), which disproportionally penalises large forecast errors.

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. A Fluctuation test addresses the question of equal predictive ability while allowing for time variation in the relative forecasting performance, and a One-Time Reversal test is designed to estimate the timing when the reversal took place. However, both tests are essentially the versions of comparing global forecasting performance though applied on a more localised scale like over the rolling windows of a fixed size in the fluctuation test and over two sub-samples around the potential reversal timing in the one-time reversal test. Therefore, these tests are prone to similar caveats as their 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.

Our paper contributes to the literature in the following way. We suggest a procedure that intends to facilitate tracking how models forecasting performance evolves over time. Rather than examining aggregate measures of the forecasting performance averaged over a certain sample period, we suggest to dissect the models' forecasting performance observation-wise, that is by scrutinising each particular forecast error. This observation-wise

approach allows us to detect multiple changes and structural breaks in the relative forecasting performance of competing models. Our procedure is based on the assessment of the models' relative forecasting performance based on the Cumulated Sum of Squared Forecast Error Differential (CSSFED) suggested earlier 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. A description of the data is provided in Section 2. In Section 3 the outline of econometric methodology is presented. Section 4 illustrates the suggested approach using GDP forecasts for Switzerland. The final section concludes.

2 Data

The forecasts used in the current paper are those from the dynamic factor model (DFM) developed in Siliverstovs and Kholodilin (2012) for Switzerland. The model is a large-scale dynamic factor model based on more than 550 economic and financial domestic indicators. Model parameters are estimated using the two-step procedure of Giannone et al. (2008).

The model was first calibrated in a simulated pseudo-real time framework using the forecast evaluation period from 2005Q1 until 2009Q2. Siliverstovs (2012) evaluates the forecasting performance of the same model in a real-time squared forecasting exercise in a more recent period that ends in 2013Q3.[1] Since 2009Q3 the DFM is used as a complimentary forecasting device to the KOF Macroeconometric model in order to generate short-term forecasts of the Swiss GDP growth rate. This allows us to extend the forecast evaluation period up to 2015Q2, such that the total forecast evaluation period is from 2005Q1 until 2015Q2. The end of our forecast evaluation sample is determined by the actual availability of the official quarterly System of National Accounts (SNA), as of the time of writing.

Our target variable is first official release of seasonally adjusted quarterly real GDP growth. The forecast origin is the beginning of the third month of each quarter when data were already released for the previous quarter. This means that our real-time forecasts or, more precisely, nowcasts precede official releases by about three months.

3 Bayesian change point analysis of CSSFED

Welch and Goyal (2008) introduces the CCSFED as a helpful graphical

[1] By the term "real-time squared" we mean that forecasts are made in genuine forecast-as-you-go manner. That is they are made in real time with real-time data vintages.

tool allowing to monitor evolution of the relative forecasting performance of equity premium regressions with respect to forecasts from a benchmark model based on a historical mean. This simple suggestion turned out to a very powerful though informal analysing tool such that its reporting is commonly used in the equity premium (Rapach et al., 2013) as well as commodity prices (Buncic, 2015) forecasting literature. At the same time its use in the macroeconomic forecasting literature still remains very limited (e.g. see Aastveit et al., 2014).

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

$$CSSFED_t = \sum_{t=1}^{T} \left[(e_{ARM,t})^2 - (e_{DFM,t})^2 \right], \tag{1}$$

where $e_{ARM,t}$ and $e_{DFM,t}$ denotes forecast errors from a benchmark and dynamic factor models. Observe that here the benchmark model is a univariate autoregression of order one which is more common in macroeconomic forecasting literature (e.g. see Gayer et al., 2014; Barhoumi et al., 2009).

Upward trending of the CSSFED reflects the tendency of the benchmark model to produce larger forecast errors than its competitor up to that point in time. Downward trending —indicates the opposite. A horizontal movement of the CSSFED implies that neither model dominates another in terms of forecast accuracy. Positive and negative values of the CSSFED observed in the last period unequivocally indicate whether the associated MSFE of the benchmark model is higher or, respectively, 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 evolution path of the relative forecasting performance. Another useful information provided by the CSSFED itself is that it allows us to verify whether the superior forecast performance of one model relative the other model is due to a continuous improvement in the forecast accuracy or a result of few influential observations, e.g. during periods of economic or financial distress like the Great Recession or, in case specific to Switzerland, the Franc shock of January 15, 2015 when the Swiss National Bank lifted the exchange rate floor of 1.20 CHF/EURO introduced on September 6, 2011. In the former case one would observed a smooth trending behaviour in CSSFED and abrupt jumps in the latter case.

As the main contribution of the paper we suggest to apply a 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 $\tau = (U_1, U_2, ..., U_T)$, 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 $U_T = 1$. The Markov chain sampling is used in order to draw values of U_t from the conditional distribution of U_t given data $\mathbf{X}$ and the current partition τ. As shown in Barry and Hartigan (1993), the transition probability of a change point p_t in a given point of time can be obtained from the following ratio:

$$\frac{p_t}{1-p_t} = \frac{P(U_t = 1|\mathbf{X}, \tau)}{P(U_t = 0|\mathbf{X}, \tau}. \tag{2}$$

Essentially, the ratio is a function of the number of blocks b for a given partition with $U_t = 0$ and two tuning parameters γ and λ that take values in the unit interval $[0,1]$. The values of γ and λ may be chosen to govern frequency and size of changes: smaller values of γ and λ result in smaller number of changes and their magnitude (Barry and Hartigan, 1993, p. 312). In the empirical application in Section 4 we set the values of the tuning parameters to their default value of 0.2 as suggested in Barry and Hartigan (1993). The Markov Chain Monte Carlo implementation of the BH algorithm delivers posterior means for each block as well as posterior probability of a break point in each period of time.[2]

4 Empirical results

The measure of forecast accuracy based on the mean squared forecast errors (MSFE) computed for the whole forecast sample is presented in Table 1. According to the reported MSFEs, the dynamic factor model produced lower forecast errors than the benchmark autoregressive model on average. The proportionate reduction in the MSFE of the benchmark model measured by the relative MSFE:

$$\text{Relative MSFE} = 100 * \frac{MSFE_{ARM} - MSFE_{DFM}}{MSFE_{ARM}} \tag{3}$$

is 48.9%. For illustration, the actual and forecast values from the ARM and DFM are shown in Figure 1.

The CSSFED together with the overlaid estimate of posterior mean determined by the change point algorithm of Barry and Hartigan (1993) is displayed in the upper panel and the corresponding posterior probability of a break point in the lower panel of Figure 2. In general, the CSSFED exhibits an upward movement characterised by tranquil periods of horizontal drift interrupted by several jumps caused by large difference in squared

[2]The BH algorithm is implemented in the R programming language in the **bcp** package (Erdman and Emerson, 2007).

Table 1: Forecast accuracy assessment: MSFE

	ARM	DFM	Relative (in%)
MSFE	0.131	0.067	48.9

forecast errors in these periods. At least in two cases, the timing of these jumps is easily recognisable. First, the ARM is too sluggish to recognise the outbreak of the Great Recession producing much larger forecast errors than the DFM in 2008Q4 and 2009Q1. The second similar episode is in 2015Q1 when the Swiss National Bank abolished the floor of the CHF/EUR exchange rate. There is a number of quarters when the DFM produces much smaller forecast errors like in 2006Q1 and 2012Q1 but it is less obvious whether this is due any specific economic event or just an artefact of autoregressive dynamics captured by the ARM.

5 Conclusion

In this paper, we suggest an approach that allow us to focus on local rather than global models' forecasting performance. Comparing the average forecasting performance based on difference in MSFEs calculated over the whole forecast evaluation sample often entails a loss of information. For example, the presence of influential observations (large forecast errors) and their exaggerated effect on difference in MSFEs may be concealed by focusing on the average forecast accuracy measures.

In this paper we suggests to combine the analysis of the relative forecasting performance based on the cumulated sum of squared forecast error difference (CSSFED) of Welch and Goyal (2008) with the Bayesian change point algorithm of Barry and Hartigan (1993). The latter procedure provides a probabilistic assessment of structural changes in the models' forecasting performance.

We provide an empirical example illustrating the use of the suggested approach by comparing forecasts of Gross Domestic Product in Switzerland produced by a large-scale dynamic factor model of Siliverstovs and Kholodilin (2012) with forecasts from the benchmark autoregressive model.

References

Aastveit, K. A., C. Foroni, and F. Ravazzolo (2014). Density forecasts with MIDAS models. Working Paper 2014/10, Norges Bank.

Barhoumi, K., S. Benk, R. Cristadoro, A. D. Reijer, A. Jakaitiene,

P. Jelonek, A. Rua, G. Rünstler, K. Ruth, and C. van Nieuwenhuyze (2009). Short-term forecasting of GDP using large datasets: A pseudo real-time forecast evaluation exercise. *Journal of Forecasting 28*(7), 595–611.

Barry, D. and J. A. Hartigan (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association 35*(3), 309–319.

Buncic, D. (2015). Forecasting copper prices with dynamic averaging and selection models. *The North American Journal of Economics and Finance 33*(1), 1 – 38.

Clark, T. E. and M. W. McCracken (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics 105*(1), 85–110.

Clark, T. E. and K. D. West (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics 138*(1), 291–311.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 253–63.

Erdman, C. and J. W. Emerson (2007). bcp: An R package for performing a Bayesian analysis of change point problems. *Journal of Statistical Software 23*(3), 1–13.

Gayer, C., A. Girardi, and A. Reuter (2014). The role of survey data in nowcasting euro area GDP growth. Directorate-General for Economic and Financial Affairs: Economic Papers 538, European Commission.

Giacomini, R. and B. Rossi (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics 25*(4), 595–620.

Giacomini, R. and H. White (2006). Tests of conditional predictive ability. *Econometrica 74*(6), 1545–1578.

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.

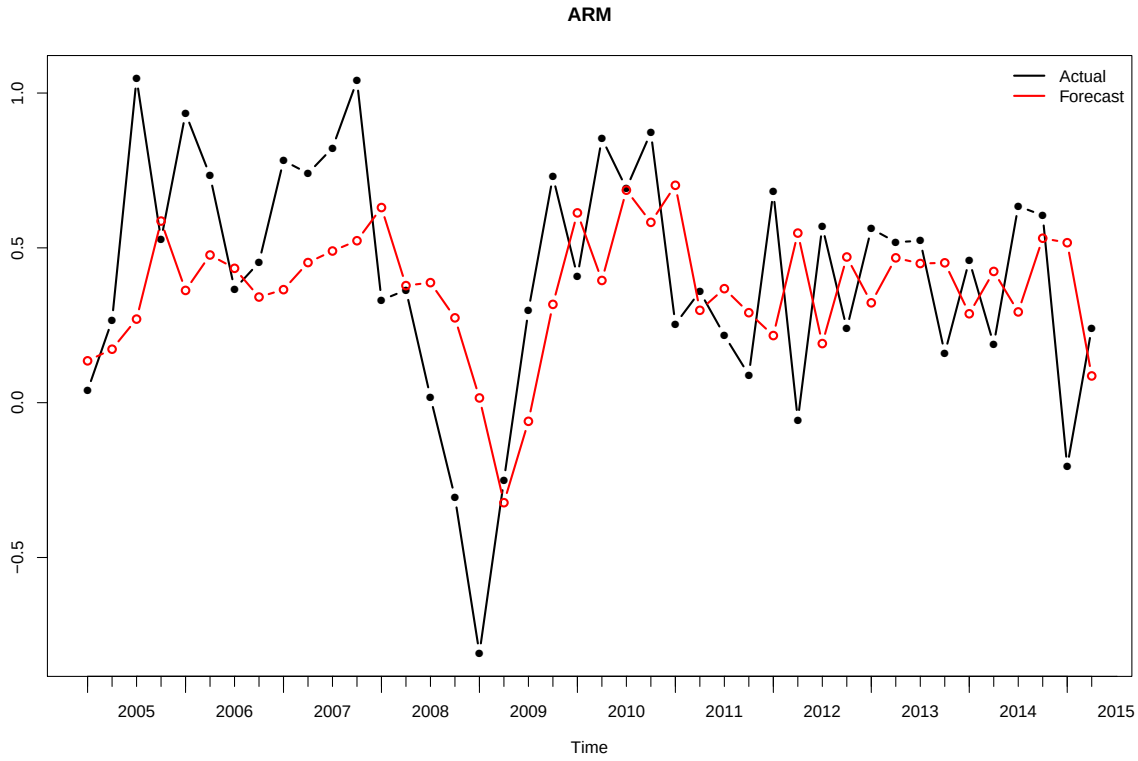
Rapach, D. E., J. K. Strauss, and G. Zhou (2013). International stock return predictability: What is the role of the United States? *The Journal of Finance 68*(4), 1633–1662.

Siliverstovs, B. (2012). Keeping a finger on the pulse of the economy: Nowcasting Swiss GDP in real-time squared. KOF Working papers 12-302, KOF Swiss Economic Institute, ETH Zurich.

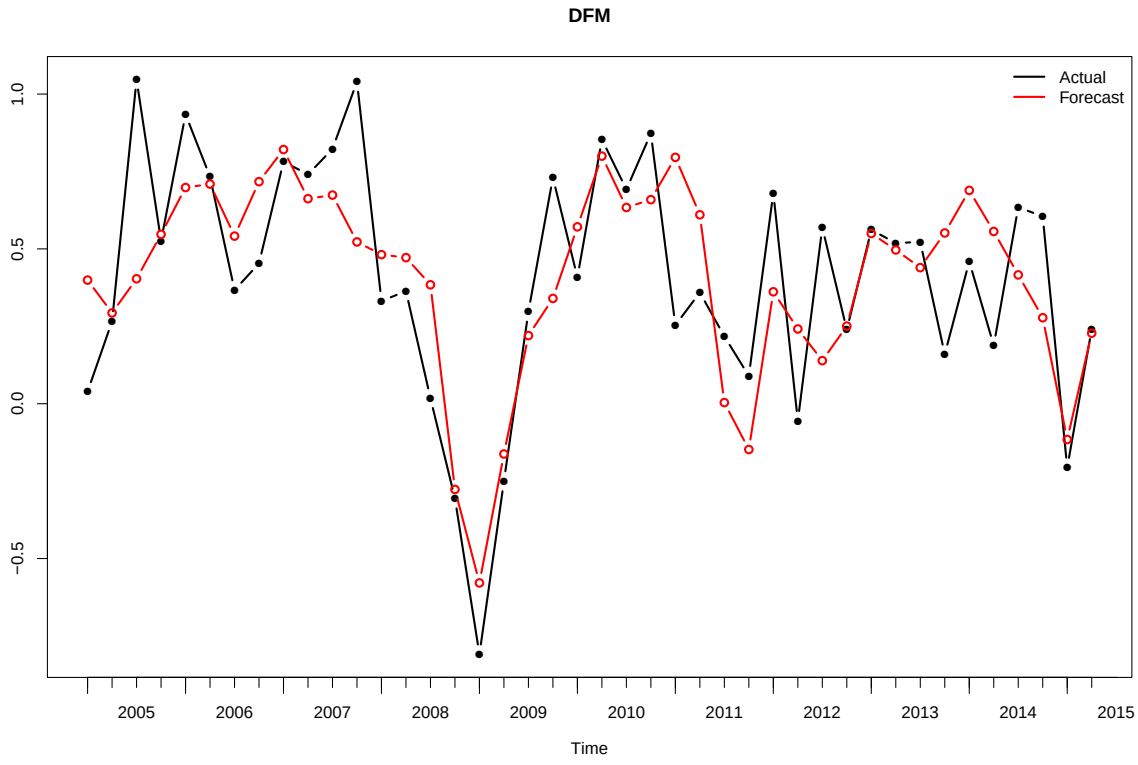
Siliverstovs, B. and K. A. Kholodilin (2012). Assessing the real-time informational content of macroeconomic data releases for now-/forecasting GDP: Evidence for Switzerland. *Jahrbücher für Nationalökonomie und Statistik 232*(4), 429–444.

Welch, I. and A. Goyal (2008). A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. *Review of Financial Studies 21*(4), 1455–1508.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica 64*(5), 1067–84.



(a) Benchmark AR(1) model



(b) Dynamic factor model of Siliverstovs and Kholodilin (2012)

Figure 1: GDP: Real quarterly growth, actual (filled circles) and forecast (empty circles) values

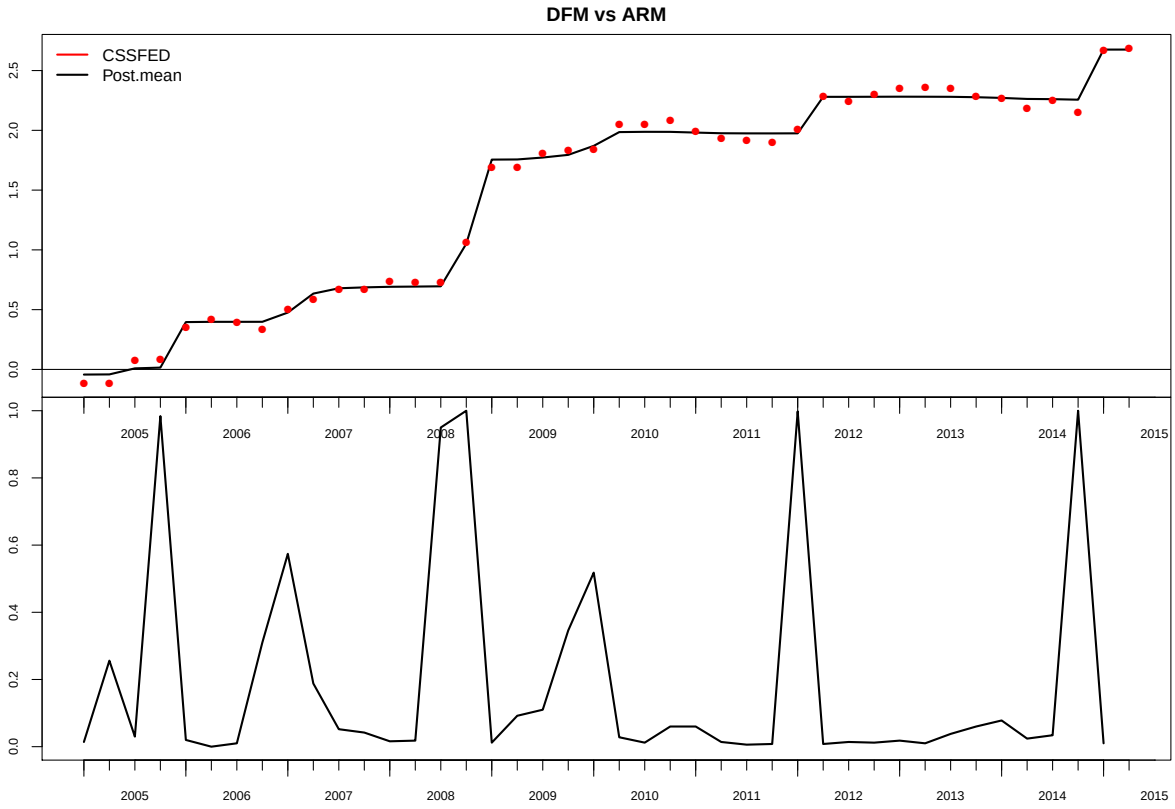


Figure 2: CSSFED