Forecasting in the Presence of Instabilities: How Do We Know Whether Models Predict Well and How to Improve Them

Barbara Rossi

This version: July 2021
(November 2019)

Barcelona GSE Working Paper Series

Working Paper nº 1162

Forecasting in the Presence of Instabilities: How Do We Know Whether Models Predict Well and How to Improve Them

Barbara Rossi*

ICREA-Univ. Pompeu Fabra,
Barcelona GSE, and CREI

This Draft: July 2021

Abstract: This article provides guidance on how to evaluate and improve the forecasting ability of models in the presence of instabilities, which are widespread in economic time series. Empirically relevant examples include predicting the financial crisis of 2007-2008, as well as, more broadly, fluctuations in asset prices, exchange rates, output growth and inflation. In the context of unstable environments, I discuss how to assess models'forecasting ability; how to robustify models'estimation; and how to correctly report measures of forecast uncertainty. Importantly, and perhaps surprisingly, breaks in models'parameters are neither necessary nor sufficient to generate time variation in models'forecasting performance: thus, one should not test for breaks in models' parameters, but rather evaluate their forecasting ability in a robust way. In addition, local measures of models' forecasting performance are more appropriate than traditional, average measures.

*Corresponding author: Barbara Rossi, CREI, Universitat Pompeu Fabra, c. Ramon Trias Fargas 25-27, Mercè Rodoreda bldg., 08005 Barcelona, Spain. E-mail: barbara.rossi@upf.edu

Acknowledgments. I thank Steven Durlauf, four anonymous referees, T. Clark, M. McCracken, G. Ganics, L. Hoetsch, J. Mitchell, T. Sekhposyan, A. Tarozzi and conference participants at the Symposium of the Spanish Economic Association, the 2019 Forecasting Conference and the 2019 Macroeconomic Workshop as well as Bologna University. I also thank A. Goyal for providing data used in this article. Supported by the ERC (grant agreement 615608) and the Cerca Programme/Generalitat de Catalunya. The BGSE acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R&D (SEV-2015-0563).

J.E.L. Codes: E4, E52, E21, H31, I3, D1.

Keywords: Forecasting, Instabilities, Time Variation, Inflation, Structural Breaks, Density Forecasts, Great Recession, Forecast Confidence Intervals, Output Growth, Business Cycles.

1 Introduction

Instabilities are widespread in economic time series. For example, casual observation points to the sharp reduction in the volatility of several macroeconomic aggregates around mid-1980s, marking the beginning of a period called “the Great Moderation”, followed a couple of decades later by a large financial crisis, “the Great Recession”, during which the relationship among macroeconomic variables changed abruptly. More broadly, in a series of papers, Stock and Watson (1996, 2003) point to the existence of instabilities in a wide variety of macroeconomic and financial time series. What are the consequences of such instabilities for forecasting and evaluating models' predictive ability?

We answer four important questions that arise when forecasting in unstable environments. The first question is: *What are forecast instabilities and why should we care about them?* We will answer this question by illustrating the importance of instabilities in four empirical and highly relevant economic examples: (i) the great recession of 2007-2008 (Ng and Wright, 2013; Alessi, Ghysels, Onorante, Peach and Potter, 2014); (ii) instabilities in forecasting asset returns and exchange rates (Goyal and Welch, 2008; Rossi, 2013); (iii) instabilities in predicting inflation; and (iv) instabilities in survey density forecasts. The examples anticipate the themes of this article: (a) the forecasting ability of economic predictors does change over time; and (b) evaluating forecasting models using traditional methods fails in the presence of the instabilities that we typically observe in macroeconomic and financial data. Additional theoretical examples help clarify what we mean by instabilities in forecasting performance. Note that, in this article, we interpret instabilities in a broad sense, including time-variation in forecasting performance. Importantly, we show that breaks in models' parameters that are of interest to the researcher are neither necessary nor sufficient to generate time variation in models' forecasting performance: thus, one should not test for breaks in such parameters, but rather evaluate their forecasting ability in a robust way.

In fact, the second important question that we address is: *How to evaluate whether a model forecasts well in the presence of instabilities?* Our emphasis is on stable and satisfactory *forecast performance*, rather than stable *forecasts*. We illustrate how to evaluate the forecasting performance of a model either in isolation ("absolute" forecasting ability) or relative to its competitors ("relative" forecasting ability).

A third, crucial question is: *"How to improve forecasts in the presence of instabilities?"*. We review strategies that help improving models' forecasting performance by either explicitly including instabilities at the model estimation stage or by exploiting big data.

Finally, the fourth question is: *"How to correctly measure and assess forecast uncertainty in unstable environments?"*. The latter is an important question for policymakers, who frequently add fan charts around their forecasts or confidence intervals around their historical path to convey an assessment of their uncertainty.

It is important to note that this survey is useful not only for forecasters and practitioners, who routinely produce and use forecasts, but more broadly for researchers whose goal is to evaluate the performance of their models. In fact, out-of-sample forecasts are not only a useful guide for investors and forecasters, but also a diagnostic tool and a reality check on models' performance. Indeed, evaluating models' performance in-sample faces the risks of over-fitting, data snooping and lack of robustness to the presence of instabilities: evaluating models' in terms of their out-of-sample forecasting ability helps alleviate these problems.[1] On the other hand, Diebold (2015) argues in favor of a more cautionary use of forecast tests: while they are appropriate for evaluating forecasts, they may not always be the best option for evaluating specific features of a model, as they may involve a loss of power: if a researcher has clearly in mind which features he/she would like to evaluate (e.g. instabilities in the model's parameters), there might be better in-sample tests to achieve such goal.

This survey is also relevant for central banks and policy institutions given the challenges they face in the presence of instabilities. Their models can be broadly divided in two categories: reduced-form (statistical) and structural (behavioral) models. The former are typically used for forecasting and the latter for policy counterfactuals. As we will explain more in detail in Section 3, while the examples in Section 2 focus on a simple linear model for clarity of exposition, all the forecast evaluation methodologies reviewed in Section 3 and the forecast combinations described in Section 4.2.3 are loosely analogous in reduced-form and structural models. For the latter, robustness to instabilities is indeed a crucial feature in light of the Marschak-Timbergen-Lucas critique. According to the critique, if the parameters of a model are not structural (i.e. not invariant to policy), then they necessarily change as a consequence of policy changes. In Lucas' (1976, p. 41) own words: "Given that the structure of an econometric model consists of optimal decision rules of economic agents, and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models." Thus, if one forecasts the effects of changes in policy using a model whose parameters endogenously change after a change in policy, the policy recommendations based on such models may potentially be misleading. Forecast evaluation may provide a useful reality check for such models. On the other hand, however, if policy is conducted with the goal of eliminating predictable variation, reduced-form models may not predict the economic variables targeted by policymakers exactly because of their own actions – see McLeay and Tenreyro (2019).

Another area where this survey is relevant is finance. Sometimes, the perception that asset returns are predictable is due to back-testing, in-sample over-fitting, or data-snooping. Evaluation methods based on out-of-sample forecasting ability provide an effective way to protect against these concerns. Thus, this survey also discusses empirical examples that are relevant in both finance and international finance.

On the other hand, the model that generated the forecasts may be unknown; such is the case for survey-based forecasts, which recently attracted a lot of attention given their good empirical performance (see Faust and Wright, 2013, and Ang, Bekaert and Wei, 2007, among others). While model-based forecasts are forecasts produced by using a model, survey-based forecasts are forecasts collected from survey participants. Although survey participants may use models to forecast, in practice such forecasts typically include a large amount of judgement (Stark, 2013). The empirical examples discussed in Section 2 include both model-based and survey-based forecasts. The evaluation of models' forecasts in Section 3 is also discussed in the context of both model-based as well as survey-based forecasts. However, we can only discuss how to improve model-based forecasts if alternative predictors

[1] See Clark and McCracken (2005). It is also true that even out-of-sample forecast evaluation procedures might be subject to strategic data mining in finite samples. See e.g. Inoue and Kilian (2004) for a discussion.

can be included or the time-variation can be explicitly modeled, and this is possible only for model-based forecasts; however, survey forecasts can be, and often are, combined with other model-based forecasts to improve their performance, as discussed in Section 4.2.3.

Finally, there are several surveys on forecasting and instabilities: how does this survey differ from the existing ones? This survey provides a broad and informal introduction to the difficulties faced by forecasters and researchers in the presence of instabilities. It aims at increasing the awareness of the challenges that they face by focusing on a series of empirical examples of interest to economists as well as researchers at economic, financial and policy institutions. It also illustrates a variety of recently developed approaches to overcome such difficulties and robustify the empirical findings. For more details on the practical implementation and the formal justification of several of the methods illustrated in this survey, see the technical review by Rossi (2014a). Other surveys target more specific topics: Goyal and Welch (2008) and Timmermann (2008) focus on asset predictability; Stock and Watson (2003) on macroeconomic variables and asset prices; Faust and Wright (2013) on inflation; Rossi (2013) on exchange rates; Ng and Wright (2013) and Alessi, Ghysels, Onorante, Peach and Potter (2014) on the great recession.

This article is organized as follows. Section 2 motivates why instabilities are important in forecasting using four key illustrative empirical examples; it also clarifies what we mean by instabilities in forecasting performance and how they differ from in-sample structural breaks. Section 3 overviews methodologies to assess whether models forecast well in the illustrative examples previously introduced. Section 4 discusses strategies for improving models' forecasting performance; Section 5 focuses on how to report measures of uncertainty around forecasts, and Section 6 concludes.

2 Why Should We Care About Instabilities When Forecasting?

Why are instabilities important when forecasting? In what follows, we discuss four empirically relevant examples, which include both model-based forecasts as well as survey's. Each one of the examples is motivated by four key areas that are of particular interest to forecasters and economists, and where instabilities are predominant: forecasting the great recession of 2007-2009 and the slow recovery afterwards; predicting inflation dynamics; forecasting asset prices; and forecasting the whole probability distribution of output growth. Each of the examples will in turn illustrate four key themes in forecasting under instabilities that will be the focus of the next sections: assessing whether a given model forecasts well in the presence of instabilities; evaluating which model forecasts best among competing models; which strategies have been most successful in improving models' predictive ability; and assessing as well as improving measures of uncertainty around forecasts, such as predictive densities and forecast confidence intervals.

2.1 Forecasting the Great Recession of 2007-2009 and the Slow Recovery

The Great Recession of 2007-2009 was the largest recession faced by the United States after WWII. Ng and Wright (2013) and Alessi, Ghysels, Onorante, Peach and Potter (2014) investigate to what extent the crisis was forecastable in real-time. In particular, Ng and Wright (2013) overview the effectiveness of a series of economic predictors as well as survey forecasts made by professional forecasters (such as the Survey of Professional Forecasters – SPF thereafter) and policy institutions (such as the Federal Reserve's Greenbook forecasts). They find that some predictors would have been potentially useful, but, at the same time, those that would have been useful at the time of the crisis were different from those in normal times. Alessi, Ghysels, Onorante, Peach and Potter (2014) focus on the macroeconomic forecasts of the Federal Reserve Board and the European Central Bank around the time of the crisis, and compare them with surveys. They note that the central banks' forecasts were aligned to survey participants' before the crisis, but that they behaved very differently at the time of the crisis: surveys underestimated the severity of the crisis more than the central bank.

We revisit the empirical evidence using the most recent data. Figure 1 plots forecasts of real output growth (measured as the growth rate of real Gross Domestic Product – GDP in short) made by the Federal Reserve (labeled "Greenbook", dashed line) as well as the SPF (labeled "SPF", dotted line) together with the actual realizations of output growth. The forecast horizon is three quarters. At any point in time, the vertical gap between the realization (depicted by the solid line) and the forecast is the forecast error: a positive gap means that the forecast is under-predicting the target variable.[2]

The figure shows several features that suggest the presence of instabilities in surveys' and Federal Reserve's forecasting performance. On the one hand, both forecasts were roughly on target between 2003 and 2007, that is, up to the Great Recession. On the other hand, the magnitude of the large recession was difficult to forecast by the Federal Reserve and the SPF, both of which substantially under-predicted the severity of the decrease in output growth. The figure confirms that, indeed, survey participants underestimated the severity of the crisis more than central banks, as noted by Alessi, Ghysels, Onorante, Peach and Potter (2014); however, since the end of the financial crisis, central banks overestimated the recovery more than survey participants.

Hence, the forecasting performance appears to have been subject to a serious deterioration at the time of the latest financial crisis. This example raises an important question: how to improve models' forecasts in the presence of instabilities? *Which estimation strategies and, more generally, which approaches to model and forecast specification can help guard against forecast breakdowns? Section 4 will tackle this issue.*

It is important to note that the poor forecasting performance is conditional on the model used by the researcher and that the forecast might be poor because either the conditional mean or the volatility are incorrectly specified (or both). In particular, when the volatility is time-varying, the model specified by the researcher may produce poor forecast densities if it

[2] The data are available from the Federal Reserve Bank of Philadelphia. The end date is constrained by the availability of the Greenbook (Tilebook) forecasts, which are only published with a delay of five years.

does not include such time-variation. While models in finance typically allow time-varying volatilities (stochastic volatility or GARCH), as the latter is widely understood to be a key feature of returns, macroeconomic models may or may not let volatilities be time-varying, and in fact a large part of the literature assumes constant volatility. Thus, if a researcher specifies a constant volatility model and the volatility turns out to be time-varying, this is a source of forecast breakdown, *conditional* on the model estimated by the researcher. In Section 5, we discuss forecast densities and review several prominent models developed in the last decade that explicitly allow for time variation in the volatility.

INSERT FIGURE 1 HERE

2.2 Forecasting Inflation

Inflation is a key variable for central banks, the private sector and financial institutions. In fact, firms forecast inflation to set their prices; financial institutions forecast inflation to predict real returns of investments; and an important concern for central banks is to target inflation to a suitable level.

The behavior of inflation has changed over time. Early on, Stock and Watson (1999a) argued that the coefficients of the Phillips curve – a model where the current level of unemployment is used to predict future inflation – are time-varying; furthermore, the Phillips curve model can be improved by including additional measures of real aggregate activity. Others have argued that the long-run inflation trend is time-varying (Cogley and Sbordone, 2008).

These findings suggest that inflation forecasts' properties changed over time. Indeed this is the case. Figure 2 shows two-quarter-ahead forecasts of US inflation in the last four decades. The measure of quarterly inflation that we consider is the (annualized) quarter-over-quarter GNP deflator growth rate up to 1991 and the (annualized) quarter-over-quarter GDP deflator growth rate afterwards, consistently with the Greenbook forecasts. The data are from the Federal Reserve Bank of Philadelphia, and are matched to the timing of SPF, also provided in their dataset. The realized values are the second revision from the real-time dataset by Croushore and Stark (2001). Again, the dashed line depicts the forecasts made by the Federal Reserve and the dotted line those made by the SPF. The solid line in the figure shows the realized inflation rate. The figure highlights interesting time-varying patterns in the Federal Reserve and SPF's forecasting performance. Both forecasts under-predict actual inflation between 1970 and 1980, a time when unforecastable oil prices persistently hit the economy; after 1980, however, both forecasts consistently over-predict inflation for two decades. Hence, the predictive ability of the SPF as well as that of the Federal Reserve did substantially change over time.

This example raises a second, important question, namely: *How can one assess models' ability to accurately predict the target variable when the predictive ability changes over time?* This is a question related to the evaluation of a forecast, rather than its construction or estimation. In other words, *if a model's forecasting performance shows time-varying features, are its forecasts "unbiased" – in the sense of displaying a good performance in tracking the ex-post realizations? We discuss this topic in Section 3.1.*

INSERT FIGURE 2 HERE

2.3 Forecasting Asset Returns and Exchange Rates

Financial returns – whether asset prices, the stock market, exchange rates or equity premia – are notoriously very difficult to forecast. Part of the difficulty in forecasting financial returns is again related to the presence of time variation. As summarized in Goyal and Welch (2008), when predicting the equity premium: "different articles use different techniques, variables, and time periods. (. . .) Some articles contradict the findings of others. Still, most readers are left with the impression that 'prediction works' – though it is unclear exactly what works. . . " Goyal and Welch (2008) find that most models that performed well in the past (especially during the oil shock period in the early 1970s) have nowadays poor out-of-sample forecasting performance relative to the historical mean: again, the predictive ability of the models has changed over time, but in relative terms. Goyal and Welch (2003) found similar results for the dividend ratio. Pesaran and Timmermann (1995) and Rapach and Wohar (2006) find that stock market return predictability is present only in a few sub-samples; Paye and Timmermann (2006) find breaks in the coefficient of stock return predictive regressions; and Pesaran and Timmermann (2000) note that oil prices were an important predictor for stock prices during the 1970s but that their importance subsequently vanished. Similarly, Rossi (2013) finds that the predictive ability of exchange rate models is ephemeral and changes over time.

The time-variation that makes asset prices so difficult to predict comes from many sources: external shocks that continuously hit the markets; dynamic changes in the economy; the low power of the economic predictors; it may also be due to predictability being endogenously affected by forecasters' own attempts to identify and exploit it.[3] According to Timmermann (2008, p.2), predictability is "a moving target that is changing over time. Just when a forecaster may think that he has figured out how to predict returns, the dynamics of market prices will, in all likelihood, have moved on – possibly as a consequence of the forecaster's own efforts". In other words, forecasters attempt a variety of different approaches to predict, and the most successful models "will be rapidly adopted by other forecasters, therefore affecting the dynamic evolution of the returns over time and dissipating whatever predictability there was in the first place." Supporting Timmermann's (2008) conjecture, Sullivan, Timmermann and White (1999) find that the performance of some trading rules disappeared shortly after their publication. The competition among forecasters implies that any forecasting method will have a poor out-of-sample performance over a long time period; however, for the same reason, any given model may work well for very short periods of time. Again, according to Timmermann (2008), "there appear to be pockets in time where there is modest evidence of local predictability; (...) the identity of the best forecasting method can be expected to vary over time, and there are likely to be periods of model breakdown where no approach seems to work".

We consider two illustrations of instabilities in financial variables. The first focuses on forecasting the equity premium. Figure 3 depicts realizations of U.S. equity premia,

[3] Data snooping may also be another source; we will discuss it separately in Section 4.2.5.

together with forecasts based on several of Goyal and Welch's (2008) economic predictors. The benchmark model is the historical mean, calculated in a rolling window over the previous twenty years. The rolling mean attempts to capture a slowly changing mean in equity returns in an agnostic way. The other forecasts, also based on a twenty-year rolling window, are based on: the book to market ratio, calculated as the ratio of the book value and the market value of the Dow Jones Industrial Average (labeled "BookToMkt"); the default yield spread, calculated as the difference between BAA and AAA-rated corporate bond yields (labeled "DFY"); the investment to capital ratio (labeled "Inv/K"); the consumption, wealth and income ratio (Lettau and Ludvigson, 2001, labeled "CAY"); the long term government bond yield (labeled "LongYield"); and the term spread, calculated as the difference between the long term yield on government bonds and the Treasury bill (labeled "Spread").[4]

Figure 3 visually confirms that, most of the time, asset return predictability seems elusive and local. For example, note how the book to market ratio seems to provide a good fit to the downturn in equity premia in 1974, and especially throughout the 1980s, but completely missed the large drops in the early 2000s; in contrast, the default yield spread did catch the large drop in 2008 and the recovery afterwards, but was not a good predictor in the 1970s and 1980s.

INSERT FIGURE 3 HERE

In international finance, a similar problem is faced by researchers attempting to predict exchange rate returns. Figure 4 depicts realizations of the rate of growth of the U.K. pound/U.S. dollar exchange rate (solid line, labeled "Realization") together with the one-month-ahead forecasts of two models. The first model is the random walk, a typical benchmark in the literature, according to which exchange rates are unpredictable. Hence, its forecasts are always zero, depicted by the dotted line. The second model is uncovered interest rate parity (dashed line, labeled "UIRP"), according to which the interest rate differential between two countries should predict their bilateral exchange rate.[5] The picture suggests that, in the first part of the sample and up to the end of 1990, UIRP forecasts track the subsequent downward and upward swings of realized exchange rates somewhat better than the random walk. However, the model's performance seems to worsen afterwards, suggesting that the predictive power of relative interest rates may have disappeared after 1990.

In both the equity premium and the exchange rate prediction examples, the empirical evidence indicates that the predictive ability of the economic variables may be time-varying relative to the benchmark model. This is an issue regarding the evaluation of the forecasting ability of a model relative to a benchmark. *In other words, the models' relative forecasting performance seems to have changed over time. But, did it actually change? How to formally evaluate whether that is the case? And how to identify the best model under these circumstances? We will tackle this topic in Section 3.2.* It is also important to note that,

[4] The data are available on A. Goyal's website: http://www.hec.unil.ch/agoyal/

[5] The data sources are: 3-Month Rates and Yields: Treasury Securities for the interest rate in the United Kingdom (available up to 2017); the 3-Month US Treasury Bill: Secondary Market Rate (TB3MS) for the US interest rate; and the U.S. / U.K. Foreign Exchange Rate (EXUSUK) for the exchange rate. All data are monthly from the Federal Reserve Bank of St. Louis' FRED database.

while the failure of economic models to beat simple benchmarks is often ascribed to instabilities, another interpretation is the following. Data snooping (or data mining) could lead researchers to find an apparent predictive relation which is not stable exactly because it is a fluke. Section 4.2.5 discusses the importance of being robust to data snooping.

INSERT FIGURE 4 HERE

2.4 Instabilities in Predictive Densities and Forecast Confidence Intervals

Instabilities are present not only in point forecasts but in predictive densities as well. Figure 5 depicts the one-year-ahead SPF prediction of output growth together with the 50%, 90% and 95% quantiles of the SPF's predictive density (the lightest, medium and darkest bands, respectively).[6] The darkest, continuous line plots realized output growth. The forecasting performance of the survey densities appears to be time-varying: it generally encompasses the realization in normal times, but is overly optimistic during the two recessions in the sample – the 2001 recession and the 2007-2009 financial crisis. Aastveit, Carriero, Clark and Marcellino (2017) also find evidence of instabilities during the recent financial crisis in both point and density forecasts. Alessi, Ghysels, Onorante, Peach and Potter (2014) find that central banks' density forecasts were aligned with survey's forecasts before 2007 but much more pessimistic afterwards, suggesting that central banks were quicker in recognizing the downside risks of the financial crisis than survey participants.

This example motivates the need for methods to assess and improve forecast densities (as opposed to point forecasts, which were the focus of the previous examples). *As the performance of predictive densities changes over time, how to assess whether predictive densities are correctly specified in the presence of time-variation? And how to compare predictive densities? These issues will be investigated in Section 5.*

INSERT FIGURE 5 HERE

2.5 What Are Instabilities in Forecasting Performance? Definitions and Examples

What do we mean by instabilities in forecasting ability? And how do they differ from breaks in models' parameters? The following theoretical examples provide illustrations as well as clarifications. Throughout this survey, we measure forecasting ability in terms of the out-of-sample square forecast error (its average will be referred to as the mean square forecast error, or MSFE in short). The sequence of forecast errors are obtained in a pseudo out-of-sample forecast environment, by mimicking what a forecaster would have done in real time: estimating the model using observations up to a certain point in time to predict the

[6] The average is calculated as the average probability attributed by the sample of forecasters to each bin; then the average predictive density over bins is smoothed out using a Gaussian distribution. The fixed-event predictive densities are transformed to a fixed-horizon by taking a weighted average across forecast horizons – see Rossi, Sekhposyan and Soupre' (2018) for details.

value of a target variables after h periods, and then taking the difference between the ex-post realization of that variable and the forecast. The resulting sequence of out-of-sample forecast errors will be denoted by $\varepsilon_{t+h|t}$, for $t = R, ..., T$. Conditional on a choice of loss function, we define forecast instabilities as instabilities in the loss. Since in our case we assume a quadratic loss,[7] forecast instabilities are instabilities in the squared forecast errors, $\varepsilon^2_{t+h|t}$. Such instabilities can be smooth and continuous or discrete and abrupt; they can take the form of structural breaks, regime switches, etc.

Instabilities in the forecasting performance are often linked to structural breaks in the models' parameters that are of direct interest to researchers – e.g., the conditional mean parameter β in the case of a linear model where $E_t y_{t+h} = \beta x_t$. However, as argued in Giacomini and Rossi (2010) and Rossi (2014a), while structural breaks in the parameters of direct economic interest to researchers (i.e. β) might cause instabilities in forecasting, they are neither necessary nor sufficient. They are not necessary because the effect of instabilities in several parameters could cancel themselves out, resulting in a constant MSFE over time. Also, it is well-known that the MSFE can be decomposed into two components: the squared bias and the variance. A bias originating from structural breaks in the parameters of one model might counterbalance a higher parameter estimation error variance in the competing model, resulting in the two models having the same MSFE. They are not sufficient either, because the model might be misspecified, or because the instabilities do not appear in the part of the model that the researcher is considering. Although this is a very simple point, it is often misunderstood or under-appreciated; therefore, this section formally discusses several examples to clarify it.

It is important to clarify at the onset that the definition of forecast instabilities taken in this article refers to the forecast performance rather than the forecasts themselves. The forecasts may be stable, yet the forecast performance may display instabilities. For example, suppose a country's output growth forecast is always the average growth for that country. This forecast is constant but its performance might be good in normal times and bad in expansions and recessions. Thus, the model's predictive ability is time varying and the performance is, at times, poor: in fact, the forecast errors are zero in normal times, positive in expansions and negative in recession. Furthermore, according to traditional procedures, the forecasts of this model may appear unbiased, as the positive forecast errors in expansions could, on average, cancel out the negative forecast errors in recessions. However, from time to time, the model is not predicting well, and the techniques robust to instabilities discussed in this article would highlight such poor forecasting performance. It is also important to clarify that, as discussed in the previous section, the forecast performance is conditional on the model estimated by the researcher. For simplicity, and in line with typical macroeconomic approaches, we distinguish between the parameter of direct interest to the researcher (that is, the parameters in the conditional mean of the forecast, β), whose time-variation may or may not be modeled, and other parameters whose time-variation is un-modeled (the volatilities).[8]

Let the variable to be forecasted at time $(t + h)$ be denoted by y_{t+h} and the forecast made

[7] Other loss functions can be used – see Elliott and Timmermann (2008). For example, the absolute loss, which corresponds to the absolute forecast error.

[8] Other examples with similar conclusions can be worked out for more general loss functions and estimated time-varying volatilities while letting other moments of the data change over time.

at time t be denoted by $y_{t+h|t}$, where $h > 0$ is the forecast horizon; thus, $\varepsilon_{t+h|t} = y_{t+h} - y_{t+h|t}$, for $t = R,\ R+1, ...,\ T$. When necessary, superscripts refer to a model, generically denoted by m, where, in this section, $m = 1,\ 2$. The first example shows that parameter instability does not necessarily result in unstable forecasting performance.

Example 1 (Parameter Instability $\nRightarrow$ Unstable Forecast Performance) *Let the data observed by the researcher be generated as follows:*

$$\begin{aligned} y_t &= \beta_t x_{t-1} + \varepsilon_t, && (1) \\ x_t &\sim \textit{independent } N(0, \sigma^2_{X,t}), && (2) \\ \varepsilon_t &\sim \textit{independent } N(0, \sigma^2_{\varepsilon,t}), && (3) \end{aligned}$$

and x_t and ε_t be independent of each other. The forecasting model assumes that y_t is unpredictable. The researcher evaluates the model according to its one-step-ahead out-of-sample forecasts. The squared forecast error at time $(t+1)$ equals $\sigma^2_{\varepsilon,t+1} + \beta^2_{t+1}\sigma^2_{X,t}$. Hence, the model's forecast performance can be constant even if the parameters of the model (β_t) are time-varying, as long as the variability in the parameter cancels out the variability in the predictor or in the error term, resulting in a constant value of $\sigma^2_{\varepsilon,t+1} + \beta^2_{t+1}\sigma^2_{X,t}$. That is, $\sigma^2_{\varepsilon,t+1} - \sigma^2_{\varepsilon,t} \simeq -\beta^2_{t+1}\sigma^2_{X,t} + \beta^2_t\sigma^2_{X,t-1}$. Note that the predictability can be constant even if the underlying true model changes over time. For example, when $\beta_t = 0$ and $\sigma^2_{\varepsilon,t} = \sigma^2_\varepsilon$ constant for $t \leq \tau$, while $\beta_t \neq 0$ and $\sigma^2_{\varepsilon,t}$, $\sigma^2_{x,t}$ are time-varying such that $\sigma^2_{\varepsilon,t+1} - \sigma^2_{\varepsilon,t} \simeq -\beta^2_{t+1}\sigma^2_{X,t} + \beta^2_t\sigma^2_{X,t-1}$ for $t > \tau$, the model switches from y being unpredictable to being predictable using past values of the $x's$.

The next example, taken from Giacomini and Rossi (2010), shows that structural breaks in the parameters need not necessarily result in unequal nor unstable relative predictive performance.

Example 2 (Parameter Instability $\nRightarrow$ Unstable Relative Forecast Performance) *The data are generated as follows:*

$$\begin{aligned} y_t &= \beta_t x_t + \varepsilon_t, && (4) \\ x_t &= .5x_{t-1} + \nu_t, \\ \nu_t &\sim iid\ N(0,1), \varepsilon_t \sim iid\ N(0, \sigma^2_\varepsilon),\ \textit{independent of each other,} \end{aligned}$$

and, for simplicity, x_{t+1} is known at time t. The researcher compares one-step-ahead out-of-sample forecasts of two models. The first model is eq. (4), where the parameter β is estimated in-sample by OLS over a rolling window using the last R observations $(\widehat{\beta}_{t,R})$, where R is finite and fixed. Hence, its forecast at time t is:

$$y^{(1)}_{t+1|t} = \widehat{\beta}_{t,R} x_{t+1}. \tag{5}$$

The second model instead assumes that Y_t is unpredictable; hence, its forecast is:

$$y^{(2)}_{t+1|t} = 0, \tag{6}$$

The relative forecasting performance, measured by the expected squared forecast error difference, is the same if:

$$E\left[\left(y_{t+1}-y_{t+1|t}^{(1)}\right)^2\right]=E\left[\left(y_{t+1}-y_{t+1|t}^{(2)}\right)^2\right]. \tag{7}$$

Condition (7) is satisfied if

$$\beta_{t+1}=\frac{\frac{\left(\sum_{j=t-R+1}^{t}\beta_j x_j^2\right)^2}{\sum_{j=t-R+1}^{t}x_j^2}+\sigma_\varepsilon^2}{2\sum_{j=t-R+1}^{t}\beta_j x_j^2},\ t=R,...,T-1.^{9} \tag{8}$$

Thus, *the models' expected relative forecasting performance can be the same at each point in time even though the parameters in the conditional mean of the true model are time-varying.*

Note that, in this case as well, the true model could switch over time, yet the relative predictive ability does not change: this would be the case if, for example, $x_t = 1$, $\beta_t = \sigma_\varepsilon^2/t$ for $t \leq \tau$ and, for $t > \tau$, x_t, β_t are time varying such that eq. (8) holds. On the other hand, one could also encounter situations where the parameters of the model that are of direct interest to the researcher are constant, yet a model's forecasting performance or its forecasting performance relative to a competitor is time-varying, as the following examples show.

Example 3 (Unstable Forecast Performance With Constant Parameters) *Let the data be generated as in Example 1, except that $\beta_t = \beta$ is constant. The researcher evaluates the one-step-ahead out-of-sample forecasts of a model that assumes that y_t is unpredictable. Its expected one-step-ahead squared forecast error at time $(t+1)$ equals $\sigma_{\varepsilon,t+1}^2+\beta^2\sigma_{X,t}^2$, which can be time-varying even if β is constant. Hence, a model's forecasting performance can be time-varying even if the parameter in the conditional mean of the model (β) is constant. Note that the model's forecasting performance will be time-varying even if the variance of the errors is constant, as long as the variance of the regressor is time-varying.*

Example 4 (Unstable Relative Forecast Performance With Constant Parameters) *Let the data be generated as follows:*

$$\begin{aligned} y_t &= \beta x_{t-1}+\varepsilon_t, \\ x_t &\sim \textit{independent } N(0,\sigma_{X,t}^2), \\ \varepsilon_t &\sim \textit{iid } N(0,\sigma_\varepsilon^2), \end{aligned} \tag{9}$$

where x_t and ε_t are independent of each other. The researcher compares the one-step-ahead out-of-sample forecasts of two models: a model where the predictor is just a constant and the true model (eq. 9) where the researcher knows the true value of β when making the forecast. The forecast error of the first model is $y_{t+1}-R^{-1}\Sigma_{j=t-R+1}^{t}y_j = \beta x_t +$

[9] See Giacomini and Rossi (2010).

$\varepsilon_{t+1} - \left[R^{-1}\Sigma_{j=t-R+1}^{t}\left(\beta x_{j-1} + \varepsilon_j\right)\right]$, *while that of the second model is* ε_{t+1}*. Since the regressors are independent of the error term and both are independent over time, the expected squared forecast error of the first model at time* $(t+1)$ *is* $\beta^2\left[\sigma^2_{X,t} + \left(R^{-1}\Sigma^t_{j=t-R+1}\sigma^2_{X,j-1}\right)\right]$ $+\sigma^2_{\varepsilon} + \sigma^2_{\varepsilon}/R$*. The second model has an expected squared forecast error equal to* σ^2_{ε}*. Hence, the expected relative forecasting performance of the two models at time* $(t+1)$*, measured by* $E\left[\left(y_{t+1} - y^{(1)}_{t+1|t}\right)^2 - \left(y_{t+1} - y^{(2)}_{t+1|t}\right)^2\right]$*, equals* $\beta^2\left[\sigma^2_{X,t} + \left(R^{-1}\Sigma^t_{j=t-R+1}\sigma^2_{X,j-1}\right)\right]$ $+\sigma^2_{\varepsilon}/R$*, and hence can be time-varying even if the parameters of the model are constant.*

This fourth example shows that *the relative forecasting performance could be time-varying even if the parameters of the model that are of direct interest to the researcher* (β) *are constant.* Instabilities in forecasting performance may also be due to model mis-specification or data snooping. For example, estimating a linear model when the model is non-linear (e.g. Markov-switching) may produce time-varying forecast errors. *Importantly, structural break tests on* β *in eq. (9) – such as Hansen's (2000) test – would not be able to signal the presence of time-variation.* The lesson to be learned from these examples is the following: *if one is interested in forecasting performance, then one should not test for structural breaks in the parameters, but rather evaluate models' forecasting ability in a way robust to instabilities.*

Similar examples hold for other non-quadratic loss functions as well as density forecasts, which will be the topic of Section 5. Regarding the latter, consider the case where: $y_{t+1} = \beta_t' x_t + \varepsilon_{t+1}$, and the researcher compares the performance of two Normal predictive densities, with mean equal to the point forecasts $y^{(1)}_{t+1|t}$ and $y^{(2)}_{t+1|t}$ in eqs. (5)-(6) in Example 2, and calibrating their variances to be one (for simplicity). Assume that the relative predictive performance of the forecast densities is evaluated using a log score, which is the logarithm of the forecast density evaluated at the actual realization. Then, the expected relative forecasting performance at time t is the same if: $E\left\{\ln(2\pi) - \frac{1}{2}\left(y_{t+1} - y^{(1)}_{t+1|t}\right)^2\right\}$ $= E\left\{\ln(2\pi) - \frac{1}{2}\left(y_{t+1} - y^{(2)}_{t+1|t}\right)^2\right\}$. The latter condition is equivalent to eq. (7); hence, the relative performance of predictive densities can be the same at each point in time even though the parameters of the conditional mean of the true model are time-varying, as in Example 2. Conclusions similar to those in Examples 1, 3 and 4 can be reached for density forecasts.

So what do we mean by instabilities in forecasting performance? It is not just, nor necessarily, time variation in the parameters of direct interest to the forecaster (β). The forecasting performance depends on the forecast errors and how they are evaluated: instabilities in the forecasting performance could originate not only from changes in the parameters in the conditional mean of the model, but also from changes in the unpredictable component (e.g. the mis-specified component in the forecasting model in Example 3), changes in the variance of the predictors, changes in the volatility of the shocks, and so forth.

In addition, there is an important distinction between the goal of *evaluating forecasts out-of-sample* and that of *evaluating models in-sample.* Forecast performance refers to the out-of-sample predictive ability of a model, judged according to a loss function (e.g. the squared out-of-sample forecast error) while model evaluation typically aims at its correct specification in-sample.

It is important to note that, as in the predictive density example above, time-varying variances may cause unstable forecast performance in models that have been estimated without taking such time variation into account. In fact, it is quite common in standard macroeconomic applications not to model volatility explicitly; thus, this implies that time-varying volatility could be a source of forecast instability in such applications. However, a researcher might choose to model time-variation in the volatility explicitly, in which case, when done appropriately, the forecasting performance may not be unstable. Time-varying volatility is typically modeled in finance via GARCH or stochastic volatility; more recently, several studies have highlighted the empirical importance of modeling time-varying volatilities for predicting densities in macroeconomic data as well. We will discuss the latter in Section 4. On the other hand, note that, should the forecaster use a loss function different from the quadratic one, instabilities in other, potentially higher moments of the data may become important as well.

3 How to Assess Whether a Model Forecasts Well in the Presence of Instabilities?

The examples above point to an important difference in how to evaluate forecasting models. It might be of interest to evaluate their forecasting performance in isolation (as in Examples 1 and 3) or, alternatively, evaluate their performance relative to a competitor (as in Examples 2 and 4). The first is referred to as "absolute forecasting performance" and the latter as "relative forecasting performance". In what follows, we illustrate how to assess both. In the next two sub-sections we first focus on forecasts aimed at targeting a scalar value of a macroeconomic variable (i.e. point forecasts), such as the inflation rate and asset prices. Section 5 discusses how to evaluate forecasts of the whole probability distribution. Notation is as in the previous section: the forecast of a variable y at time $t+h$ made at time t is denoted by $y_{t+h|t}$; the realized value is denoted by y_{t+h}; $h>0$ is the forecast horizon; and the forecast error is $\varepsilon_{t+h|t} = y_{t+h} - y_{t+h|t}$. When necessary, superscripts denote models.

3.1 Absolute Forecasting Ability

A minimum requirement that forecasts should satisfy is to track the realizations – that is, the forecast error should be small. The forecast error should also not be predictable on the basis of any variable observed at the time the forecast is made (including the forecast itself). Forecast rationality tests are designed exactly to evaluate these minimal requirements. Forecast rationality tests are F-tests to evaluate whether $\alpha = \beta = 0$ in the regression:

$$y_{t+h} - y_{t+h|t} = \alpha + \beta y_{t+h|t} + u_{t,h} \tag{10}$$

where $u_{t,h}$ is the regression error.

We evaluate rationality of the Federal Reserve and SPF inflation forecasts, described in Section 2.2. Given the potential presence of instabilities, we implement the Fluctuation Rationality test. The test involves repeatedly testing forecast rationality in rolling windows, then comparing the largest value with the appropriate critical value (Rossi and Sekhposyan, 2016). That is, one calculates the F-statistic ($W_{t,m}$) for testing $a = \beta = 0$ in rolling windows of, say, size m centered around the forecast time t, and repeats the test for all available points in time t; then, the Fluctuation Rationality test is $\sup_t W_{t,m}$. More details are provided in Algorithm 6 in the Appendix.

Figure 6 plots $\mathcal{W}_{t,m}$ for the SPF (continuous line, labeled "SPF") and for the Federal Reserve forecasts (dashed line, labeled "Greenbook"). Both are above the 5% critical value (depicted by the dotted line, labeled "cv"); hence, forecast rationality is indeed rejected.

Conversely, the value of the traditional forecast rationality test not robust to instability (Mincer and Zarnowitz, 1969)[10] is 0.41 for the Federal Reserve forecasts (its p-value equals 0.81) and 0.44 for the SPF (its p-value is 0.80). That is, there is no evidence against forecast rationality according to the non-robust test. Yet, the forecasts depicted in Figure 2 clearly suggest a systematic bias: both the central bank and the survey forecasts under-predict inflation in the 1970s and 1980s and over-predict inflation in the last three decades. The reason why non-robust tests are unable to find evidence of biases is that the positive forecast errors in the 1970-1980s compensate the negative forecast errors in the last three decades, so that forecast errors are, on average, zero. Simply put, non-robust tests may lack the ability to detect deviations from rationality when there are instabilities in the forecasting performance – see Rossi (2005) for a formal discussion.[11] Thus, *local measures (tests) of forecast performance are more appropriate than average measures (tests) in the presence of instabilities.*

One might hope that structural models' forecasts would be rational; however, the empirical evidence shows that popular structural models' forecasts are also not rational (Edge and Gürkaynak, 2010; Gürkaynak, Kısacıkoğlu and Rossi, 2013).

INSERT FIGURE 6 HERE

3.2 Relative Forecasting Ability

Similar problems affect the comparison of competing forecasts in the presence of instabilities. In Section 2.3, we discussed the predictability of asset prices and exchange rates. We noted that economic predictors may, at times, forecast equity premia or exchange rates above and beyond the benchmark model – the historical mean for equity premia and the random walk for exchange rate returns. In this sub-section, we formally investigate whether this is the case using the Fluctuation test (Giacomini and Rossi, 2010).

The Fluctuation test is based on the relative forecast error loss of a model versus a benchmark's, scaled by a measure of its variance, calculated locally in rolling windows to follow their relative forecasting performance over time. Thus, the Fluctuation test is appropriate for small and continuous reversals in forecasting performance. In our example, the loss is the difference between the squared forecast errors of the two competing models in rolling

[10] This is F-statistic for testing $\alpha = \beta = 0$ in-sample using all the observations. See West and McCracken (1998) for forecast rationality tests that explicitly take into account models' parameter estimation error.

[11] Other forecast rationality requirements are that no other variables should predict the forecast errors or that the forecast errors be uncorrelated. Such requirements can be tested in a similar fashion.

windows ($\Delta\mathcal{L}_{t,h} \equiv \left(y_{t+h} - y^{(1)}_{t+h|t}\right)^2 - \left(y_{t+h} - y^{(2)}_{t+h|t}\right)^2$). Consistently with the previous sections, the average relative forecast error loss will be the difference of the MSFEs of the two models. Unlike the recursive R-squared or the recursive MSFE measures sometimes used in the literature (cfr. Goyal and Welch, 2008; Timmermann, 2008), the Fluctuation test offers a procedure to formally evaluate the relative out-of-sample forecasting performance of competing models. Formally, the test involves repeatedly calculating the t-test on α in the regression $\Delta L_{j,h} = \alpha + u_j$, where u_j is the regression error. The t-test ($F_{t,m}$) is calculated in rolling windows of, say, size m centered around the forecast time t, and then repeated for all available points in time t. Then, the Fluctuation Rationality test is $\sup_t F_{t,m}$ and its critical values are in Giacomini and Rossi (2010). More details are provided in Algorithm 7 in the Appendix.

Figure 7 plots the Fluctuation test for comparing the economic predictors versus the past historical average benchmark in forecasting the equity premium. Each panel in the figure reports the values of $F_{t,m}$ for one of the economic predictors considered in Section 2.3 (the continuous line), together with the critical value of the test (the dashed line). When the Fluctuation test is above the critical value, there is evidence that the model with predictors performs better than the historical average benchmark.

The results clearly show the presence of "pockets of predictability" (Timmermann, 2008): using the DFY, for example, it is possible to beat the historical mean; however, the predictability only appears sporadically (especially in the 1950s and 1960s). Results are similar across predictors with the exception of the spread, which never shows any ability to forecast better than the benchmark.

INSERT FIGURE 7 HERE

As a second example, consider the problem of forecasting exchange rates. It is well-known that the random walk is the toughest benchmark when forecasting exchange rates. For example, using monthly data from 1973:3 for the British pound/U.S. dollar, one finds that the random walk MSFE is 0.0245 while the MSFE for UIRP is 0.0249. But, has the random walk consistently been a better predictor? Figure 8 reports the Fluctuation test results: it plots $F_{t,m}$ (dark line) together with its critical value (light line). Positive values indicate that UIRP produces more competitive forecasts than the random walk. Again, it appears that there are "pockets" of predictability: UIRP was a better predictor than the random walk in the mid-1980s, but not after that.

The Fluctuation test is designed for small, continuous changes in models' forecasting performance; the One-time Reversal test (Giacomini and Rossi, 2010) is instead more appropriate in case of large, discrete reversals. The test evaluates whether the forecast performance is the same at each point in time against the alternative of a large reversal in the relative forecasting performance at some point in time, causing one of the two models to predict better.[12] In other words, the relative performance of the models may switch at an unknown point in time; under this assumption, the test can pinpoint the time of the switch. The test is implemented by jointly checking whether the (expected) squared forecast error differences

[12] The test is robust to the presence of multiple reversals in the relative forecasting performance.

in any two sub-samples of the data are constant and equal to zero via an F-test, then taking the largest value across the F-tests over time. That is, one calculates the F-statistic (W_t) for testing $a = \delta = 0$ in the regression $\Delta L_{j,h} = \alpha + \delta (d_{j,t} - t) + u_j$, where u_j is the regression error and $d_{j,t}$ is a dummy variable equal to unity if $j \leq t$. Then, the test is repeated for various values of t. The One-time Reversal test is $\sup_t W_t$ and its critical values are reported in Giacomini and Rossi (2010). More details are provided in Algorithm 8 in the Appendix.

Again, it is important to use tests that are robust to the presence of instabilities: in fact, when applied to the US dollar/UK pound exchange rate data, tests not robust to instabilities (e.g. Diebold and Mariano, 1995; West, 1996, 2006; Giacomini and White, 2006) conclude that UIRP is not significantly different from a random walk – the reason being that such tests are not robust to reversals in relative forecasting performance.

INSERT FIGURE 8 HERE

The relative forecasting performance of structural models against a constant mean or against central banks' own forecasts has also been the subject of several studies. Edge and Gürkaynak (2010) find that structural models forecast worse than a simple mean in the last two decades before the financial crisis, and Gürkaynak, Kısacıkoğlu and Rossi (2013) find that they forecast inflation worse than a simple autoregressive model.

In this survey, we focus on out-of-sample measures of forecast evaluation; alternatively, as in Stock and Watson (1999b,c), one could evaluate forecasting ability in-sample via Granger-causality tests (Granger, 1969). A Granger-causality test robust to instability has been developed by Rossi (2005).[13]

3.3 Special Issues and Technicalities

The previous section discussed evaluation strategies in the simplest and most intuitive way, i.e. via simple regressions. However, this required glancing over a variety of technicalities, which are discussed in this sub-section.

A distinction that runs through the forecast evaluation literature is that between forecast accuracy "in population" (i.e. at the true parameter values)[14] and "in finite samples" (i.e. at the estimated parameter values). Forecast accuracy in population measures predictive ability that would be present if the researcher knew the parameters of the model – or, if he/she could estimate them in a large enough sample so that they could be treated as known for practical purposes. Forecast accuracy in finite samples is the actual predictive ability that the researcher obtains in finite samples.

Forecast accuracy in population is useful typically in academic contexts, when the researcher is interested in discovering which economic model is the correct description of reality. Forecast accuracy in finite samples is useful typically for practitioners who are interested in the actual forecasting ability that a model has in the finite samples that are available – no

[13] See also Rossi and Wang (2019) for VAR-based tests of Granger-causality and their implementation in Stata as well as Giacomini and Rossi (2006) for applications and Giacomini and Rossi (2016) for in-sample model comparisons in unstable environments.

[14] Or, at the pseudo-true parameter values, if the model is misspecified. The discussion in this sub-section applies equally to true and pseudo-true parameter values.

matter how well the model could have performed if one had been able to estimate the true parameters.

There are two practical differences between the two approaches. The first is the economic interpretation of the empirical results, as discussed above: since they test two different hypotheses, the results have to be interpreted in the appropriate way. The second is that the estimate of the variance has to be corrected for parameter estimation error in the former approach, but not in the latter. The Fluctuation and Fluctuation Rationality tests are valid in both approaches by using the appropriate variance estimate. This implies that, as we anticipated in the introduction, while the example in this section focus on a simple linear model for simplicity of exposition, all the methodologies are equally applicable to both reduced-form as well as structural models. In fact, for example, in the finite sample approach, versions of the Fluctuation, Fluctuation Rationality and One-time Reversal tests can be directly used on forecasts produced by any model – be it reduced-form, structural, survey-based, DSGE-based, regime switching or time-varying parameter, under some technical assumptions.[15] On the other hand, version of the Fluctuation and Fluctuation Rationality tests can also be in principle applied to all these models in the "population approach" as well, although their practical implementation might be challenging.[16]

Another issue that requires special care involves the presence of real-time data (see Croushore, 2006, for an introduction). Out-of-sample forecasts are typically implemented in order to verify the actual predictive ability that would have been achieved by a researcher when producing the forecasts in real-time. When using a model with predictors, it is important that not only the forecasting model be re-estimated in real-time, but also that the predictors be exactly those available at the time the forecast was made. In fact, the value of, say, GDP in December 2000 that would be used by a researcher using historical data retrieved in December 2001 might be different from the value he/she would have used if he/she had retrieved the data in December 2000. The reason is that there are data revision that continuously update GDP, due to seasonal adjustment, new sources, changes in definitions such as base-year changes, etc. For the US, Croushore and Stark (2001, 2003) discuss datasets of real-time vintages that collect predictors' values exactly as they were available at each point in time. The ECB has a similar dataset for European data.

Clearly, the use of real-time vintages of predictors is an important step in the out-of-sample validation of models' forecasting ability. However, at the same time, the use of real-time data may introduce additional challenges in the evaluation of the forecasts. In particular, when evaluating forecasts in population, the variance adjustment has to be specifically tailored to the real-time nature of the data. See Clark and McCracken (2009c) for the details on implementing such adjustments, which will result in a correct implementation

[15] In particular, we refer to the versions of the Fluctuation, One-time Reversal and Fluctuation Rationality tests in Proposition 1 and 2 of Giacomini and Rossi (2010) and Proposition 6 in Rossi and Sekhposyan (2016), respectively. The technical assumptions require that, for example, the squared out-of-sample forecast error differences (in the case of a MSFE evaluation) should not have unit roots – see the cited papers for more details.

[16] In particular, the versions of the Fluctuation and Fluctuation Rationality tests in Algorithm 1 of Giacomini and Rossi (2010) and Theorem 5 in Rossi and Sekhposyan (2016), respectively, can be applied to all the cited models, although their practical implementation might be challenging due to the parameter estimation error component. A similar reasoning holds for the One-time Reversal test.

of the Fluctuation and Fluctuation Rationality tests. When evaluating forecasts in finite samples, no adjustment is required.

4 How to Improve Models' Forecasting Performance: Strategies for Forecasting in the Presence of Instabilities

Instabilities are known for being harmful to forecasting ability. Clements and Hendry (1999b, 2006) importantly argue that deterministic breaks, in particular location ones, are the most detrimental in terms of forecast accuracy.

There are two main estimation strategies to guard against instabilities. The first is to model instabilities explicitly. The second is to guard against instabilities by using additional "dimensions" of the data, such as enlarging the set of predictors or exploiting "external information", such as survey forecasts. In what follows, we consider each of these two options in detail and highlight their advantages and disadvantages.

4.1 Modeling Instabilities Explicitly: Estimation Methods Robust to Instabilities

One might think that, in the presence of instabilities, modeling the instability explicitly in the models' parameters might improve their forecasting performance. Note that, however, the examples in Section 2.5 show that this may not necessarily be the case. In fact, as the examples suggest, instabilities in models' forecasting performance are not necessarily equivalent to instabilities in models' parameters; therefore, by the same token, explicitly modeling instabilities in models' parameters may not necessarily improve their forecasting performance. In other words, while the main goal of the literature on time-varying parameters models is to provide the best in-sample fit, a good in-sample fit does not automatically guarantee good out-of-sample forecasting performance – see Clements and Hendry (1998, 1999a) on discrepancies between in-sample fit and out-of-sample forecasting performance (i.e. "forecast breakdowns") and Giacomini and Rossi (2009) for formal tests of forecast breakdowns.

More in detail, the reasons why a satisfactory in-sample fit may not translate into a satisfactory out-of-sample performance are discussed in Pesaran and Timmermann (2007), Clark and McCracken (2009b) and Giacomini and Rossi (2009), among others. Pesaran and Timmermann (2007) made the important, but often under-appreciated, point that, in the presence of structural breaks, including too distant information may increase the bias but reduce the forecast variance, while the opposite is true when discarding the oldest observations. As we discuss below, they formally show that it is not necessarily optimal to discard data before the breakpoint if the objective is to minimize the MSFE. This suggests that tests for structural breaks may not necessarily be the best approach for estimating the parameters when forecasting; rather, it is best to estimate the parameters directly targeting the out-of-sample performance. A similar point was made by Clark and McCracken (2009b), who note that identifying breaks and using the most recent data since the break should theoretically lead to good forecasts; however, practical difficulties in detecting breaks and their timing in finite samples may prevent such improvements. Giacomini and Rossi (2009) find that causes of forecast breakdowns include several kinds of estimation uncertainty and overfitting, in addition to parameter instabilities. Again, this implies that, if one wants to improve models' out-of-sample forecasting ability, the objective is not to fit the model with the highest likelihood (i.e. a measure of in-sample fit), but to find the model that minimizes the squared forecast errors (or, more generally, a measure of out-of-sample forecasting performance). In other words, if the forecasting ability is judged by the MSFE, even a correctly specified model might produce poor forecasts if the estimation results in a sufficiently high variance. Furthermore, one of the biggest difficulties in dealing with forecast instabilities is that they are easier to detect in hindsight than in real-time, at the end of the sample, when they are most useful.[17] This is the reason why, in this review, we focus on strategies for modeling instabilities with the explicit goal to improve the out-of-sample forecasting ability, and thus de-emphasize forecasts obtained after fitting models with breaks, non-linearities or time-varying parameters in-sample.[18] The latter, however, might offer common baseline models, and we discuss them at the end of this section. We will refer to the former approach as "out-of-sample, forecast-based" and to the latter as "in-sample, model fit-based".

4.1.1 Out-of-Sample Forecast-based Approaches to Modeling Instabilities

Forecasters typically guard against the presence of time variation in the parameters by re-estimating them each time a forecast is made. At each point in time, they use past observations within a certain window of data, potentially assigning different weights to recent/older observations according to a chosen "weight" function. Again, throughout this section, we assume that their objective is to minimize the expected squared forecast error.[19]

We illustrate these techniques in the Pesaran, Pick and Pranovich (2013) framework. Assume that the forecaster uses a linear model with $(N \times 1)$ exogenous predictors (x_t) to forecast a target variable h periods into the future (y_{t+h}). The model is:

$$y_{t+h} = \beta'_{t,h} x_t + \varepsilon_{t+h}, t = 1, 2, ..., T, \tag{11}$$

where $\beta_{t,h}$ is a $(N \times 1)$ vector of potentially time-varying parameters and ε_{t+h} are the unforecastable disturbances.[20] The parameters are estimated locally using a weighting scheme with possibly time-varying weights $\omega_{t,j;R}$ that may depend on time t, the initial observation

[17] See Andrews (2003) for end-of-sample instability tests. Again, the latter focuses on parameter instability rather than out-of-sample forecasting ability. For real-time procedures to monitor predictive ability, see Inoue and Rossi (2005) and Harvey, Leybourne, Sollis and Taylor (2019).

[18] For a classic review of in-sample estimation of models with structural breaks, see Stock (1986).

[19] The reason is that the majority of the theoretical results are derived for this case.

[20] Forecasting models typically include a constant, lags of the dependent variable as well as lags of the predictors. For simplicity, we will ignore the latter, although the general conclusions in the chapter do carry over to more general linear as well as nonlinear and more complicated structural model.

j and the window size used for estimation, R:

$$\widehat{\beta}_{t,h} = \left(\sum_{j=1}^{t-h} \omega_{t,j;R} x_j x_j' \right)^{-1} \left(\sum_{j=1}^{t-h} \omega_{t,j;R} x_j y_{j+h} \right). \qquad (12)$$

For example, by choosing $\omega_{t,j;R} = (1/R) \cdot 1\,(t-h-R+1 \leq j \leq t-h)$, a function that gives weight equal to $1/R$ to each of the past R observations, one obtains the "rolling" window estimation scheme: $\beta_{t,h;R} = \left(\sum_{j=t-h-R+1}^{t-h} x_t x_t' \right)^{-1} \left(\sum_{j=t-h-R+1}^{t-h} x_t y_{t+h} \right)$. The rolling estimation scheme estimates the parameter locally over a window of R observations, giving each observation the same weight. When $\omega_{t,j;R} = 1/\,(t-h)$, the estimation scheme is called "recursive" and all available past observations are used to estimate the parameters at each point in time, although more distant observations receive a lower weight.

Clearly, the choice of the weights is crucial, and the optimal choice depends on the (unknown) type of instability in the data. The literature has developed several techniques to choose the weights, either by deriving explicit formulas under specific assumptions or by developing robust approaches to choose them; we review each of these below.

Large, Discrete Breaks Clements and Hendry (2006) and Pesaran, Pick and Pranovich (2013) explicitly derive theoretical results in the presence of large, discrete and deterministic breaks for specific models. Ideally, if the break date were known, only observations after the most recent break should be used. In fact, Pesaran, Pick and Pranovich (2013) show that, ideally, the same weight should be given to observations within the same stable subsample/regime.[21] However, the break date is typically unknown. An alternative approach is intercept correction (Clements and Hendry, 1996, 1998, 1999b). Since large, discrete breaks introduce a bias in the forecast, intercept corrections adjust the model's forecast ($y_{t+h|t}$) with an estimate of the intercept based on the most recently estimated forecast error ($\widehat{\varepsilon}_{t|t-h}$). The forecast after the intercept correction is: $y^c_{t+h|t} = y_{t+h|t} + \widehat{\varepsilon}_{t|t-h}$.[22] The variance of the intercept-corrected forecast will be larger than that of the uncorrected forecast, in particular because the adjustment is estimated using only one observation. One possibility to decrease the variability is to estimate the model using all the data since the most recent break. However, the estimated break date may not be precisely estimated; also, if the post-break sample is small, the strategy of using only post-break data may not minimize the MSFE, as pointed out by Pesaran and Timmermann (2007). Thus, the latter recommend using observations both before and after the break, and provide explicit formulas to optimally weight the observations for specific models. They focus on the rolling estimation scheme, where is only one tuning parameter, the estimation window size R; thus, the challenge is to choose the optimal window size, R^*. They propose several methods to estimate R^*. A first method is analytical: under some assumptions,[23] it is possible to work out explicit formulas to choose an optimal window size to minimize the MSFE by managing the trade-off between bias and variance. Another method is cross validation, which reserves a fraction of the

[21] Their assumptions include strictly exogenous regressors and independent errors as well as $h = 1$.

[22] Hendry (2006) recommends double-differencing if the data are already in differences.

[23] Their framework assumes serially uncorrelated errors and strictly exogenous regressors, and sets $h = 1$.

most recent observations to select the window in a pseudo-out-of-sample exercise. Other methods involve estimating the parameters using all possible values of R and then average the forecasts using cross-validation or equal weights. In their Monte Carlo simulations, they find that combination methods perform well, especially when the break is small and difficult to detect. Pesaran, Pick and Pranovich (2013) instead suggest weighting across the regimes when the break date is not known.

Small, Continuous Breaks Models' parameters can change continuously and smoothly[24] instead of abruptly and discretely. In that case, a large literature (Holt, 1957; Brown, 1959; and Harvey 1989) recommends estimation strategies where all past observations are used but are down-weighted depending on how relevant they are for forecasting. The weight might be monotonic – that is, the farthest in the past, the less weight the observations receive. However, they need not be monotonic: if regimes repeat over time, it might be beneficial to give higher weight to observations that are further in the past if they belong to the regime that helps forecast the best. But how should the weights be chosen?

Giraitis, Kapetanios and Price (2013) propose a cross-validation method. Their model does not include regressors; that is, in eq. (12), x_t is the constant, and $h = 1$. To simplify notation, for the one-step-ahead forecast horizon, we let $\beta_{t,1;R}$ be denoted by $\beta_{t;R}$. Thus, in their model, the forecast is a simple mean of past observations,

$$\widehat{\beta}_{t;R} = \sum_{j=1}^{t-1} \omega_{t,j;R} y_{j+1}, \tag{13}$$

where $\omega_{t,j;R} \geq 0$, $\sum_{j=1}^{t-1} \omega_{t,j;R} = 1$, and the weights depend on a continuous and differentiable kernel, $K(u)$, such that $\omega_{t,j;R} = \frac{K(j/R)}{\sum_{s=1}^{t-1} K(s/R)}$. The simple constant rolling window approach weights equally the most recent R observations, i.e. the kernel is flat: $K(u) = 1\,(0 \leq u \leq 1)$. The exponentially weighted moving average approach (EWMA), instead, uses all the observations but increasingly down-weights the more distant ones, such that $K(u) = e^{-u}$ and, thus, $\omega_{t,j;R} = \frac{e^{-j/R}}{\sum_{s=1}^{t-1} e^{-s/R}}$; see Harvey (1989).[25] Triangular window weights are such that $K(u) = 2(1-u)\,1\,(0 \leq u \leq 1)$. Giraitis, Kapetanios and Price (2013) estimate R by minimizing the MSFE of all the forecasts.[26] Pesaran, Pick and Pranovich (2013) demonstrate that, in the presence of small and continuos breaks, the optimal weight is indeed EWMA.[27] In Monte Carlo simulations, Eklund, Kapetanios and Price (2010) also find that EWMA performs well when breaks are not deterministic.

Farmer, Schmidt and Timmermann (2019) consider the more general linear model with predictors in eq. (11). They focus on a smooth kernel that weights all observations, but progressively down-weights those further in the past. The estimator is thus a weighted least-square estimator, a generalization of eq. (12), where $\omega_{t,j;R}$ depends on the kernel weight,

[24] This kind of break is necessarily stochastic.

[25] Exponential smoothing (Holt, 1957; Brown, 1959) implies $\widehat{\beta}_{t;R} = \alpha \widehat{\beta}_{t-1;R} + (1-\alpha)\, y_t$.

[26] Generalizations based on Kalman filters were proposed by Hyndman et al. (2008).

[27] Their assumptions include strictly exogenous regressors and independent errors, as well as $h = 1$.

$K(.)$.[28] Alternatively, Hirano and Wright (2018) propose a cross-validation method to choose the window size at the end of the sample, in a situation where the instabilities captured by (12) are small enough that they cannot be detected with standard methods.

All the above papers rely on one-step-ahead forecasts and cannot be used for multi-step-ahead ones (i.e. $h > 1$).[29] An approach that can be used in the more general model, eq. (11), and for multi-step-ahead forecasts is Inoue, Jin and Rossi (2017). They focus on the flat kernel, and propose to estimate the window size R^* to minimize the approximate conditional MSFE in a situation where the parameters are modeled as smooth functions of time and the functional form is unknown.[30] Clark and McCracken (2009a) suggest instead combining rolling and recursive parameter estimates; their method can be used for multiple-step-ahead forecasts as well.

4.1.2 In-sample, Model Fit-based Approaches

A researcher might think that, if the source of instability is time-variation in the parameters, nonlinear or time-varying parameter models could offer a robust approach: we refer to this practice as the "in-sample, model fit"-based approach. This approach, that explicitly models non-linearities and time-variation in the parameters to fit models in-sample, hence markedly differs from the previously discussed "out-of-sample forecast-based" approach, whose goal is explicitly to produce good out-of-sample forecasts.

Examples of in-sample, model fit-based approaches to instabilities can be summarized by a general non-linear model (Teräsvirta, 2006):

$$y_{t+h} = \alpha x_t + \theta x_t G(\gamma, c; s_t) + \varepsilon_{t+h}, \tag{14}$$

where $\varepsilon_{t+h} \sim iid(0, \sigma^2)$ and $\sigma^2 = var(\varepsilon_{t+h})$.[31] Threshold, smooth-transition, Markov-switching and time-varying parameter models are special cases of the general model in eq. (14), corresponding to specific choices of the function $G(.)$.[32]

[28] In practice, they focus on the Epanechnikov kernel.

[29] Multiple-step ahead forecasts can be implemented using direct or iterated methods. Direct estimation means that h-period ahead forecasts are based on parameters estimated by regressing on predictors lagged h-periods. Iterated estimation means that h-period ahead forecasts are based on parameters estimated by regressing on predictors lagged one period and then iterating the procedure. Direct multi-step procedures are more robust than iterated under some DGP designs (Chevillon, 2016).

[30] Their framework is also more general since the error term and the regressors can be weakly dependent, and the regressors could potentially include both exogenous and lagged dependent variables.

[31] Specific assumptions on the error term, such as iid-ness, are often necessary in order to estimate such models. As a consequence, models are typically estimated for $h = 1$, then iterated to produce forecasts at longer horizons.

[32] For example: (i) Markov-switching models (Hamilton, 1989), where $G(\gamma, c; s_t) = 1(s_t = 1)$ and $s_t = \{0, 1\}$ is a regime indicator that follows a Markov chain. In this case, the parameter changes randomly between α and $\alpha + \theta$ between regimes. (ii) (Logistic) Smooth Transition Regression models (Bacon and Watts, 1971; Teräsvirta, 1994, 1998), where $G(\gamma, c; s_t) = (1 + \exp[-\gamma(s_t - c)])^{-1}$, and s_t is the transition variable. In this case, the parameter changes from α to $\alpha + \theta$ as a function of the transition variable. (iii) Switching Regression/Threshold models (Tong, 1990), where $G(\gamma, c; s_t) = 1(s_t \geq c)$. In this case, the parameter switches when s_t is above c. (iv) Time-varying Parameter models, where $\theta_t \equiv \theta G(\gamma, c; s_t)$ and θ_t can be either deterministic or stochastic – as an example of the latter: $\theta_t = \theta_{t-1} + \eta_t$, where the volatility

In the in-sample, model fit-based approach, the best nonlinear/in-sample model is first fit in-sample via maximum likelihood or Bayesian methods;[33] the model is subsequently validated with either in-sample predictive ability tests or out-of-sample forecasts. In-sample predictive ability tests evaluate whether the predictor is significant. Rewrite the general non-linear model in eq. (14) as follows (Teräsvirta, 2006, p. 418): $y_{t+h} = \beta_t x_t + \varepsilon_{t+h}$, where $\beta_t \equiv [\alpha + \theta G(\gamma, c; s_t)]$. In linear models, where $G(.) = 0$, the predictor's significance can be evaluated via a simple t-test on β, i.e. the Granger-causality test, as β is constant; in non-linear and time-varying parameter models, this requires testing $\beta_t = \beta = 0$; Rossi (2005) proposes such a robust Granger-causality specifically for models with instabilities. The latter tests differ from model specification tests, which evaluate whether the model is linear versus non-linear, or time-varying versus constant; in fact, as Example 4 clarified, testing for time-variation in the models' parameters (i.e. $\beta_t = \beta$) is not necessarily the same as testing for instability in the forecasting performance: structural break tests that detect instabilities in models' parameters do not necessarily detect instabilities in forecasting performance. In addition, as previously mentioned, models may overfit in-sample and result in poor out-of-sample forecasts, resulting in forecast breakdowns. See Giacomini and Rossi (2009) for forecast breakdown tests.[34]

More complicated models are available too. Pesaran, Pettenuzzo and Timmermann (2006) consider the case of stochastic breaks. They assume that, if a break happened in the past, it might likely happen again in the future, and explicitly model the break process allowing the information from past break dynamics to guide forecasts of future breaks. By modeling the break process itself, their method not only produces forecasts but also information about how many breaks are likely to occur in the future, how large they are, and when they are most likely to happen. Koop and Potter (2007) also estimate models with stochastic breaks for forecasting.[35]

can also be stochastic if $var(\varepsilon_{t+h}) = \sigma^2_{t+h}$. For simplicity of exposition, we have focused on models with one transition or one regime. The models described below may also impose restrictions on the parameters (e.g. $\gamma > 0$ in the smooth transition model) or focus on specific special cases (e.g. Hamilton, 1989, focuses on the case where x_t is the lagged value of y). The reader is referred to the cited references for more details and a discussion of more general models.

[33] This typically requires making distributional assumptions. For example, see D'Agostino, Gambetti and Giannone (2013) estimate models with time-varying parameters that, subsequently, perform well out-of-sample.

[34] They define surprise losses as the difference between the in-sample sum of squared residual and the out-of-sample squared forecast error, and propose tests for forecast breakdowns that evaluate whether the surprise loss is zero. Note that we have adapted the discussion to the quadratic loss function considered in this article as the main example – see Giacomini and Rossi (2009) for the applicability to more general loss functions.

[35] Among the Bayesian approaches, Giordani, Kohn and Van Dijk (2007) propose models with threshold, smooth transition and Markov-switching that include structural changes in the parameters as well as outliers. Models without regressors but with time-varying parameters are often referred to as unobservable component stochastic volatility models – see Koop and Potter (1998), Harvey (1989), Stock and Watson (2007) and Giordani and Villani (2010), among others, for applications and references.

4.2 Exploiting Additional Dimensions and Big Data

There exists a vast literature on methods whose goal is to improve models' forecasting ability by exploiting information from additional dimensions of the data, such as large datasets of predictors, frequencies, cross-sectional data, or, more generally, models – in short, "big data".[36] Since our focus is on the dangers of time variation in predictive ability, the omission of potentially important predictors is especially problematic when their ability to forecast shows up only in sub-samples, thus making it difficult to detect and exploit. As we mentioned, Timmermann (2008) refers to this situation as "pockets of predictability". Large dimensional models allow researchers to simultaneously use many predictors, thus avoiding missing important ones. Big data may also help in protecting against mis-specification more generally. Furthermore, by allowing parameters to be time-varying in addition to considering large dimensional models, researchers can also potentially follow the predictors' forecasting ability over time, including them at times in which they are useful and discarding them when they are not.

However, estimating models with many predictors comes at a cost. First, when jointly including all the variables, the OLS estimator in eq. (11) is $\left(\overline{\Sigma}_j x_j x_j'\right)^{-1}\left(\overline{\Sigma}_j x_j y_j'\right)$, where $\overline{\Sigma}_j$ denotes the sample average in the chosen window of observations up to time $t-h$. Hence, it involves inverting $\left(\overline{\Sigma}_j x_j x_j'\right)$, which might simply be infeasible when the number of predictors N is large. In fact, when N is much larger than T, one cannot invert $\left(\overline{\Sigma}_j x_j x_j'\right)$ altogether. In addition, multicollinearity among the predictors becomes more likely the larger the number of predictors,[37] and a high number of lags might worsen the problem as well.[38] Furthermore, in finite samples, jointly including all the variables has a cost in terms of forecasting, as it implies less precise estimates. In fact, recall that the MSFE can be decomposed into the sum of (squared) bias and variance: by estimating models with many predictors, one minimizes the mis-specification bias from omitting potentially important predictors – at the same time, however, one generates a high variance due to the larger parameter estimation error incurred while estimating large models. Aggregation and dimensionality reduction are key to prevent parameter proliferation from negatively affecting the forecasting process, and the choice of how to perform the aggregation process is crucial. Again, we will direct our attention mainly to the strand of the literature explicitly focusing on forecasting in the presence of instabilities. Aggregation can be performed before, during or after the forecasting process; hence, we will distinguish among the following strategies in the next sub-sections: (a) "aggregate then forecast"; (b) "forecast while aggregating"; (c) "forecast (the disaggregates) then aggregate". Note that aggregation does not need to necessarily be among large datasets of predictors, or models, but also across frequencies, as in the MIDAS approach (Andreou, Ghysels and Kourtellos, 2013) – we consider such cases in Section 4.2.4.

[36] In the case of structural models, "big data" means either a large dimensional structural model or a large number of models.

[37] In fact, including more predictors increases the possibility that the additional predictors might contain similar (or highly correlated) information to that in the predictors that are already included in the model.

[38] For example, in a VAR with N dependent variables and p lags, the number of parameters is at least N^2p.

4.2.1 "Aggregate then Forecast": (Unsupervised) Factor Models

The strategy of "aggregating then forecasting" when dealing with many predictors can be summarized as follows: first, summarize the information contained in a large dataset of predictors in a parsimonious number of indices, thus reducing the dimensionality; then, use the indices for forecasting. One common way to aggregate information in a large-dimensional dataset of predictors is to use factor models:

$$\begin{aligned} x_{i,t} &= \lambda_i' f_t + \eta_{i,t} \qquad (15) \\ y_{t+h} &= \beta' f_t + v_{i,t}, \qquad (16) \end{aligned}$$

where f_t is an $(r \times 1)$ vector, $\eta_{i,t}, v_{i,t}$ are uncorrelated disturbances and r is much smaller than N.[39] Factor models summarize the information in a large cross-section of predictors (the $x's$) in a few principal components (the $f's$), which explain the largest amount of the variability of the predictors.

An important parameter to choose when estimating factor models is how many factors to use: the larger the number of factors, the lower the bias but, also, the lower the benefits from the dimensionality reduction. In fact, using all the factors is equivalent to OLS using all the predictors.[40] Typically, the number of factors is selected in order to explain a large portion of the variance of the predictors while excluding irrelevant factors – see Bai and Ng (2002) for widely-used information criteria to correctly estimate r.

Modeling instabilities could potentially be important, and one may want to give more weight to predictors at times in which they forecast well. Most of the literature focuses on in-sample approaches, although the next section discusses recent methodologies that focus on forecasting.[41]

4.2.2 "Forecast while Aggregating": Big Data, Model Selection and Shrinkage Methods

An alternative approach to that described in the previous subsection is to use a large number of predictors (i.e. "big data") directly in the forecasting procedure (rather than aggregating the predictors prior to forecasting): we will refer to this approach as "Forecast while Aggregating". This is an emerging field: while we focus on applications in forecasting with

[39] Many of the methods discussed in this section perform better if the regressors are orthogonalized; we therefore assume in this section that the predictors are orthogonalized and that a constant is included in eq. (16). See Stock and Watson (2002) and Doz, Giannone and Reichlin (2012) on forecasting with factor models.

[40] This follows from the fact that the factors are rotations of the predictors.

[41] A large literature proposes in-sample methodologies to allow for time variation in the parameters of the factor models. Breitung and Eickmeier (2011), Corradi and Swanson (2014), Han and Inoue (2015), Chen, Dolado and Gonzalo (2014) and Cheng, Liao and Schorfheide (2016) develop tests for instabilities in factor models, and Stock and Watson (2012) analyze the stability of factor models around the Great Recession. Corradi and Swanson (2014) disentangle forecast failure of factor-augmented models into instability in factor loading and instability in regression coefficients, and propose techniques to jointly test for both. Bates, Plagborg-Moller, Stock and Watson (2013) derive theoretical results on the size and magnitude of instabilities that induce inconsistency in parameter estimates in factor models.

instabilities in economics and finance, there are many uses of big data in fields such as population dynamics, crime, energy, media, environmental and biomedical sciences – see Hassani and Silva (2015), among others, for examples.

The relationship between the "forecast while aggregating" approach and the "aggregate then forecast" approach described in the previous sub-section is as follows. Traditional factor models assume that each predictor is potentially correlated with the factors; the dimension reduction is achieved by selecting only a small number of factors. Since the factors are typically selected with the objective to explain the largest portion of the variance of the predictors in eq. (15), the information is not directly extracted with the goal of forecasting the target variable – namely y_{t+h} in eq. (16). In other words, when forecasting output or inflation, the selected factor would be the same. In addition, the model may or may not have an underlying factor structure, in which case traditional information criteria would select the wrong number of factors. The alternative way to aggregate information in the "big data" literature is to use all available data while performing shrinkage – that is, replacing selected parameters with smaller, or even zero, values – to handle the large dimensionality of the data, where the amount of shrinkage is directly chosen to lower the MSFE. The estimate of β may thus be biased, but the variance would decrease because of the reduction in dimensionality; when the shrinkage is done appropriately to trade-off bias and variance, the resulting estimator has a lower MSFE than OLS.[42]

There are a variety of methods available to perform shrinkage; often, such methods are based on machine learning. In forecasting, machine learning refers to automated predictive algorithms, especially in out-of-sample contexts, typically dealing with a large number of models and predictors in complex environments.[43] Ridge (Hoerl and Kennard, 1970), lasso (Tibshirani, 1996), elastic nets (Zou and Hastie, 2005; Zou and Zhang, 2009) and neural networks (White, 1992; Swanson and White 1997a,b) are all shrinkage methods that have machine learning features.[44] The use of big data in forecasting economic time series, pioneered by Swanson and White (1995, 1997a,b), raises important questions, such as whether including hundreds of predictors in a predictive regression or a Vector Autoregression (VAR) using automated "machine learning" methods improves forecasts, and, in particular, whether it may result in less evidence of forecast instabilities or breakdowns. At the same time, there are several challenges when forecasting with big data: (i) predictors may appear significant just by chance if the researcher tests many specifications without taking into account the search across specifications;[45] (ii) the chances of overfitting the data might be higher, given the larger number of predictors; (iii) the large number of predictors might create more noise and more signal distortions than when the numbers of predictors is small (Banbura and Modugno, 2014).[46] Although, in what follows, we will review several machine-learning tech-

[42] As previously noted, the MSFE equals bias squared plus variance.

[43] For example, textual analyses – see Choi and Varian (2012) for an early use of trends in google searches for particular words when forecasting.

[44] Other machine learning techniques include boosting (Schapire and Freund, 2012), bagging (Breiman, 1996; Inoue and Kilian, 2008), adaptive Lasso (Zou, 2006) and non-parametric regression trees and forests (Breiman, Friedman, Stone and Olshen, 1984; Breiman, 2001). Several of the techniques we review in the next section also have machine learning features.

[45] See sub-section 4.2.5 for a discussion of this issue.

[46] A possibility is to filter out the noise first (e.g. Hassani et al., 2009, 2013), while however being vigi-

niques among the most-widely used in economics and finance, a comprehensive discussion is beyond the scope of this article – see e.g. Hastie, Tibshirani and Friedman (2016) for a thorough overview.

Model selection assigns weights that are either zero or one to each of the OLS coefficients. Thus, model selection induces sparsity, in the sense that only some predictors are used and others are discarded. A traditional way to discard predictors is via multiple testing, i.e. by starting with the largest possible set of predictors and then discarding the unimportant ones via repeated testing procedures. It then becomes important to control the probability of erroneously finding *at least* one predictor.[47] While this approach might be feasible in small dimensions, it presents difficulties when a large number of tests needs to be performed. A more sophisticated approach (Bai and Ng's, 2008, "targeted predictors") selects a subset of predictors based on the outcome of individual t-tests, then extracts principal components from this subset.[48] Model selection can be automated in adaptive ways (see Swanson and White, 1995, 1997a,b, for a pioneering example in economics). Another way to perform model selection is via information criteria that select factors with the purpose of directly targeting improvements in models' forecasting ability. Information criteria minimize the (log) sum of squared residuals from eq. (16) while penalizing less parsimonious models, i.e. models that include too many factors. This way, the choice of factors is "supervised". Traditional information criteria can be used, such as AIC or BIC (Stock and Watson, 2002). Carrasco and Rossi (2016) propose information criteria based on generalized cross-validation and Mallows (1973), and show that they improve forecasts in the presence of structural breaks. While AIC and BIC assume a factor structure, cross validation and Mallows' (1973) criteria are valid no matter whether the large dataset of predictors can be well-approximated by a factor model or not.[49]

Shrinkage methods, on the other hand, apply weights to OLS coefficients that are not necessarily zero or one. Shrinkage can be imposed either in a frequentist or in a Bayesian way. Frequentist shrinkage methods include, among others: ridge (Hoerl and Kennard, 1970); lasso (Tibshirani, 1996); and elastic nets (Zou and Hastie, 2005; Zou and Zhang, 2009). A useful representation of these estimators is as penalized OLS estimators:

$$\widehat{\beta}_t = \arg\min_{\beta} \overline{\Sigma}_j \left(y_{j+h} - \beta'_{t,h} x_j \right)^2 + g\left(\beta\right), \tag{17}$$

where the penalty function $g\left(\beta\right)$ equals $\gamma \sum_{i=1}^{N} \left(\beta_i - 0\right)^2$ for ridge;[50] $\gamma \sum_{i=1}^{N} \left|\beta_i - 0\right|$ for lasso; and $\gamma \sum_{i=1}^{N} \left(\beta_i - 0\right)^2 + \left(1-\gamma\right) \sum_{i=1}^{N} \left|\beta_i - 0\right|$ for elastic nets. It follows that ridge shrinks all coefficients towards zero by a similar amount, while not setting any of them exactly to zero. Therefore, ridge imposes shrinkage but does not perform model selection. Lasso, instead,

lant that the filtering is done in real-time to avoid contaminating the out-of-sample forecasting exercise by including future data in the estimation.

[47] Procedures that control the probability of making at least one false discovery (such as Bonferroni) are said to control the family-wise error rate – see also Section 4.2.5 below.

[48] Giovannelli and Proietti (2015) instead choose the factors that best correlate with the target variable, after controlling for the error rate.

[49] See Carrasco and Rossi (2016) for a comparison of information criteria for forecasting purposes.

[50] Note that the sample mean corresponds to $\gamma \to \infty$ (when a constant is included) and OLS to $\gamma \to 0$.

combines the features of both model selection and shrinkage, as it shrinks some coefficients to zero and sets others to zero, thus creating sparsity.

An alternative representation where the forecasts are viewed as special cases of weighted principal components (as in Carrasco and Rossi, 2016) provides an illuminating way to understand the relationship among some of these estimators. The information in the second moments of the N predictors can be summarized by its eigenvalues $\hat{\lambda}_i^2$ and eigenfunctions $\hat{\psi}_i$: $\left(\Sigma_j x_t x_t'\right) \hat{\psi}_i = \hat{\lambda}_i^2 \hat{\psi}_i$, $i = 1, 2, ...N$. The forecast, written as a function of a weighted average of the eigenfunctions (instead of the predictors), is:

$$y_{t+h|t} = \frac{1}{T} \sum_{i=1}^{\min(N,T)} \hat{q}_i \hat{\psi}_i \hat{\psi}_i' y, \tag{18}$$

where y is the vector of in-sample $y_t's$ and $\hat{q}_i \equiv q\left(\alpha, \hat{\lambda}_i^2\right)$. Note that $\hat{q}_i = 1$ corresponds to forecasts obtained using the usual OLS estimator with all the predictors. Ridge is a special case when $\hat{q}_i = \frac{\lambda_j^2}{\lambda_j^2 + \alpha}$. Factor models assume $\hat{q}_i = I\left(\lambda_i^2 \geq \alpha\right)$. Landweber and Friedman (see Carrasco, Florens and Renault, 2007) corresponds to $\hat{q}_i = 1 - \left(1 - d\lambda_j^2\right)^{1/\alpha}$.[51] In factor models with r principal components, $\hat{q}_i = I\left(i \leq r\right)$.

In the Bayesian approach, shrinkage is achieved by using priors in Bayesian VARs to control parameter proliferation. Computational issues can also be resolved by choosing an appropriate prior structure. Banbura, Giannone and Reichlin (2010) are among the first to attempt to use large dimensional Bayesian VARs, while Koop (2013a) focuses on investigating the role played by the priors. There is an interesting connection between ridge, lasso and elastic nets, on the one hand, and Bayesian shrinkage on the other. Under some assumptions,[52] the ridge estimator is the posterior mode using a Gaussian prior (De Mol, Giannone and Reichlin, 2008). This result provides an interesting interpretation in factor models: if the prior has the same variance on all the $\beta's$ in eq. (11), that is uniform shrinkage on all the $\beta's$, then the prior on the principal components version of the model is proportional to the eigenvalues. Thus, the prior shrinks more the most unimportant factors even if the prior on the $\beta's$ assigns equal weight to each of them. Under the same assumptions, the lasso estimator is the posterior mode using a Laplace prior.[53] The elastic net estimator is the posterior mode obtained using a combination of a Gaussian and Laplace priors.

It is also possible to induce parsimony in models with factors as well, for example by allowing the factor loading matrices to be sparse (i.e. have zeros), resulting in so-called sparse factor models. Kim and Swanson (2014) comprehensively compare the performance of sparse factor models with traditional (unsupervised) factor models, and find that the former

[51] α and d are tuning parameters – see Carrasco and Rossi (2016) for practical suggestions on how to choose them. For partial least squares with r factors, the formula for $\hat{q}_i$ is provided in Carrasco and Rossi (2016, eq. 13).

[52] The assumptions include i.i.d., Gaussian data and a linear model. Note that eq. (11) for $h = 1$ is one equation of a VAR system, where x_t incldues the same number of lags for both y_t and x_t.

[53] Note that, in the Bayesian approach, the posterior mode is sparse while the posterior itself is never sparse.

are better at forecasting at short horizons, while the latter are better at longer horizons.[54] As they argue, this result might be due to the fact that traditional factor models could be more robust to structural changes since they always use all the predictors, while sparse models would perform well only if their coefficients swiftly adapt to structural changes. In this sense, traditional (unsupervised) factor models are akin to forecast combinations, which we discuss in the next sub-section and which have been shown to perform very well in forecasting.[55]

Yet another possibility is to use both a large dimension of predictors as well as time-varying parameters, where, typically, the emphasis is on letting the volatility to also change over time. The latter becomes especially advantageous when modeling predictive densities. We will therefore tackle this issue in Section 5.2.2.

Finally, note that, on the one hand, the shrinkage methods we discussed have "machine learning" features to them; on the other hand, they all have different properties, which we have overviewed. However, it is important to notice that the theoretical properties of many of these methods are really not known in a time-series context, which is in the context that is relevant for forecasting – let alone in the context of forecasting in the presence of instabilities. A large component of the current theory has been developed for independent observations – and a large part of it under the assumption of orthogonal regressors. While this section has reviewed what is currently known, this is an area in need of theoretical analyses and will most likely see important developments in the near future.

4.2.3 "Forecast then Aggregate": Forecast Combination

Forecasters face substantial uncertainty regarding which model to use in practice, which means that they most likely end up combining a variety of forecasting models. The strategy of "forecasting then aggregating" means using the models one-at-a-time to produce a set of forecasts, then aggregating (combining) the forecasts via a weighted average:

$$y_{t+h|t} = \sum_{m=1}^{M} \omega_{m,h} y_{t+h|t}^{(m)},\ \omega_{m,h} \in [0,1]\,,\ \sum_{m=1}^{M} \omega_{m,h} = 1 \tag{19}$$

where $y_{t+h|t}^{(m)}$ is the forecast of model m, $m = 1, 2, ..., M$, $\omega_{m,h}$ is the weight (between zero and one) associated to the forecast of model m, and the forecast is obtained by re-estimating the parameters over time. One common way to implement forecast combinations in a dataset with N predictors is to include them one-at-a-time, that is, $y_{t+h} = \beta_i x_{i,t}$, for $i = 1, 2, ..., N$; see Stock and Watson (2003).[56] Relative to Bayesian VARs with time-varying parameters, forecast combinations are relatively straightforward and easy to implement. The weights can be chosen to be either: (i) equal across models ($\omega_{m,h} = 1/M$, as in Bates and Granger, 1969, and, more recently, Stock and Watson, 2003); (ii) trimmed ($\omega_{m,h} = 0$ for models with

[54] They also consider independent component analysis, which constructs factors with non-Gaussian distributions.

[55] Groen and Kapetanios (2016) compare factor models, ridge and PLS when the model has a weak factor structure. Kelly and Pruitt (2015) propose instead to use proxies for forecasting. See also Kim and Swanson (2015, 2018) and Swanson, Xiong and Yang (2020) for other big data approaches to forecasting.

[56] Again, models may include a constant and lags.

past MSFEs above a certain threshold, and the rest of the weights equal to one divided by the number of remaining models); or estimated by either: (iii) the inverse of the past MSFE of each model (as in Stock and Watson, 2004, and Timmermann, 2006); (iv) each model's posterior probability, as in Bayesian model averaging (Wright, 2008, 2009); (v) minimizing the Mallows (1973) and cross-validation criteria in factor-augmented VAR models (Cheng and Hansen, 2015); (vi) the probability that the squared forecast error from the alternative model is smaller than the benchmark (Granziera and Sekhposyan, 2019). As an alternative, Elliott, Gargano and Timmermann (2013) propose estimating the weights $\omega_{m,h}$ using complete subset regressions; that is, first combining forecasts from all possible linear regression models that have the same number of predictors using equal weight; then, in a second step, using an optimal procedure to determine the number of predictors. The latter provide an indication of model complexity and can be selected to trade-off bias and variance of the forecast errors. Complete subset regressions impose different degrees of shrinkage on each predictor.[57] Koop, Korobilis and Pettenuzzo (2019) instead propose randomly compressing the predictors first to shrink their dimension and produce forecasts, then aggregating the forecasts using Bayesian model averaging.[58]

Instead of weighting each model with a weight between zero and one, each model can also be either included or excluded using weights that are either zero or one: $\omega_{m,h} = \{0, 1\}$. This is the approach taken by Groen, Paap and Ravazzolo (2013).[59] They find that there is substantial uncertainty in the model specification and that allowing for breaks results in a much smaller number of predictors in the individual regressions. Thus, large-dimensional models can be thought of as a substitute for time-varying parameters, and time-variation in the parameters might be due to the model mis-specification incurred by a small model. They also find that more parsimonious models forecast (inflation) better at longer horizons.

Survey forecasts are themselves forecast combinations, as typically one uses the average across a cross-section of forecasters, each of which, in turn, may be obtained using different models. Survey datasets, however, may be available only for selected variables or countries. Importantly, survey-based forecasts can be, and often are, used as one of the models to combine. In fact, Faust and Wright (2013) and Ang, Bekaert and Wei (2007), among others, demonstrate that survey forecasts are very competitive and can enhance combined forecasts.

Time-variation can be built explicitly in the forecast combinations as well using time-varying weights: $y_{t+h|t} = \sum\limits_{m=1}^{M} \omega_{m,t,h} y_{t+h|t}^{(m)}$, where the inclusion of each model is itself time-varying. Elliott and Timmermann (2005) let the weights be driven by a regime-switching process in a latent state variable. Hoogerheide, Kleijn, Ravazzolo, van Dijk and Verbeek (2009) focus on the case of smooth transition across models over time by letting the time-

[57] Interestingly, ridge is a special case of complete subset regressions.

[58] That is, they create several "models" m, $m = 1, ..., M$, where $x_t^{(m)} = \Psi^{(m)} x_t$ and $\Psi^{(m)}$ is an $(r \times N)$ non-estimated matrix with coefficients randomly simulated. Forecasts $y_{t+h|t}^{(m)}$ are then generated using $x_t^{(m)}$, and then aggregated as in eq. (19) using Bayesian model averaging.

[59] In addition, in their framework, the forecasts of each model are obtained by letting the parameters to be possibly time-varying, $y_{t+h|t,m} = \beta_{t,m}' x_{t,m}$, and the parameter evolution follows a random walk when the parameter changes. The parameter is allowed to change at each point in time, although it does not have to. They also allow the variances of each model to change over time.

varying weights $\omega_{m,t,h}$ follow a random walk.[60] Koop and Korobilis (2012) instead allow for swift changes across models over time in addition to smooth time-variation; to avoid parameter proliferation, they use forgetting factors, which operate like an exponential smoothing.[61] Guerron-Quintana and Zhong (2018) instead propose a clustering approach based on machine learning techniques where the data are first divided in sub-samples (or clusters), then the forecasts are adjusted using forecast errors from past clusters that are most similar to recent observations. A similar idea is followed by Dendramis, Kapetanios and Marcellino (2019), who cluster past data which are the best match for the current economic conditions and, hence, could be more informative for forecasting.

4.2.4 Other Dimensions

The above discussion focused on using large datasets of models/predictors to guard against instabilities. There are alternative dimensions that can be exploited as well. One such important dimension is mixed frequencies. For data available at different frequencies (such as daily and monthly, for example), the "aggregate then forecast" strategy involves first aggregating the data at the lowest common frequency (monthly, in the example), then forecasting the target variable at the lowest frequency (e.g. Stock Watson, 2003). Alternatively, in the spirit of the "forecasting while aggregating" approach, one could use both high and low frequency variables directly to forecast. MIDAS models are frequently used to produce forecasts when the predictor is sampled at a frequency higher than the target variable (Andreou, Ghysels and Kourtellos, 2013). To allow the predictive ability to change over time across regimes, Galvao (2013) proposes a smooth transition MIDAS model that also allows to take into account nonlinearities; she finds gains when forecasting output growth using high-frequency financial variables. Carriero, Clark and Marcellino (2015) use Bayesian regressions with stochastic volatility and mixed frequencies to improve nowcasts of GDP growth, and show that stochastic volatility is an essential ingredient. Alternatives to MIDAS models include jointly modeling variables at different frequencies via a Kalman filter approach: this is the strategy pursued in the nowcasting literature – see Banbura, Giannone, Modugno and Reichlin (2013).

Forecast combinations can also be used to robustify forecasts over the choice of the estimation window size. Pesaran and Timmermann (2007) propose to combine forecasts based on many window sizes; the combination can be obtained either by weighting the forecasts of each window size by the inverse of its past MSFE or by using equal weights. Their "average" forecast at time T is simply: $y_{t+h|t} = \sum_{r=\underline{R}}^{\overline{R}} \omega_{r,t,h} y_{t+h|t}^{(r)}$, where $\omega_{r,t,h}$ are the weights, and $y_{t+h|t}^{(r)}$ is the forecast based on window size r, where r ranges from a minimum ($\underline{R}$) to a maximum ($\overline{R}$). The latter method turns out to perform empirically well (Rossi, 2013).[62]

[60] See also Ravazzolo, Verbeek and Van Dijk (2007) for forecast combinations under time-varying weights.

[61] Both Hoogerheide, Kleijn, Ravazzolo, van Dijk and Verbeek (2009) and Koop and Korobilis (2012) allow for breaks in models' parameters as well as the variance of the overall combination.

[62] See also Hubrich and Hendry (2011) for the interplay between structural breaks and disaggregate data.

4.2.5 The Dangers of Data Snooping

When considering a large number of predictors, the dangers of multiple testing and data snooping become particularly relevant. According to White (2000): "Data snooping occurs when a given set of data is used more than once for purposes of inference or model selection. When such data reuse occurs, there is always the possibility that any satisfactory result obtained may simply be due to chance rather than any merit inherent in the method yielding the results." Here is an illuminating example of how data snooping may generate inexistent predictive ability. The example is taken from White (2000).

Example 5 (Data Snooping) *Consider the following newsletter scam: the forecaster "selects a large number of individuals to receive a free copy of a stock market newsletter; to half the group one predicts the market will go up next week; to the other, that the market will go down. The next week, (the forecaster) sends the free newsletter only to those who received the correct prediction; again, half are told the market will go up and half down. The process is repeated (... and) after several months (there is a) group of people who received the perfect prediction and is willing to pay for such "good" forecasts". (White, 2000) Clearly, the forecaster did not identify the true model that generates stock market returns, and any forecasts he produces are random; hence, the predictive ability of his "model" is a fluke.*

While out-of-sample forecasts guard to some extent against data snooping, as they avoid evaluating models in-sample, still, if many models are repeatedly evaluated using the same dataset, the empirical findings may not be robust to data snooping. White (2000) and Hansen (2005) propose bootstrap procedures to protect against the dangers of data snooping when evaluating the best performing models (against a benchmark) within a large set of competing models. Sullivan, Timmermann and White (1999) evaluate the performance of trading rules when forecasting the stock market, and conclude that the predictability from the best trading rules did disappear over time. Similarly, Sullivan, Timmermann and White (2001) find no predictability from calendar effects in stock returns once data snooping has been taken into account. Hansen, Lunde and Nason (2011) propose ways to construct confidence sets of the best forecasting models when considering a large set of them.

5 Measures of Forecast Uncertainty and Predictive Densities in Unstable Environments

Predictions are typically communicated via "point forecasts"; that is, by reporting the expected, average value of the future target variable. However, the uncertainty around point forecasts is crucial. In fact, in a highly uncertain environment, forecasts may be far away from their target and could be highly unreliable. Measures of uncertainty around forecasts – such as confidence intervals, quantiles or density forecasts – are more informative and, at the same time, provide hedging in an uncertain world.[63] After all, as Neils Bohr reportedly said, "it is very difficult to predict — especially the future."

[63] For a basic introduction to density forecasts in economics and policymaking, see Rossi (2014b).

Several central banks routinely communicate measures of "confidence intervals" around their predictions via fan charts.[64] Fan charts depict percentiles of the forecast distribution over a sequence of forecast horizons. In general, central banks' fan charts are the result of convoluted methodologies that involve a variety of models and subjective assessments, although fan charts can be based on specific models as well. For example, a model that has been used by the Bank of England is the split-normal distribution, a non-symmetric distribution that is completely characterized by three parameters, thus reducing the subjective assessment to only a few parameters and avoiding the necessity that policymakers agree on all the forecast percentiles. For example, the skewness summarizes the monetary policy committee's assessment of the balance of risk on upside/downside uncertainty (Wallis, 1999, 2004). Other institutions provide measures of predictive uncertainty based on the RMSFEs of historical forecasts.[65]

There are two main ways to report forecast uncertainty: one is to report uncertainty around a forecast over time, for a given forecast horizon (forecast confidence intervals); the second is to report uncertainty around a sequence of forecasts across horizons, fixing the point in time in which the forecast is made (fan charts). Both measures of uncertainty can, in principle, be obtained as percentiles from a sequence of forecast densities or, alternatively, as quantiles or forecast confidence intervals. A predictive density is the conditional distribution of the target variable, say y_{t+h}, given a conditioning set of variables, say x_t, and will be denoted by $\phi_{t+h|t} \equiv \Pr\left(y_{t+h}|x_t\right)$. Predictive densities provide a complete description of the uncertainty associated with a forecast. They can be obtained from parametric models or non-parametrically (e.g. survey density forecasts).

Note, however, that it is not obvious how one should construct fan charts based on the percentiles of the predictive density. Symmetric percentiles may only be appropriate under special assumptions (Diebold, Gunther and Tay, 1998; Mitchell and Weale, 2019). For example, some central banks[66] report the "best critical region", i.e. the interval of shortest length with a given target coverage rate, which differs from the central interval unless the predictive density is symmetric and uni-modal. Predictive densities can also be obtained from best critical regions. Mitchell and Weale (2019) show that best critical regions result in censored predictive densities where no probability is defined on the outer tails. The latter might induce robustness against instabilities and extreme events that cannot be quantified, and could be related to Knightian uncertainty (see Rossi, Sekhposyan and Soupre, 2018, and Galvao and Mitchell, 2019, for attempts to link Knightian uncertainty and predictive densities).

How to correctly quantify forecast uncertainty in the presence of instabilities? And how to evaluate these measures of uncertainty? This section addresses both of these two issues.

[64] E.g., the Bank of England Inflation Report, the Economic Bulletin of the Bank of Italy, and the publications by the Bank of Canada, the Reserve Bank of Australia and the European Central Bank

[65] For example the US FOMC Summary of Economic Projections, which calculates them in rolling windows over the previous twenty years.

[66] E.g. the Bank of England.

5.1 Density Forecast Evaluation

When attempting to evaluate density forecasts,[67] an important issue is how to evaluate whether they are correctly specified (calibrated). Testing for the correct specification of a predictive density means understanding whether the description of uncertainty provided by the forecast model is accurate given the ex-post realizations. Diebold, Gunther and Tay (1998), Corradi and Swanson (2006a,b,c) and Rossi and Sekhposyan (2019) develop methodologies based on the result that, if densities are correctly calibrated, the areas below each predictive density up to the subsequent realization (the Probability Integral Transforms, or PITs) should be a sequence of i.i.d. Uniform observations (Diebold, Gunther and Tay, 1998). Thus, the latter papers test properties of the PITs, such as independence and uniformity, which imply the correct specification of the predictive density. This ensures that the percentage of realizations in each quantile of the predictive density corresponds to the probability assigned by the model. For example, uniformity follows from the fact that, in the 5-th quantile of the density, one should observe about 5% of the realizations, ex-post. The papers differ depending on whether and how parameter estimation error is taken into account when evaluating the forecasts: in the pioneering approach by Diebold, Gunther and Tay (1998) and Diebold, Tay and Wallis (1999), parameter estimation error is ignored, while it is taken into account in Corradi and Swanson (2006b,c) and Rossi and Sekhposyan (2019), who use expanding and fixed parameter estimation windows, respectively. Alternative approaches include tests on raw moments of the distribution, such as Berkowitz's (2001) likelihood ratio and Knueppel's (2015) GMM. The trade-off between PIT-based and raw-moments-based approaches is that the former jointly test for the correct specification of the whole predictive density (hence, all the moments at the same time), while the latter focus on a selected subset of the moments – if the researcher does not select all the relevant moments, the raw-moments-based tests suffer from mis-specification, while they are more powerful than PIT-based tests if the correct subset of moments is selected. Importantly, Mitchell and Wallis (2011) emphasize that models that correctly condition on an incomplete information set – and hence might be poor forecasting models – would still have uniform PITs. Thus, they propose using additional criteria beyond the uniformity of the PITs, such as checks of serial correlation of the PITs.[68]

However, the presence of instabilities is potentially problematic in all the tests above, which are not robust. Rossi and Sekhposyan (2013) develop PIT-based tests robust to instabilities, which we use to evaluate the correct calibration of GDP growth forecasts from the SPF (described in Section 2.4). Their κ_P test statistic is 2.7907 and is significant at the 5% level, thus highlighting strong evidence in favor of lack of correct calibration. The instability detected by the test is in 2008:1, exactly in the middle of the financial crisis.[69]

If SPF forecasts do not predict the recession well, are there other econometrics models that can? And which ones? We consider a Bayesian VAR model with stochastic volatility

[67] We will use the terms predictive densities and forecast densities interchangeably,

[68] They also recommend KLIC-based measures as well as correlations of log-score differences with external explanatory variables – see Section 5.2.1 for details on the latter.

[69] If one implemented tests ignoring potential instabilities, such as Rossi and Sekhposyan (2019), in this example one would still find evidence of incorrect calibration: the value of the Rossi and Sekhposyan (2019) test statistics is 1.7041 and its critical value is 0.7258.

(Clark and Ravazzolo, 2015) and compare it to the SPF predictive density.[70] We evaluate their relative forecasting performance using Giacomini and Rossi's (2010) Fluctuation test. To implement the Fluctuation test, one needs to choose a loss function, such as the Continuous Rank Probability Score (CRPS), the log score or quantile scores (Manzan, 2015); here we use the CRPS – see Gneiting and Raftery (2007) for a discussion of scoring rules.[71] The results are reported in Figure 9, which depicts the difference between the SPF and the Bayesian VAR's performance: negative values indicate that the SPF performs best. The results indicate that, in the early 2000s, the SPF and the Bayesian VAR performed similarly; however, closer to the financial crisis in 2007-2008 and in 2015, the SPF starts forecasting better than the Bayesian VAR, and its relative performance improves over time.[72]

INSERT FIGURE 9 HERE

5.2 Strategies for Improving Predictive Density Performance in the Presence of Instabilities

Similarly to the case of point forecasts, strategies for improving density forecasts in the presence of instabilities include either using large datasets of predictors or modeling instabilities explicitly. We will consider each of these strategies in turn in what follows.

5.2.1 Exploiting Additional Dimensions and Big Data

Combining densities from a large set of models or predictors is one way to guard against instabilities in density forecasts.[73] The combined predictive density ($\phi_{t+h|t}$) is obtained from aggregating each of the M individual models' predictive densities ($\phi_{t+h|t}^{(m)}$, $m = 1, 2, .., M$):

$$\phi_{t+h|t} = \sum_{m=1}^{M} w_{m,t,h} \phi_{t+h|t}^{(m)}, \tag{20}$$

[70] We consider fixed-horizon SPF predictive densities constructed using optimal weights that target correct specification; see Ganics, Rossi and Sekhposyan (2019). The predictive density is smoothed using a Gaussian distribution.

[71] The MSFE is only appropriate for point forecasts. Density forecasts are instead evaluated by scoring rules. In the log score case, the performance of a model is measured by the logarithm of the predictive density evaluated at the actual realization; thus, a higher log score implies a better predictive density, as the best forecast density associates the highest ex-ante probability to the value that indeed realized ex-post. The CRPS is instead the average (quadratic) distance between the cumulative distribution functions of the predictive density and that of the perfect forecast – a step function equal to zero for values below the realization and one above it; thus, a model with a better predictive density has a lower CRPS. In both the log score and the CRPS cases, the relative performance of two models is measured by the difference of their respective log scores or CRPS.

[72] Amisano and Giacomini (2007) propose a widely-used forecast density comparison test that is not robust to instabilities.

[73] Again, we consider combinations of different predictors as a special case of combinations of different models, where $m = i$ and $i = 1, 2, ..., M$.

where $w_{m,t,h} \geq 0$ and $\sum_{m=1}^{M} w_{m,t,h} = 1$.[74]

Again, the choice of the weights is crucial for obtaining successful forecasts. The weights considered in the literature can be either constant or time-varying. Constant weights include: (i) equal weights ($w_{m,t,h} = 1/M$), in which case the resulting forecast density is known as the "linear opinion pool" (Hall and Mitchell, 2007; Timmermann, 2006); (ii) trimming ($\omega_{m,t,h} = 0$ for the worst-performing models and equal weights for the remaining models). Estimated, time-varying weights include: (iii) recursive log-score weights (Jore, Mitchell and Vahey, 2010), where the weights are estimated in recursive windows: $w_{m,t,h} = \frac{\exp\left(\sum_{j=1}^{t-h} \ln \phi_{j+h|j}^{(m)}(y_{t+h})\right)}{\sum_{m=1}^{M} \exp\left(\sum_{j=1}^{t-h} \ln \phi_{j+h|j}^{(m)}(y_{t+h})\right)}$; (iv) posterior probabilities (Bayesian model averaging); (v) KLIC-based weights (Mitchell and Hall, 2005);[75] (vi) copula-based weights (Smith and Vahey, 2016).[76] An appealing property of the linear opinion pool is that, if all models have the same forecast density, then that is also the combined forecast density. Another property is that, if forecast densities are different from each other, then the combined density can be quite different from each one of them.[77] The latter can either be viewed as an advantage or a disadvantage; it may be an advantage since it produces a flexible combined density; it could be a disadvantage since the combined density does not mimic any model. In the latter case, Garratt, Henckel and Vahey (2019) propose an empirically-transformed linear opinion pool that better preserves the characteristics of the individual predictive densities. The appealing property of log-score weights is that they give a higher weight to densities that assign a high probability to the ex-post realizations. As discussed in Jore, Mitchell and Vahey (2010), this is equivalent to Bayesian model averaging using equal prior weights across models. A possible concern with log score weights is that they might be sensitive to tail events (Gneiting and Raftery, 2007).[78] KLIC-based weights instead minimize the distance between $\phi_{t+h|t}$ in eq. (20) and the true predictive density. When, due to model mis-specification, none of the forecasting models is the true model, KLIC-based forecast combinations provide forecasts from the model closest to the true one.[79] It is important to note that different weighting procedures may result in density combinations with very different properties, especially in the presence of time variation: a topic that will be important to explore in the future. Al-

[74] Imposing non-negative weights that sum to unity ensures that the combined predictive density is a density.

[75] In the KLIC-based approach: $w_{m,t,h} = \frac{\exp(-(KLIC_m - \min_m KLIC_m))}{\sum_{m=1}^{M} \exp(-(KLIC_m - \min_m KLIC_m))}$, where $KLIC_m$ is the Kullback-Leibler measure for model m.

[76] Alternatively, Ganics (2018) proposed PIT-based weights and Bassetti, Casarin and Ravazzolo (2018) propose forecast combinations of predictive densities that also takes into account calibration.

[77] For example, since the weighted average of Gaussian distributions is Gaussian, then the linear opinion pool of independent Gaussian distributions (with different means and variances) with fixed weights is a Gaussian distribution. However, when the weights are stochastic and have to be estimated, the forecast combination of independent Gaussian distributions is a mixture of normals, and hence can have skewness and kurthosis that the Gaussian distribution does not have.

[78] See Geweke and Amisano (2011) for theoretical results on log-score weights.

[79] In Bayesian model averaging, the latter might be problematic since one typically has to assume that one of the combined models is the true model. Aastveit, McAlinn, Nakajima and West (2019) introduce a multivariate combination that allows all models to be mis-specified as well as interdependence between both variables and models/forecasts. Such features deliver relatively large gains in terms of forecast accuracy during the Great Recession.

ternatively, one could also model the time-variation in the weights parametrically: Billio, Casarin, Ravazzolo and Van Dijk (2013) propose to use multivariate time-varying combinations of predictive densities, where the weight dynamics is driven by the densities' past performance using learning. Del Negro, Hasegawa and Schorfheide (2016) propose instead dynamic prediction pools to combine predictive densities using time-varying weights, which are treated as unobservable components and where the degree of time-variation in the parameters is driven by the data. Waggoner and Zha (2012) let the combination weights follow a Markov-switching model.

It is also possible to exploit different frequencies as additional dimensions. Carriero, Clark and Marcellino (2015) and Aastveit, Foroni and Ravazzolo (2017) consider predictive densities based on mixed frequency data. Aastveit, Ravazzolo and van Dijk (2018) introduce a combined density nowcasting approach to factor models that takes into account the time-varying uncertainty of the models and explicitly allows for model incompleteness. Their approach provides accurate and complete density nowcasts of U.S. GDP growth, especially for the two first months of the quarter, where the data uncertainty is relatively high and model miss-specification is more likely. Pettenuzzo and Ravazzolo (2016) instead perform combinations with the objective of choosing portfolios. See Aastveit, Mitchell, Ravazzolo and van Dijk (2018) for a detailed survey.[80]

Survey-based density forecasts are among the most used non-parametric predictive densities. For example, in the U.S., the Philadelphia Fed maintains the SPF; in Europe, the ECB maintains a European SPF. In the case of survey forecasts, $\phi^{(m)}_{t+h|t}$ is the predictive density of the m-th forecaster; hence, aggregation is across forecasters, whose models and information sets could potentially differ. Survey forecasts provide both aggregate predictive densities, from which one can obtain actual measures of aggregate forecast uncertainty, and individual forecasters' predictions, from which one can obtain measures of disagreement that are sometimes interpreted as uncertainty. Note, however, that, as Lahiri and Sheng (2010) showed, dispersion across forecasters is not a measure of uncertainty in their average forecast. De Bruin, Manski, Topa and Van der Klaauw (2011) show that, in their data, predictive densities elicited via surveys result to be internally consistent and provide reliable information on the individuals' actual perceived uncertainty; they also find that dispersion across forecasters may over-estimate the uncertainty associated with predictive densities, confirming that the two are substantially different concepts. One drawback of predictive densities from surveys is that they are often conducted for "fixed-events".[81] For example, in each quarter panelists are asked to forecast GDP growth and inflation for the current calendar year and the next, implying that the forecast horizon shrinks over time as they approach the end of the year. The fixed-event nature limits the usefulness of survey density predictions for policymakers and market participants, who often wish to characterize uncertainty a fixed number of periods ahead ("fixed-horizon"). Ganics, Rossi and Sekhposyan (2019) develop fixed-horizon density forecasts from combining fixed-event probabilistic predictions. It is also possible to combine density and point forecasts, as in Krüger, Clark and Ravazzolo (2017), who min-

[80] Alternative combination strategies involve using disaggregate data. For example, Proietti, Marczak and Mazzi (2017) combine density estimates from GDP sub-components to predict GDP growth.

[81] For example, the US SPF is conducted for fixed-events; the BCEI is also conducted for fixed-events, for some variables. However, the European SPF is conducted for both fixed-events and fixed-horizons.

imize the distance between the true distribution and the estimated density subject to the constraints imposed by the point forecasts.[82]

5.2.2 Modeling Instabilities Explicitly And Technical Aspects in Large-Dimensional Models

Most, if not all, of the literature on instabilities in density forecasts takes an in-sample, model fit-based approach to predictive densities. Hence, we will focus on that, notwithstanding the caveats expressed in the previous sections. Predictive densities can be easily derived from parametric models after making assumptions on the distribution of the forecast errors. For example, assume that $y_{t+h} = \beta_t' x_t + \varepsilon_{t,t+h}$. A conditional predictive density can be obtained by assuming a parametric distribution for the error term, $\varepsilon_{t,t+h}$. For example, suppose that, conditional on information at time t, the errors are Gaussian, i.e. $N\left(0, \sigma_{t,h}^2\right)$. Then, $\phi_{t+h|t} = N\left(\beta_t' x_t, \sigma_{t,h}^2\right)$ is the predictive density. Note that, in real-time forecasting, the predictive density at time t is estimated using only data up to time t and then a forecast is made for time $(t+h)$, similarly to point forecasts. For a technical introduction to density forecasts from parametric models, see Elliott and Timmermann (2016, Chp. 13).

There is widespread evidence that the MSFEs are time-varying – e.g. Stock and Watson (2003) and Rossi (2006, 2014a). Hence, it is important to allow the variance of the forecast errors to change over time. Note that the MSFEs could be time-varying in spite of the parameters of the model being constant $-$ for example, if the volatility changes over time and its evolution is not included in the model. Changes in macroeconomic volatility are particularly important when producing density forecasts: unlike point forecasts, where the mis-specification of the volatility may result in inefficient estimates, ignoring changes in volatilities may result in a mis-specified predictive density, and, hence, the wrong assessment of uncertainty around point forecasts. Clark (2011) shows that density forecasts from small-dimensional Bayesian VARs with stochastic volatility predict well relative to models with constant volatility.

One could combine large-dimensionality and time-varying parameters and take advantage of both. This is typically done in linear regression models with many predictors or large VARs, allowing for time-variation in the coefficients. However, estimation is complicated by the fact that the number of parameters increases drastically: in addition to having N predictors, one also has to include parameters describing the forecast distribution and model their variation over time, which introduces additional parameters, computation problems, potential overfitting and large uncertainty in the parameter estimates. Hence, dimensionality-reduction is key when forecasting with such models.

Again, this can be done via either forecast combinations or shrinkage in large-scale models. In the shrinkage approach, proposed methodologies to handle time-varying parameters differ depending on how the shrinkage is performed and which parameters are allowed to be time-varying, although, overall, the estimation is typically computationally intensive and typically relies on Bayesian methods. In large-dimensional VARs with time-varying volatility but constant mean and autoregressive coefficients, Carriero, Clark and Marcellino (2016)

[82] Additional dimensions include disaggregate data (Ravazzolo and Vahey, 2014).

achieve dimensionality-reduction by letting volatilities be driven by a single common factor, while Carriero, Clark and Marcellino (2019) allow for a non-factor structure letting volatilities evolve according to a random walk. The latter develop a computationally clever method that triangularizes large-dimensional covariance matrices and makes them easily tractable. Banbura and van Vlodrop (2018) instead develop methods to estimate VARs with time-variation in both the mean and the variance, while maintaining the autoregressive parameters constant; the mean is a random walk and the volatility is modeled via stochastic volatility.[83] Koop and Korobilis (2013, 2019) allow for time variation in either the conditional mean coefficients or in the volatilities. Koop and Korobilis (2013) focus on large dimensional VARs using forgetting factors as a way to discard predictors when they become unimportant. Their specification implies, roughly speaking, an exponential smoothing over the time-varying volatilities that makes estimation in large dimensions feasible.[84] Koop and Korobilis (2019) and Korobilis (2019) instead focus on single equation estimation, rather than VARs. They develop computationally efficient algorithms for estimating large, time-varying parameter linear regression models: in their largest empirical application, they can handle a dependent variable along with as many as one hundred regressors. The method performs dynamic variable selection at each point in time, searching for the best predictors and discarding the rest. Relative to Koop and Korobilis (2013), their approach can provide a full characterization of the distribution of the volatility process, instead of a point estimate. Korobilis (2019) uses graphical approaches which, in selected Monte Carlo simulations, performs well even with hundreds of predictors. Eisenstat, Chan and Strachan (2016) estimate VARs with time-varying parameters and stochastic volatility using an indicator that chooses whether a parameter is constant or time-varying using shrinkage priors with lasso, focusing, however, on small-dimensional VARs.

The computational costs of including time-varying parameters and estimating large-dimensional Bayesian VARs are daunting and especially challenging in small samples, however, where researchers face the perils of over-parameterization. Thus, the choice of the prior becomes very important. There have been many types of priors used in Bayesian forecasting, with an important distinction being between traditional subjective priors (such as the Minnesota prior – see Koop, 2013b, for a general introduction) and more modern automatic variable selection priors, especially when modeling parameter changes in the volatility – see Chan (2019b) for a recent survey that covers these developments. For example, Carriero, Clark and Marcellino (2016) and Chan (2019a) design algorithms that adapt the choice of priors to model time-varying and heteroskedastic variances, respectively, in parsimonious ways,[85] while Korobilis and Pettenuzzo (2019) explore the role of adaptive priors in large dimensional settings.

[83] Allowing the coefficients on the lagged variables to be constant keeps the problem tractable. In their empirical analysis, they also incorporate survey expectations to reduce the uncertainty on the conditional mean.

[84] As noted in Carriero, Clark and Marcellino (2019), Koop and Korobilis' (2013) approach is not fully Bayesian, and hence cannot be used to estimate the uncertainty in the volatility in a coherent way.

[85] The latter considers non-gaussianity and heteroskedastic and serially correlated errors, while the former lets the time-varying volatility be driven by a common factor.

5.2.3 MSFE-based and Quantile-based Confidence Intervals

An alternative way to report measures of forecast uncertainty is to use the historical forecast errors, like several central banks do. For example, the FOMC SEP includes fan charts with uncertainty bands computed using the MSFEs of past historical forecasts, assuming uncertainty is constant within a certain rolling window of past data. In the latter case, the time-variation is not directly modeled. Clark, McCracken and Mertens (2018) improve such estimates by explicitly modeling the time variation in the forecast error variances using a multiple-horizon stochastic volatility model. Their model includes the forecast error from the previous quarter and forecast updates for subsequent quarters to summarize the information in the set of forecast errors at all horizons.

Another way to obtain forecast confidence intervals is to directly model the quantiles of the distribution using quantile autoregressive models. Quantile autoregressions directly model the quantile of a distribution as a function of the lags of the predictors, where the lag coefficients may differ depending on the quantile, as different variables may be important in different parts of the distribution. Manzan (2015) uses quantile autoregressive models, including information from a large dataset of predictors, such as the factors extracted from the dataset or a subset of variables selected by lasso. He finds considerable improvements in the tail of the distribution, especially at long horizons, when forecasting output and employment. Adrian, Boyarchenko and Giannone (2019) instead use quantile models as a first step to obtain predictive densities, which are estimated by subsequently smoothing across quantiles. They focus, however, on a small number of models. Lerch, Thorarinsdottir, Ravazzolo and Gneiting (2017) directly model tail densities.[86]

Note that confidence intervals are summary statistics of a distribution, hence they contain less information than a predictive density. Only in special cases they are as informative as a predictive density: for example, when the predictive density is Gaussian, knowing a confidence interval implies knowledge of the mean (which corresponds to the center of the confidence interval) and the variance of the distribution (as the extremes of a, say, 95% confidence interval equal the mean plus/minus 1.96 times the standard deviation), and, hence, knowledge of the whole predictive density.

In practice, data revisions are also important and affect the uncertainty around point forecasts as well as policymaking (Orphanides, 2001; Croushore, 2011). Galvao, Mitchell and Runge (2019) find that the public does understand that output growth point forecasts are uncertain, due to data revisions, and that communicating uncertainty improves their understanding – especially when using intervals, quantiles and bell curves.[87]

6 Conclusions

This survey article aimed at answering four main questions. The first question was: "What are forecast instabilities and why should we care about them?". As discussed in Section 2,

[86] Delle Monache and Petrella (2017) consider instead score-driven models.

[87] Clements and Galvao (2017) and Galvao and Mitchell (2019) study how professional forecasters quantify data uncertainty due to data revisions. More generally, Haldane and McMahon (2018) find that monetary policy communication might affect inflation expectations by the public.

forecast instabilities are structural changes (smooth and continuous or abrupt and discrete) in forecasting ability, defined as a function of the models' forecast errors (e.g. the squared forecast error). It is important to carefully take them into account because, in the presence of such instabilities, standard tests for forecast evaluation are invalid, as are methods to measure uncertainty around those forecasts. After all, evaluating models according to their out-of-sample predictive ability – as opposed to their in-sample fit – is an important "reality check".

The second question was: "How to assess whether a model forecasts well in the presence of instabilities?". In the presence of instabilities, it is not appropriate to test models' forecasts by using methods that are not robust to instabilities. In fact, as we showed in Section 3, traditional tests may be invalid in the presence of forecast instabilities, and more powerful tests should be used – see Table 1 for a summary of robust tests. Importantly, changes in models' forecasting ability may be due to time-variation in the parameters, but the latter are neither necessary nor sufficient; thus, if a researcher worries about time-variation in the models' forecasting performance, he/she should use forecast evaluation methods robust to instabilities rather than tests for instabilities in models' parameters.

INSERT TABLE 1 HERE

The third question was: "How to improve forecasts in the presence of instabilities?". We overviewed two main approaches: a first strategy is to allow time-variation at the model estimation stage, with the explicit goal to improve forecasting ability. The second is to use "big data", i.e. to include a large dataset of predictors/models to "guard" against instabilities, where again the choice of which predictors/models to include is designed with the ultimate goal of improving forecasting performance. Table 2 summarizes the two approaches. Both are empirically successful options. As we mentioned, however, a lot of the theoretical properties of many of the shrinkage methods developed to handle "big data" are really not known in the forecasting context – let alone in the presence of instabilities. As we discussed in Section 4.2.2, this is an area in need of theoretical analysis and will most likely see important developments in the near future. In terms of empirical findings, several works (among which Clark and McCracken, 2008; Clark and McCracken, 2010; Rossi, 2014a; Kotchoni, Leroux and Stevanovich, 2019; and Coulombe, Leroux, Stevanovich and Surprenant, 2019) empirically evaluate the performance of machine learning and large-dimensional methods for forecasting; combinations are among the best performers. Although the best method often varies, depending on which target variable and horizon is being considered, some general patterns arise: forecast combinations and Bayesian shrinkage consistently perform among the best forecasting methods across a wide variety of variables.[88] Furthermore, and perhaps not surprising, out-of-sample forecast accuracy is not strongly correlated with measures of in-sample fit.

INSERT TABLE 2 HERE

Finally, the fourth question was: "How to correctly measure and assess forecast uncertainty in unstable environments?". Again, allowing for instabilities is crucial. Section 5

[88] Forni, Giovannelli, Lippi and Soccorsi (2018) find empirical evidence in favor of factor models.

overviewed several methodologies for reporting predictive densities and confidence intervals in the presence of instabilities, as well as evaluating their correct calibration and performing comparisons. See Table 3 for a summary. The construction of predictive densities is a lively area of research that has recently attracted a lot of interest. It is important especially (but not only) for policymakers who wish to convey their assessment of uncertainty around their projections and increase public's confidence in their assessment. In terms of empirical findings, several papers (e.g. Jore, Mitchell and Vahey, 2010; Clark, 2011; Aastveit, Carriero, Clark and Marcellino, 2017, among others) find that allowing for time variation in the conditional variance is crucial for obtaining accurate density forecasts in a wide set of macroeconomic variables, and especially during the latest financial crisis. Among alternative models of time-varying volatility, VARs with stochastic volatility are typically among the best models (Clark and Ravazzolo, 2015) while, among density combination methods, equal-weight combinations are among the best for macroeconomic data (Kasha and Ravazzolo, 2010; Rossi and Sekhposyan, 2014). Overall, the empirical findings suggest that either forecast combinations or a careful modeling of time-varying volatilities are important ingredients for securing successful predictive densities in unstable environments – although with some exceptions, most of the evaluation is still performed using forecast evaluation methods that are not robust to instabilities and more work needs to be done to draw broader conclusions.

INSERT TABLE 3 HERE

Overall, we have shown that instabilities in forecasting are extremely important, both empirically and theoretically; therefore, it is crucial to take them into account when forecasting as well as when evaluating forecasts and their uncertainty. Thus, the topics reviewed in this article are important for practitioners, policy institutions and researchers alike: after all, forecasting is not just predicting the future, but also offers the ultimate model validation.

References

Aastveit, K.A., A. Carriero, T.E. Clark and M. Marcellino (2017), "Have Standard VARs Remained Stable Since the Crisis?", *Journal of Applied Econometrics* 32(5), 931-951.

Aastveit, K.A., C. Foroni and F. Ravazzolo (2017), "Density Forecasts with MIDAS Models", *Journal of Applied Econometrics* 32(4), 783-801.

Aastveit, K.A., K. McAlinn, J. Nakajima and M. West (2019), "Multivariate Bayesian Predictive Synthesis in Macroeconomic Forecasting", *Journal of the American Statistical Association*, forthcoming

Aastveit, K.A., F. Ravazzolo and H.K. van Dijk (2018), "Combined Density Nowcasting in an Uncertain Economic Environment", *Journal of Business and Economic Statistics* 36(1), 131-145.

Aastveit, K.A., J. Mitchell, F. Ravazzolo and H.K. van Dijk (2018), "The Evolution of Forecast Density Combinations in Economics", *Oxford Research Encyclopedia of Economics and Finance*, forthcoming.

Adrian, T., N. Boyarchenko and D. Giannone (2019), "Vulnerable Growth", *American Economic Review* 109(4), 1263-1289.

Alessi, L., E. Ghysels, L. Onorante, R. Peach and S. Potter (2014), "Central Bank Macroeconomic Forecasting During the Financial Crisis: the European Central Bank and Federal Reserve Bank of New York Experiences", *Journal of Business and Economic Statistics* 32(4), 510-514.

Amisano, G. and R. Giacomini (2007), "Comparing Density Forecasts via Weighted Likelihood Ratio Tests", *Journal of Business and Economic Statistics* 25(2), 177-190.

Andreou, E., E. Ghysels and A. Kourtellos (2013), "Should Macroeconomic Forecasters Use Daily Financial Data and How?", *Journal of Business and Economic Statistics* 31(2), 240-251.

Andrews, D.W. (2003), "End-of-Sample Instability Tests", *Econometrica* 71(6), 1661-1694.

Ang, A., G. Bekaert and M. Wei (2007), "Do Macro Variables, Asset Markets or Surveys Forecast Inflation Better?", *Journal of Monetary Economics* 54(4), 1163–1212.

Bacon, D.W. and D.G. Watts (1971), “Estimating the Transition Between Two Intersecting Straight Lines”, *Biometrika* 58(3), 525–534.

Bai, J. and S. Ng (2002), "Determining the Number of Factors in Approximate Factor Models", *Econometrica* 70(1), 191–221.

Bai, J. and S. Ng (2008), "Forecasting Economic Time Series Using Targeted Predictors", *Journal of Econometrics* 146(2), 304-317.

Banbura, M. and M. Modugno (2014), "Maximum Likelihood Estimation of Factor Models on Data Sets with Arbitrary Pattern of Missing Data", *Journal of Applied Econometrics* 29(1), 133-160.

Banbura, M., D. Giannone, M. Modugno and L. Reichlin (2013), "Nowcasting and the Real Data Flow", in: Elliott, G. and A. Timmermann (eds.), *Handbook of Economic Forecasting* Vol. 2, North Holland: Elsevier, 195-237.

Banbura, M., D. Giannone and L. Reichlin (2010), "Large Bayesian Vector Auto Regressions", *Journal of Applied Econometrics* 25(1), 71-92.

Banbura, M. and A. van Vlodrop (2018), "Forecasting with Bayesian Vector Autoregressions with Time Variation in the Mean," *Tinbergen Institute Discussion Paper* TI 2018-025/IV.

Bassetti, F., R. Casarin and F. Ravazzolo (2018), “Bayesian Nonparametric Calibration and Combination of Predictive Distributions”, *Journal of American Statistical Association* 113(522), 675-685

Bates, J.M. and C.W.J. Granger (1969), "The Combination of Forecasts", *Operational Research Quarterly* 20, 451-468.

Bates, B.J., M. Plagborg-Møller, J.H. Stock and M.W. Watson (2013), "Consistent Factor Estimation in Dynamic Factor Models with Structural Instability", *Journal of Econometrics* 177(2), 289-304.

Berkowitz, J. (2001), "Testing Density Forecasts, With Applications to Risk Management", *Journal of Business and Economic Statistics* 19(4), 465-474.

Billio, M., R. Casarin, F. Ravazzolo and H.K. van Dijk (2013), "Time-Varying Combinations of Predictive Densities using Nonlinear Filtering", *Journal of Econometrics* 177(2), 213–232.

Breiman, L., J. Friedman, C.J. Stone and R.A. Olshen (1984), *Classification and Regression Trees.* CRC press.

Breiman, L. (1996), “Bagging Predictors,” *Machine Learning* 24(2), 123-140.

Breiman, L. (2001), "Random Forests", *Machine Learning* 45(1), 5-32.

Breitung, J. and S. Eickmeier (2011), "Testing for Structural Breaks in Dynamic Factor Models", *Journal of Econometrics* 163(1), 71-84.

Brown, R. (1959), *Statistical Forecasting for Inventory Control.* McGraw-Hill, New York.

Carrasco, M. Florens, J.P. and E. Renault (2007), "Linear Inverse Problems in Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization", in: Heckman, J.J. and E.E. Leamer (eds.), *Handbook of Econometrics*, Vol. 6B, North Holland: Elsevier.

Carrasco, M. and B. Rossi (2016), "In-Sample Inference and Forecasting in Misspecified Factor Models", *Journal of Business and Economic Statistics* 34(3), 313-338.

Carriero, A., T.E. Clark and M. Marcellino (2015), "Real-time Nowcasting with a Bayesian Mixed Frequency Model with Stochastic Volatility", *Journal of the Royal Statistical Society Series* A 178(4), 837-862.

Carriero, A., T.E. Clark and M. Marcellino (2016), "Common Drifting Volatility in Large Bayesian VARs", *Journal of Business and Economic Statistics* 34(3), 375-390.

Carriero, A., T.E. Clark and M. Marcellino (2019), “Large Bayesian Vector Autoregressions with Stochastic Volatility and Non-Conjugate Priors”, *Journal of Econometrics* 212(1), 137-154.

Chan, J.C.C. (2019a), "Large Bayesian VARs: A Flexible Kronecker Error Covariance Structure", *Journal of Business and Economic Statistics*, forthcoming.

Chan, J.C.C. (2019b), "Large Bayesian Vector Autoregressions", *CAMA Working Paper* No. 19/2019.

Chen, L., J. Dolado and J. Gonzalo (2014), "Detecting Big Structural Breaks in Large Factor Models", *Journal of Econometrics* 180(1), 30-48.

Cheng, X. and B.E. Hansen (2015) "Forecasting with Factor-Augmented Regression: A Frequentist Model Averaging Approach", *Journal of Econometrics* 186(2), 280-293.

Cheng, X., Z. Liao and F. Schorfheide (2016), "Shrinkage Estimation of High-Dimensional Factor Models with Structural Instabilities," *Review of Economic Studies* 83(4), 1511-1543.

Chevillon, G. (2016), "Multistep Forecasting in the Presence of Location Shifts", *International Journal of Forecasting* 32(1), 121-137.

Choi, H. and H. Varian (2012), "Predicting the Present with Google Trends", *Economic Record* 88(s1), 2–9.

Clark, T.E. (2011), "Real-Time Density Forecasts from Bayesian Vector Autoregressions with Stochastic Volatility", *Journal of Business and Economic Statistics* 29(3), 327-341.

Clark, T.E. and M.W. McCracken (2005), "The Power of Tests of Predictive Ability in the Presence of Structural Breaks", *Journal of Econometrics* 124(1), 1-31.

Clark T.E. and M.W. McCracken (2008), "Forecasting with Small Macroeconomic VARs in the Presence of Instability". In: Rapach D. E., Wohar M. E. (eds.), *Forecasting in the Presence of Structural Breaks and Model Uncertainty*, Emerald Group Publishing: Bingley, U.K., 93–147.

Clark, T.E. and M.W. McCracken (2009a), "Improving Forecast Accuracy by Combining Recursive and Rolling Forecasts", *International Economic Review* 50(2), 363-395.

Clark, T.E. and M.W. McCracken (2009b), "Forecasting with Small Macroeconomic VARs in the Presence of Instabilities". In: M. Wohar and D. Rapach, *Forecasting in the Presence of Structural Breaks and Model Uncertainty*, Amsterdam: Elsevier, 93-147.

Clark, T. and M. McCracken (2009c), "Tests of Equal Predictive Ability With Real-Time Data", *Journal of Business and Economic Statistics* 27(4), 441-454.

Clark, T.E. and M.W. McCracken (2010), "Averaging Forecasts from VARs with Uncertain Instabilities", *Journal of Applied Econometrics* 25(1), 5-29.

Clark, T.E., M.W. McCracken and E. Mertens (2018), "Modeling Time-Varying Uncertainty of Multiple-Horizon Forecast Errors", *Review of Economics and Statistics,* forthcoming.

Clark, T.E. and F. Ravazzolo (2015), "The Macroeconomic Forecasting Performance of Autoregressive Models with Alternative Specifications of Time-Varying Volatility", *Journal of Applied Econometrics* 30(4), 551-575.

Clark, T.E. and K.D. West (2006), "Using Out-of-Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis", *Journal of Econometrics* 135(1-2), 155-186.

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.

Clements, M. and A. Galvao (2017), "Predicting Early Data Revisions to US GDP and the Effects of Releases on Equity Markets", *Journal of Business and Economic Statistics* 35(3), 389-406.

Clements, M.P. and D.F. Hendry (1996), "Intercept Corrections and Structural Change", *Journal of Applied Econometrics* 11(5), 475–494.

Clements, M.P. and D.F. Hendry (1998), *Forecasting Economic Time Series*, Cambridge: Cambridge University Press.

Clements, M.P. and D.F. Hendry (1999a), "Some Methodological Implications of Forecast Failure", *mimeo*, Warwick University and Nuffield College.

Clements, M.P. and D.F. Hendry (1999b), *Forecasting Non-stationary Economic Time Series*. Cambridge, MA: The MIT Press.

Clements, M.P. and D.F. Hendry (2006), "Forecasting with Breaks". In: G. Elliott, C. Granger, A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Vol. 1, 605–658, North Holland: Elsevier.

Cogley, T. and A. Sbordone (2008), "Trend Inflation, Indexation and Inflation Persistence in the New Keynesian Phillips Curve", *American Economic Review* 95(5), 2101-2126.

Corradi, V. and N.R. Swanson (2006a), "Bootstrap Conditional Distribution Tests in the Presence of Dynamic Mis-specification", *Journal of Econometrics* 133(2), 779-806.

Corradi, V. and N.R. Swanson (2006b), "Predictive Density Evaluation", in: Elliott, G., C. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting* Vol. 1, North Holland: Elsevier, 197-284.

Corradi, V. and N.R. Swanson (2006c), "Predictive Density and Conditional Confidence Interval Accuracy Tests", *Journal of Econometrics* 135(1–2), 187-228.

Corradi, V. and N.R. Swanson (2014), "Testing for Structural Stability of Factor Augmented Forecasting Models", *Journal of Econometrics* 182(1), 100-118.

Coulombe, P.G., M. Leroux, D. Stevanovich and S. Surprenant (2019), "How is Machine Learning Useful for Macroeconomic Forecasting?", *mimeo*.

Croushore, D. (2006), "Forecasting with Real-Time Macroeconomic Data". In: Elliott, G., C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Vol.1, Elsevier, Chp. 17, 961-982.

Croushore, D. (2011), "Frontiers of Real-Time Data Analysis", *Journal of Economic Literature* 49(1), 72-100.

Croushore, D. and T. Stark (2001), "A Real-Time Data Set for Macroeconomists", *Journal of Econometrics* 105(1), 111–130.

Croushore, D. and T. Stark (2003), "A Real-Time Data Set for Macroeconomists: Does the Data Vintage Matter?", *Review of Economics and Statistics* 85(3), 605-617.

D'Agostino, A., L. Gambetti and D. Giannone (2013), "Macroeconomic Forecasting and Structural Change", *Journal of Applied Econometrics* 28(1), 81-101.

De Bruin, W.B., C.F. Manski, G. Topa and W. Van Der Klaauw (2011), "Measuring Consumer Uncertainty about Future Inflation", *Journal of Applied Econometrics* 26(3), 454-78.

Delle Monache, D. and I. Petrella (2017), "Adaptive Models and Heavy Tails with an Application to Inflation Forecasting", *International Journal of Forecasting* 33(2), 482-501.

Del Negro, M., R.B. Hasegawa and F. Schorfheide (2016), "Dynamic Prediction Pools: An Investigation of Financial Frictions and Forecasting Performance", *Journal of Econometrics* 192(2), 391-405.

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.

Dendramis, Y., G. Kapetanios, and M. Marcellino (2019), "A Similarity-based Approach for Macroeconomic Forecasting," *mimeo.*

Diebold, F.X. (2015), "Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests", *Journal of Business and Economic Statistics* 33(1), 1-50.

Diebold, F. X., T.A. Gunther, and A.S. Tay (1998), "Evaluating Density Forecasts with Applications to Financial Risk Management", *International Economic Review* 39(4), 863-883.

Diebold, F.X. and R. S. Mariano (1995), "Comparing Predictive Accuracy", *Journal of Business and Economic Statistics* 13(3), 253-263.

Diebold F.X., A.S. Tay and K.F. Wallis (1999), "Evaluating Density Forecasts of Inflation: the Survey of Professional Forecasters", in: Engle R.F. and H. White, (Eds.), *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger*. Oxford University Press, pp. 76-90.

Doz, C., D. Giannone and L. Reichlin (2012), "A Quasi–Maximum Likelihood Approach for Large, Approximate Dynamic Factor Models," *Review of Economics and Statistics* 94(4), 1014-1024.

Edge, R.M. and R.S. Gürkaynak (2010), "How Useful Are Estimated DSGE Model Forecasts for Central Bankers?", *Brookings Papers on Economic Activity* 41(2), 209-259.

Eisenstat, E., J.C.C. Chan and R.W. Strachan (2016), "Stochastic Model Specification Search for Time-Varying Parameter VARs," *Econometric Reviews* 35(8-10), 1638-1665.

Eklund, J., G. Kapetanios and S. Price (2010), "Forecasting in the Present of Recent Structural Change", *Bank of England Working Paper* No. 406.

Elliott, G., A. Gargano and A. Timmermann (2013), "Complete Subset Regressions", *Journal of Econometrics* 177(2), 357-373.

Elliott, G. and A. Timmermann (2005), "Optimal Forecast Combination Under Regime Switching", *International Economic Review* 46(4), 1081-1102.

Elliott, G. and A. Timmermann (2008), "Economic Forecasting", *Journal of Economic Literature* 46(1), 3-56.

Elliott, G. and A. Timmermann (2016), *Economic Forecasting*, Princeton University Press.

Farmer, L., L. Schmidt and A. Timmermann (2019), "Pockets of Predictability", *mimeo*, UCSD.

Faust, J. and J. Wright (2013), "Forecasting Inflation", in: Elliott, G., C. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting* Vol. 2, North Holland: Elsevier, 2-56.

Forni, M., A. Giovannelli, M. Lippi, S. Soccorsi (2018), "Dynamic Factor Model with Infinite-Dimensional Factor Space: Forecasting", *Journal of Applied Econometrics* 33(5), 625-642

Galvao, A.B. (2013), "Changes in Predictive Ability with Mixed Frequency Data", *International Journal of Forecasting* 29(3), 395-410.

Galvao, A. and J. Mitchell (2019), "Measuring Data Uncertainty: An Application using the Bank of England's Fan Charts for Historical GDP Growth", *ESCoE Discussion Paper* 2019-08.

Galvao, A., J. Mitchell and J. Runge (2019), "Communicating Data Uncertainty: Experimental Evidence for U.K. GDP", *mimeo.*

Ganics, G. (2018), "Optimal Density Forecast Combinations, *Bank of Spain Working Paper* No. 1751.

Ganics, G., B. Rossi and T. Sekhposyan (2019), "From Fixed-Event to Fixed-Horizon Density Forecasts: Professional Forecasters' View on Multi-horizon Uncertainty", *mimeo.*

Garratt, A., T. Henckel and S.P. Vahey (2019), "Empirically Transformed Linear Opinion Pools", *mimeo*, Warwick University.

Geweke, J. and G. Amisano (2011), "Optimal Prediction Pools", *Journal of Econometrics* 164(1), 130-141.

Giacomini, R. and B. Rossi (2006), "How Stable is the Forecasting Performance of the Yield Curve for Output Growth?", *Oxford Bulletin of Economics and Statistics* 68(s1), 783-795.

Giacomini, R. and B. Rossi (2009), "Detecting and Predicting Forecast Breakdown", *The Review of Economic Studies* 76(2), 669-705.

Giacomini, R. and B. Rossi (2010), "Forecast Comparisons in Unstable Environments", *Journal of Applied Econometrics* 25(4), 595-620.

Giacomini, R. and B. Rossi (2016), "Model Comparisons in Unstable Environments", *International Economic Review* 57 (2), p. 369-392

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

Giordani, P., R. Kohn and D. van Dijk (2007), "A Unified Approach to Nonlinearity, Outliers and Structural Breaks," *Journal of Econometrics* 137(1), 112–137.

Giordani, P. and M. Villani (2010), "Forecasting Macroeconomic Time Series With Locally Adaptive Signal Extraction," *International Journal of Forecasting* 26(2), 312–325.

Giovannelli, A. and T. Proietti (2015) "On the Selection of Common Factors for Macroeconomic Forecasting", *mimeo*, University Tor Vergata.

Giraitis, L., G. Kapetanios and S. Price (2013), "Adaptive Forecasting in the Presence of Recent and Ongoing Structural Change", *Journal of Econometrics* 177(2), 53-170.

Gneiting, T. and A.E. Raftery, (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation", *Journal of the American Statistical Association* 102(477), 359-378.

Goyal, A. and I. Welch (2003), "Predicting the Equity Premium with Dividend Ratios", *Management Science* 49(5), 639-654.

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

Granger, C.W.J. (1969), "Investigating Causal Relations by Econometric Models and Cross-spectral Methods", *Econometrica* 37(3), 424–438.

Granziera, E. and T. Sekhposyan (2019), Predicting Relative Forecasting Performance: An Empirical Investigation", *International Journal of Forecasting*, forthcoming.

Groen, J.J. and G. Kapetanios (2016), "Revisiting Useful Approaches to Data-Rich Macroeconomic Forecasting", *Computational Statistics and Data Analysis* 100, 221-239.

Groen, K.J., R. Paap and F. Ravazzolo (2013), "Real-time Inflation Forecasting in a Changing World", *Journal of Business and Economic Statistics* 31, 29-44.

Guerron-Quintana, P. and M. Zhong (2018), "Macroeconomic Forecasting in Times of Crises", *mimeo.*

Gürkaynak, R., B. Kısacıkoğlu and B. Rossi (2013), "Do DSGE Models Forecast More Accurately Out-of-Sample than VAR Models?", in: Fomby, T., L. Kilian and A. Murphy (eds.), "VAR Models in Macroeconomics - New Developments and Applications: Essays in Honor of Christopher A. Sims", *Advances in Econometrics* 32, 27-80.

Haldane, A. and M. McMahon (2018), "Central Bank Communications and the General Public", *American Economic Review P&P* 108, 578-583.

Hall, S.G. and J. Mitchell (2007), "Combining Density Forecasts", *International Journal of Forecasting* 23(1), 1-13.

Hamilton, J.D. (1989), "A New Approach to the Economic Analysis of Non-stationary Time Series and the Business Cycle", *Econometrica* 57(2), 357–384.

Han, X. and A. Inoue (2015), "Tests for Parameter Instability in Dynamic Factor Models", *Econometric Theory* 31(5), 1117-1152.

Hansen, B. (2000), "Testing for Structural Change in Conditional Models", *Journal of Econometrics* 97(1)*, 93-115.*

Hansen, P.R. (2005), "A Test for Superior Predictive Ability", *Journal of Business and Economic Statistics* 23(4), 365-380.

Hansen, P.R., A. Lunde and J.M. Nason (2011), "The Model Confidence Set", *Econometrica* 79(2), 453-497.

Harvey, A. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.

Harvey, D.I., Leybourne, S.J., R. Sollis and A.M.R. Taylor (2019), "Detecting Regimes of Predictability in the U.S. Equity Premium", *mimeo.*

Hassani, H., S. Heravi and A. Zhigljavsky (2009), "Forecasting European Industrial Production with Singular Spectrum Analysis", *International Journal of Forecasting* 25(1), 103–118.

Hassani, H., S. Heravi and A. Zhigljavsky (2013), "Forecasting UK Industrial Production with Multivariate Singular Spectrum Analysis", *Journal of Forecasting* 32(5), 395–408.

Hassani, H. and E.S. Silva (2015), "Forecasting with Big Data: A Review", *Annals of Data Science* 2(1), 5–19.

Hastie, T., R. Tibshirani and J. Friedman (2016), *The Elements of Statistical Learning.* New York: Springer.

Hendry, D.F. (2006), "Robustifying Forecasts from Equilibrium-Correction Models", *Journal of Econometrics* 135(1-2), 399-426.

Hirano, K. and J. Wright (2018), "Analyzing Cross-validation for Forecasting with Structural Instability", *mimeo.*

Hoerl, A.E. and R.W. Kennard (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems", *Technometrics* 12, 55-67.

Holt, C. (1957), "Forecasting Trends and Seasonals by Exponential Weighted Averages", *ONR Memorandum* 52/157, Carnegie Mellon University.

Hoogerheide, L., R. Kleijn, F. Ravazzolo, H. K. Van Dijk, M. Verbeek (2009), "Forecast Accuracy and Economic Gains from Bayesian Model Averaging using Time-varying Weights", in: Special Issue: Advances in Business Cycle Analysis and Forecasting", *Journal of Forecasting* 29 (1-2), p. 251-269.

Hyndman, R.J., A. Koehler, J.K. Ord and R. Snyder (2008), *Forecasting with Exponential Smoothing: The State Space Approach*, Berlin: Springer Verlag.

Hubrich, K. and D. Hendry (2011), "Combining Disaggregate Forecasts or Combining Disaggregate Information to Forecast an Aggregate", *Journal of Business and Economic Statistics* 29(2), 216-227.

Inoue, A., L. Jin and B. Rossi (2017), "Optimal Window Selection in the Presence of Possible Instabilities", *Journal of Econometrics* 196(1), 55-67.

Inoue, A. and L. Kilian (2004), "In-Sample or Out-of-Sample Tests of Predictability? Which One Should We Use?", *Econometric Reviews* 23(4), 371-402.

Inoue, A. and L. Kilian (2008), "How Useful Is Bagging in Forecasting Economic Time Series? A Case Study of U.S. Consumer Price Inflation", *Journal of the American Statistical Association* 103(482), 511-522.

Inoue, A. and B. Rossi (2005), "Recursive Predictability Tests for Real-time Data", *Journal of Business and Economic Statistics* 23(3), 336-345.

Jore, A.S., Mitchell, J. and S. P. Vahey (2010), "Combining Forecast Densities From VARs With Uncertain Instabilities", *Journal of Applied Econometrics* 25(4), 621–634.

Kascha, C. and F. Ravazzolo (2010), "Combining Inflation Density Forecasts", *Journal of Forecasting* 29(1-2), 231-250.

Kelly, B. and S. Pruitt (2015), "The Three-Pass Regression Filter: A New Approach to Forecasting Using Many Predictors", *Journal of Econometrics* 186(2), 277–476.

Kim, H.H. and N.R. Swanson (2014) "Forecasting Financial and Macroeconomic Variables using Data Reduction Methods: New Empirical Evidence", *Journal of Econometrics* 178(2), 352-367.

Kim, H.H. and N.R. Swanson (2015) "Methods for Past-casting, Nowcasting and Forecasting Using Factor-MIDAS with an Application to Real-Time Korean GDP", *mimeo*, Rutgers University.

Kim, H.H. and N.R. Swanson (2018) "Mining Big Data Using Parsimonious Factor and Shrinkage Methods", *International Journal of Forecasting* 34(2), 339-354.

Knueppel, M. (2015), "Evaluating the Calibration of Multi-Step-Ahead Density Forecasts Using Raw Moments", *Journal of Business and Economic Statistics* 33(2), 270-281.

Koop, G. (2013a), "Forecasting with Medium and Large Bayesian VARs", *Journal of Applied Econometrics* 28(2), 177-203.

Koop, G. (2013b), *Bayesian Econometrics.* Wiley.

Koop, G. and D. Korobilis (2012), "Forecasting Inflation Using Dynamic Model Averaging", *International Economic Review* 53(3), p. 867-886.

Koop, G. and D. Korobilis (2013), "Large Time-varying Parameter VARs", *Journal of Econometrics* 177(2), 185-198.

Koop, G. and D. Korobilis (2019), "Variational Bayes Inference in High-Dimensional Time-Varying Parameter Models," *mimeo.*

Koop, G., D. Korobilis and D. Pettenuzzo (2019), "Bayesian Compressed Vector Autoregressions", *Journal of Econometrics* 210(1), 135-154

Koop, G. and S.M. Potter (2007), "Estimation and Forecasting in Models with Multiple Breaks", *Review of Economic Studies* 74, 763-789.

Korobilis, D. (2019), "High-Dimensional Macroeconomic Forecasting Using Message Passing Algorithms," *Journal of Business and Economic Statistics*, forthcoming.

Korobilis, D. and D. Pettenuzzo (2019), "Adaptive Hierarchical Priors for High-dimensional Vector Autoregressions," *Journal of Econometrics* 212(1), 241-271.

Kotchoni, R., M. Leroux and D. Stevanovic (2019), "Macroeconomic Forecast Accuracy in a Data-rich Environment", *Journal of Applied Econometrics*, forthcoming.

Krüger, F., T.E. Clark and F. Ravazzolo (2017), "Using Entropic Tilting to Combine BVAR Forecasts With External Nowcasts", *Journal of Business & Economic Statistics* 35(3), 470-485.

Lahiri, K. and X. Sheng (2010), "Measuring Forecast Uncertainty by Disagreement: The Missing Link", *Journal of Applied Econometrics* 25(4), 514-538.

Lerch, S., T. Thorarinsdottir, F. Ravazzolo and T. Gneiting (2017), "Forecaster's Dilemma: Extreme Events and Forecast Evaluation", *Statistical Science* 32(1), 106-127.

Lettau, M. and S. Ludvigsson (2001), "Consumption, Aggregate Wealth, and Expected Stock Returns", *Journal of Finance* 56(3), 815-850.

Lucas, R. (1976). "Econometric Policy Evaluation: A Critique". In: Brunner, K. and A. Meltzer (eds.), *The Phillips Curve and Labor Markets.* Carnegie-Rochester Conference Series on Public Policy, Elsevier, 19–46.

Mallows, C.L. (1973) "Some Comments on C_p", *Technometrics* 15, 661-675.

Manzan, S. (2015) Forecasting the Distribution of Economic Variables in a Data-Rich Environment, *Journal of Business and Economic Statistics* 33(1), 144-164.

McLeay, M. and S. Tenreyro (2019), "Optimal Inflation and the Identification of the Phillips curve," *NBER Macro Annual*, forthcoming.

Mincer, J. and V. Zarnowitz (1969), "The Evaluation of Economic Forecasts". In: Mincer, J., *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance.* New York, National Bureau of Economic Research, 1-46.

Mitchell, J. and S. 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.

Mitchell, J. and K. Wallis (2011), "Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness", *Journal of Applied Econometrics* 26(6), 1023-1040.

Mitchell, J. and M. Weale (2019), "Forecasting with Unknown Unknowns: Censoring and Fat Tails on the Bank of England's Monetary Policy Committee," *EMF Research Papers* 27.

Ng, S. and J. Wright (2013), "Facts and Challenges from the Great Recession for Forecasting and Macroeconomic Modeling", *Journal of Economic Literature* 51(4), 1120–1154.

Orphanides, A. (2001), "Monetary Policy Rules based on Real-Time Data", *American Economic Review* 91(4), 964-985.

Paye, B.S. and A. Timmermann (2006), "Instability of Return Prediction Models", *Journal of Empirical Finance* 13(3), 274-315.

Pesaran, M.H., D. Pettenuzzo and A. Timmermann (2006), "Forecasting Time Series Subject to Multiple Structural Breaks", *Review of Economic Studies* 73(4), 1057-1084.

Pesaran, M.H., A. Pick and M. Pranovich (2013), "Optimal Forecasts in the Presence of Structural Breaks", *Journal of Econometrics* 177(2), 134-152.

Pesaran, M.H. and A. Timmermann (1995), "Predictability of Stock Returns: Robustness and Economic Significance", *Journal of Finance* 50(4), 1201-1228.

Pesaran, M.H. and A. Timmermann (2000), "A Recursive Modeling Approach to Predicting U.S. Stock Returns", *Economic Journal* 110(460), 159-191.

Pesaran, M.H. and A. Timmermann (2007), "Selection of Estimation Window in the Presence of Breaks", *Journal of Econometrics* 137(1), 134-161.

Pettenuzzo, D. and F. Ravazzolo (2016), “Optimal Portfolio Choice under Decision-Based Model Combinations”, *Journal of Applied Econometrics* 31(7), 1312-1332.

Proietti, T., M. Marczak and G. Mazzi (2017), "Euromind-D: A Density Estimate of Monthly Gross Domestic Product for the Euro Area," *Journal of Applied Econometrics* 32(3), 683-703.

Rapach, D. and M. Wohar (2006), "Structural Breaks and Predictive Regression Models of Aggregate U.S. Stock Returns", *Journal of Financial Econometrics* 4(2), 238-274.

Ravazzolo, F. and S. P. Vahey (2014), "Forecast Densities for Economic Aggregates from Disaggregate Ensembles", *Studies of Nonlinear Dynamics and Econometrics* 18(4), 367–381.

Ravazzolo, F., M. Verbeek and H.K. Van Dijk (2007), "Predictive Gains from Forecast Combinations Using Time Varying Model Weights", *mimeo*.

Rossi, B. (2005), "Optimal Tests for Nested Model Selections with Underlying Parameter Instabilities", *Econometric Theory* 21(5), 962-990.

Rossi, B. (2006), "Are Exchange Rates Really Random Walks? Some Evidence Robust to Parameter Instability", *Macroeconomic Dynamics* 10(1), 20-38.

Rossi, B. (2013), "Are Exchange Rates Predictable?", *Journal of Economic Literature* 51(4), 1063-1119.

Rossi, B. (2014a), "Advances in Forecasting Under Model Instability", in: Elliott, G. and A. Timmermann (eds.), *Handbook of Economic Forecasting Vol. 2*, North Holland: Elsevier, 1203-1324.

Rossi, B. (2014b), "Density Forecasts in Economics and Policymaking", *CREI Opuscles*.

Rossi, B. and T. Sekhposyan (2013), "Conditional Predictive Density Evaluation in the Presence of Instabilities", *Journal of Econometrics* 177(2), 199-212.

Rossi, B. and T. Sekhposyan (2014) "Evaluating Predictive Densities of U.S. Output Growth and Inflation in a Large Macroeconomic Data Set", *International Journal of Forecasting* 30(3), 662-682.

Rossi, B. and T. Sekhposyan (2016), "Forecast Rationality Tests in the Presence of Instabilities, With Applications to Federal Reserve and Survey Forecasts", *Journal of Applied Econometrics* 31(3), 507-532.

Rossi, B. and T. Sekhposyan (2019), "Alternative Tests for Correct Specification of Conditional Predictive Densities", *Journal of Econometrics* 208(2), 638-657.

Rossi, B., T. Sekhposyan and M. Soupre (2018), “Understanding the Sources of Macroeconomic Uncertainty”, *CEPR Discussion Paper* 11415.

Rossi, B. and Y. Wang (2019), "VAR-based Granger-causality Tests in the Presence of Instabilities", *Stata Journal*, forthcoming.

Schapire, R.E. and Y. Freund (2012), *Boosting: Foundations and Algorithms.* MIT Press.

Smith, M.S. and S.P. Vahey (2016), "Asymmetric Forecast Densities for U.S. Macroeconomic Variables from a Gaussian Copula Model of Cross-Sectional and Serial Dependence", *Journal of Business and Economic Statistics* 34(3), 416-434.

Stark, T. (2013), "SPF Panelists' Forecasting Methods: A Note on the Aggregate Results of a November 2009 Special Survey," Philadelphia Fed, *mimeo*.

Stock, J.H. (1986), "Unit Roots, Structural Breaks and Trends", in: Engle, R.F. and D. McFadden (eds.), *Handbook of Econometrics* Vol. 4, Chp. 46, North Holland: Elsevier, 2739-2841.

Stock, J.H. and M.W. Watson (1996), "Evidence on Structural Stability in Macroeconomic Time Series Relations", *Journal of Business and Economic Statistics* 14(1), 11-30.

Stock, J.H. and M.W. Watson (1999a), "Forecasting Inflation", *Journal of Monetary Economics* 44(2), 293-335.

Stock, J.H. and M.W. Watson (1999b), "Vector Autoregressions", *Journal of Economic Perspectives* 15(4), 101-116.

Stock, J.H. and M.W. Watson (1999c), "Business Cycle Fluctuations in U.S. Macroeconomic Time Series", in: Taylor, J. and M. Woodford, *Handbook of Macroeconomics* Vol. 1A, 3-64.

Stock, J. H. and M.W. Watson (2002), "Macroeconomic Forecasting Using Diffusion Indexes", *Journal of Business and Economic Statistics* 20(2), 147-162.

Stock, J.H. and M.W. Watson (2003), "Forecasting Output and Inflation: The Role of Asset Prices", *Journal of Economic Literature* XLI, 788-829.

Stock, J.H. and M.W. Watson (2004), "Combination Forecasts of Output Growth in a Seven Country Data Set", *Journal of Forecasting* 23, 405-430.

Stock, J.H. and M.W. Watson (2007), "Has Inflation Become Harder to Forecast?", *Journal of Money, Credit and Banking* 39(1), 3-33.

Stock, J.H. and M.W. Watson (2012), "Disentangling the Channels of the 2007-09 Recession", *Brookings Papers on Economic Activity* 43(1), 81-156.

Sullivan, R., A. Timmermann and H. White (1999), "Data-Snooping, Technical Trading Rules and the Bootstrap", *Journal of Finance* 54(5), 1647-1692.

Sullivan, R., A. Timmermann and H. White (2001), "Dangers of Data Mining: The Case of Calendar Effects in Stock Returns", *Journal of Econometrics* 105(1), 249-286.

Swanson, N.R. and H. White (1997a), "A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks", *Review of Economics and Statistics* 79(4), 540-550.

Swanson, N.R. and H. White (1997b), "Forecasting Economic Time Series Using Adaptive Versus Nonadaptive and Linear Versus Nonlinear Econometric Models", *International Journal of Forecasting* 13(4), 439-461.

Swanson, N.R. and H. White (1995), "A Model Selection Approach to Assessing the Information in the Term Structure Using Linear Models and Artificial Neural Networks", *Journal of Business and Economic Statistics* 13(3), 265-275.

Swanson, N., W. Xiong and X. Yang (2020), "Predicting Interest Rates Using Shrinkage Methods, Real-Time Diffusion Indexes, and Model Combinations", *Journal of Applied Econometrics*, forthcoming.

Teräsvirta, T. (1994), "Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models", *Journal of the American Statistical Association* 89(425), 208–218.

Teräsvirta, T. (1998), “Modeling Economic Relationships with Smooth Transition Regressions”. In: Ullah, A. and D.E. Giles (eds.), *Handbook of Applied Economic Statistics.* Dekker, New York, pp. 507–552.

Teräsvirta, T. (2006), "Forecasting Economic Variables with Nonlinear Models". In: Elliott, G., C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Volume 1, 414-457.

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso", *Journal of the Royal Statistical Society Series B* 58, 267-288.

Timmermann, A. (2006), "Forecast Combinations". In: Elliot, G., C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting* Vol. 1, North Holland: Elsevier, 135–196.

Timmermann, A. (2008), "Elusive Return Predictability", *International Journal of Forecasting* 24, 1-18.

Tong, H. (1990), *Non-Linear Time Series. A Dynamical System Approach*, Oxford University Press, Oxford.

Waggoner, D. and T. Zha (2012), "Confronting Model Mis-specification in Macroeconomics", *Journal of Econometrics* 171(2), 167-184.

Wallis, K. (1999), "Asymmetric Density Forecasts of Inflation and the Bank of England’s Fan Chart", *National Institute Economic Review* 167(1), 106-112.

Wallis, K. (2004), "Assessment of Bank of England and National Institute Inflation Forecast Uncertainties", *National Institute Economic Review* 189(1), 64-71.

West, K.D. (1996), "Asymptotic Inference about Predictive Ability", *Econometrica* 64(5), 1067-1084.

West, K.D. (2006), "Forecast Evaluation", in: Elliott, G., C. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting* Vol. 1, North Holland: Elsevier, 99-134.

West, K.D., and M.W. McCracken (1998), "Regression-Based Tests of Predictive Ability", *International Economic Review* 39(4), 817-840.

White, H. (1992), *Artificial Neural Networks: Approximation and Learning Theory.* Blackwell Publishers.

White, H. (2000), "A Reality Check for Data Snooping", *Econometrica* 68(5), 1097-1126.

Wright, J.H. (2008), "Bayesian Model Averaging and Exchange Rate Forecasts", *Journal of Econometrics* 146(2), 329-341.

Wright, J.H. (2009), "Forecasting U.S. Inflation by Bayesian Model Averaging", *Journal of Forecasting* 28(2), 131-144.

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties", *Journal of the American Statistical Association* 101, 1418-1429.

Zou, H. and T. Hastie (2005), "Regularization and Variable Selection via the Elastic Net", *Journal of the Royal Statistical Society Series B* 67, 301-320.

Zou, H. and H.H. Zhang (2009), "On the Adaptive Elastic-Net with a Diverging Number of Parameters", *Annals of Statistics* 37, 1733-1751.

Technical Appendix

This Appendix collects details on how to calculate the tests discussed in Section 3.

Algorithm 6 (Fluctuation Rationality test (Rossi and Sekhposyan, 2016)) *The Fluctuation Rationality test is:* $\sup_t \mathcal{W}_{t,m}$*, where* $\mathcal{W}_{t,m}$ *is a sequence of F-statistics for testing* $\alpha = \beta = 0$ *calculated in rolling samples of size* m *centered at time* t*. (If the first forecast is made at time R, i.e. the first forecast is* $y_{R+h|R}$*, then* $t = R + m/2, ..., T - m/2$*). The test reject forecast rationality when* $\sup_t \mathcal{W}_{t,m}$ *is larger than the critical values reported in Rossi and Sekhposyan (2016).*

Algorithm 7 (Fluctuation test (Giacomini and Rossi, 2010)) *The Fluctuation test is* $sup_t \left|F_{t,m}\right|$*,where* $F_{t,m}$ *is a t-test on* α *in the regression* $\Delta\mathcal{L}_{j,h} = \alpha + u_j$ *estimated in rolling samples of size* m *centered at time* t*, and* u_j *is the regression error.*[89] *(If the first forecast is made at time R, i.e. the first forecast is* $y_{R+h|R}$*, then* $t = R + m/2, ..., T - m/2$*). The test rejects equal predictive ability when* $sup_{t=R+m/2,...,T-m/2} \left|F_{t,m}\right|$ *is larger than the critical values reported in Giacomini and Rossi (2010).*

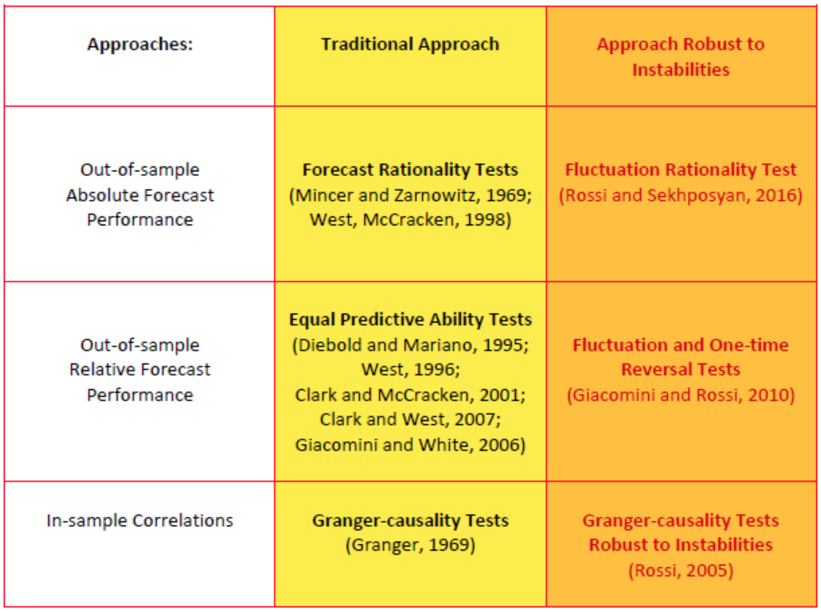
Algorithm 8 (One-time Reversal test (Giacomini and Rossi, 2010)) *The One-time Reversal test is* $sup_t W_t$*,where* W_t *is a joint F-test on* α *and* δ *in the regression* $\Delta\mathcal{L}_{j,h} = \alpha + \delta\left(d_t - t\right) + u_j$*,* u_j *is the regression error, and* d_t *is a dummy variable equal to unity if* $j \leq t$*, for* $t = \left[0.15T\right], \left[0.15T\right] + 1, ..., \left[0.85T\right]$*. (If the first forecast is made at time R, i.e. the first forecast is* $y_{R+h|R}$*, then* $t = R + \left[0.15T\right], ..., R + \left[0.85T\right]$*).*[90] *The test rejects equal predictive ability when* $sup_t W_t$ *is larger than the critical values reported in Giacomini and Rossi (2010).*

[89] HAC standard errors are recommended. Since the models are nested, in practice we perform a correction to the out-of-sample squared error differences ($\Delta\mathcal{L}_{j,h}$) due to Clark and West (2006, 2007). See Giacomini and Rossi (2010) for the derivation of the test and its critical values.

[90] Note that $[0.15T]$ denotes the largest integer of 15% of the total sample size. HAC standard errors are recommended. The test is equivalent to the one presented in Giacomini and Rossi (2010).

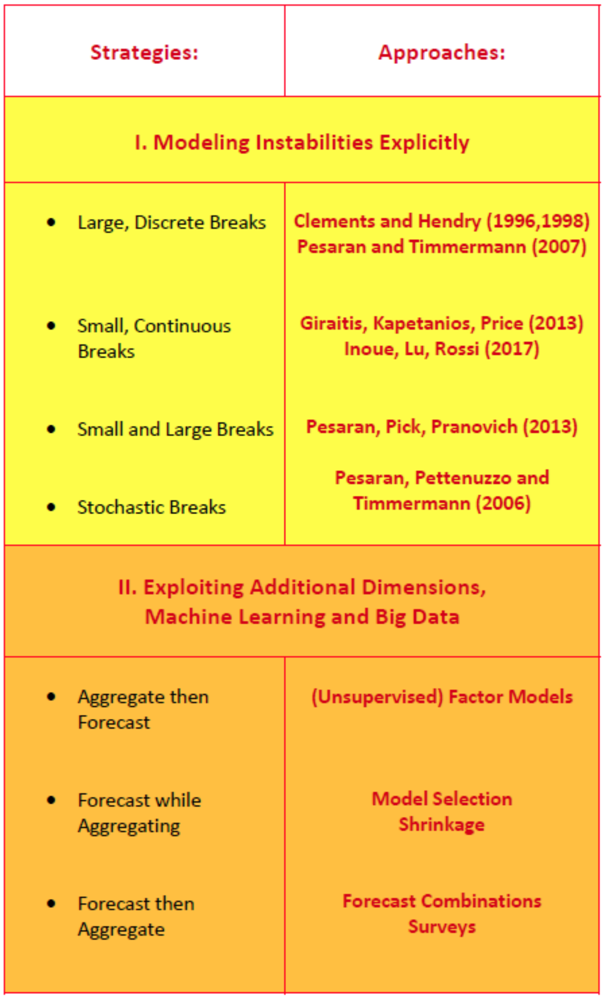
Tables and Figures

Table 1. Forecast Evaluation Tests



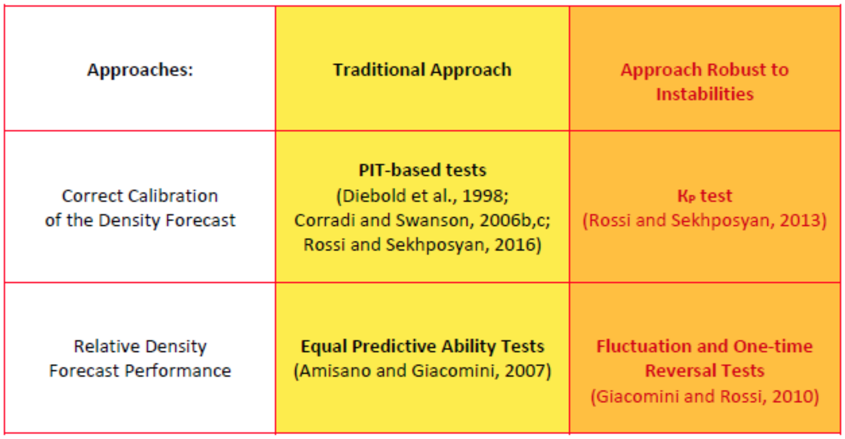
Approaches:	Traditional Approach	Approach Robust to Instabilities
Out-of-sample Absolute Forecast Performance	**Forecast Rationality Tests** (Mincer and Zarnowitz, 1969; West, McCracken, 1998)	**Fluctuation Rationality Test** (Rossi and Sekhposyan, 2016)
Out-of-sample Relative Forecast Performance	**Equal Predictive Ability Tests** (Diebold and Mariano, 1995; West, 1996; Clark and McCracken, 2001; Clark and West, 2007; Giacomini and White, 2006)	**Fluctuation and One-time Reversal Tests** (Giacomini and Rossi, 2010)
In-sample Correlations	**Granger-causality Tests** (Granger, 1969)	**Granger-causality Tests Robust to Instabilities** (Rossi, 2005)

Table 2. Strategies for Forecasting in the Presence of Instabilities



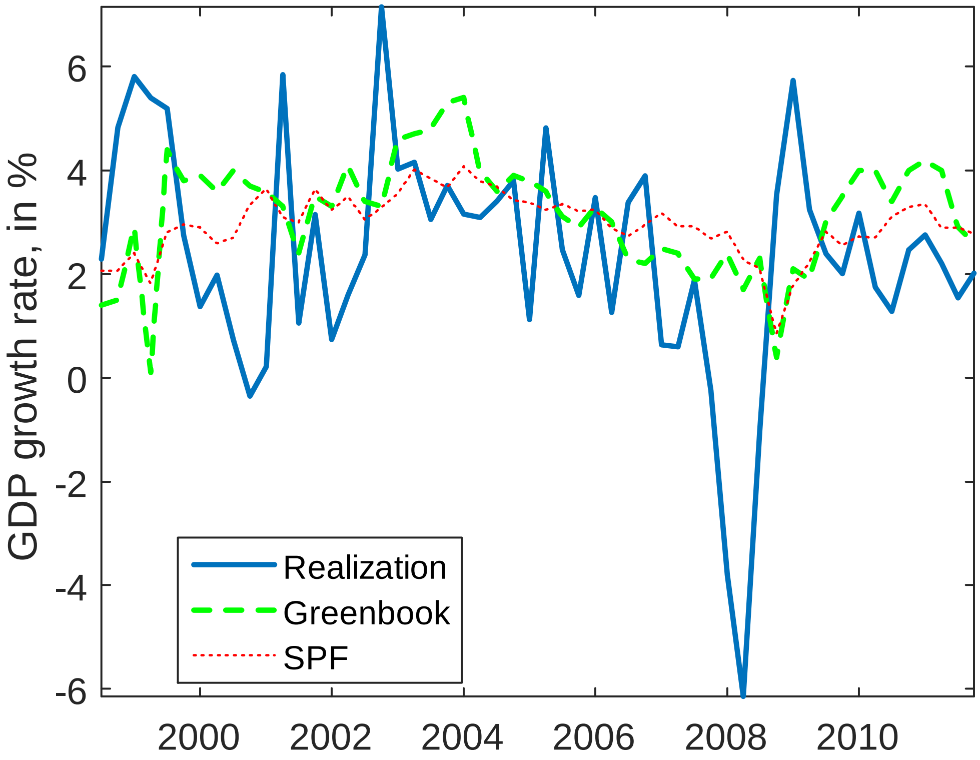
Strategies:	Approaches:
I. Modeling Instabilities Explicitly	
- Large, Discrete Breaks	Clements and Hendry (1996,1998) Pesaran and Timmermann (2007)
- Small, Continuous Breaks	Giraitis, Kapetanios, Price (2013) Inoue, Lu, Rossi (2017)
- Small and Large Breaks	Pesaran, Pick, Pranovich (2013)
- Stochastic Breaks	Pesaran, Pettenuzzo and Timmermann (2006)
II. Exploiting Additional Dimensions, Machine Learning and Big Data	
- Aggregate then Forecast	(Unsupervised) Factor Models
- Forecast while Aggregating	Model Selection Shrinkage
- Forecast then Aggregate	Forecast Combinations Surveys

Table 3. Forecast Density Evaluation Tests



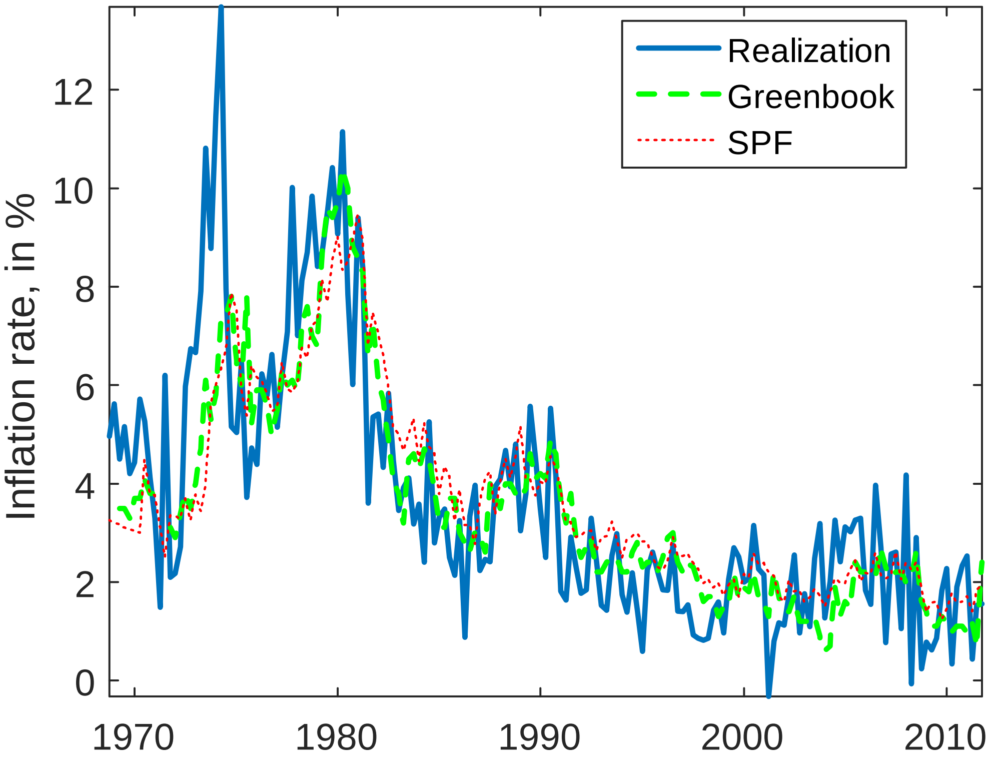
Approaches:	Traditional Approach	Approach Robust to Instabilities
Correct Calibration of the Density Forecast	**PIT-based tests** (Diebold et al., 1998; Corradi and Swanson, 2006b,c; Rossi and Sekhposyan, 2016)	K_P test (Rossi and Sekhposyan, 2013)
Relative Density Forecast Performance	**Equal Predictive Ability Tests** (Amisano and Giacomini, 2007)	Fluctuation and One-time Reversal Tests (Giacomini and Rossi, 2010)

Figure 1. Forecasting the Great Recession



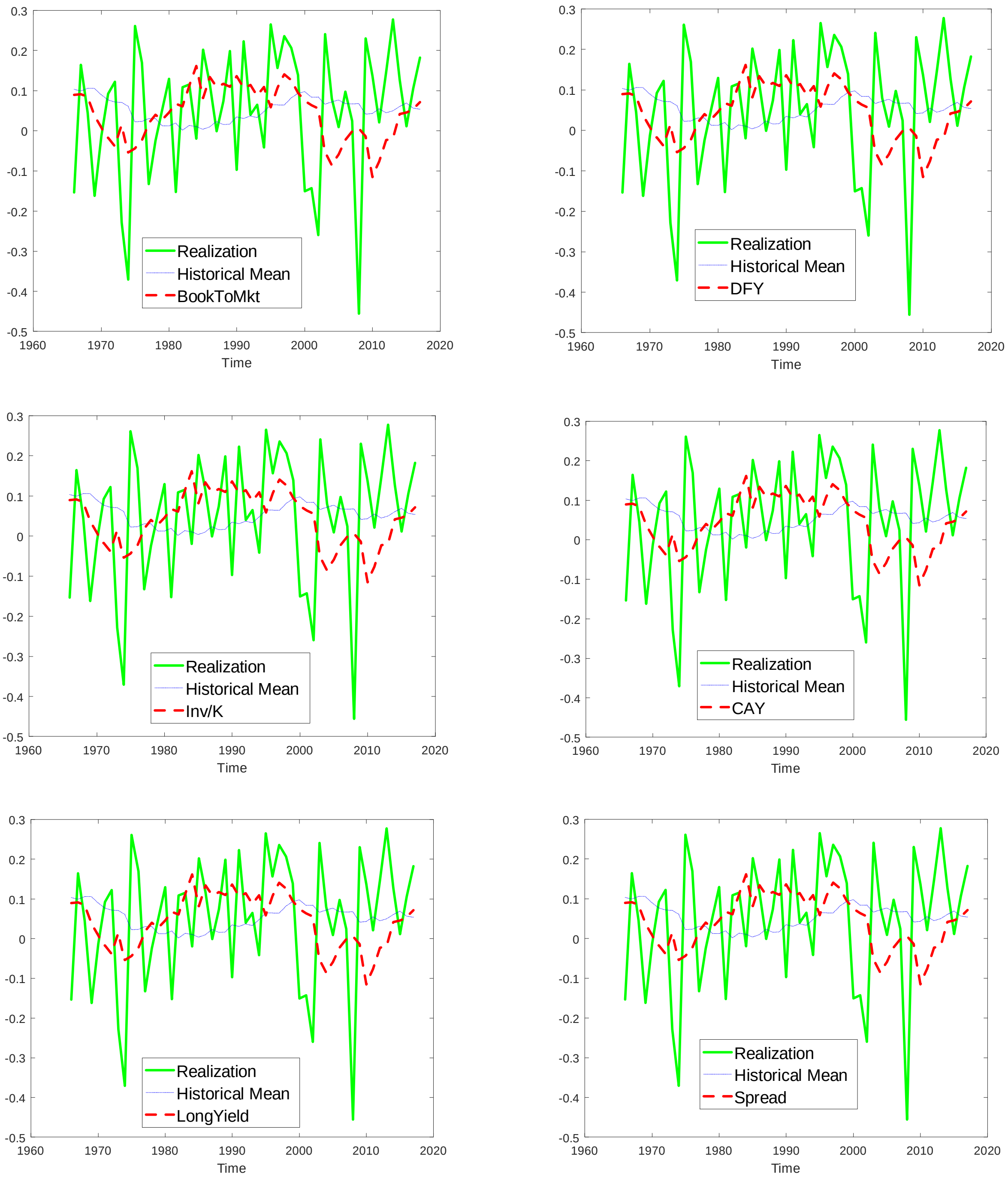
Note. The figure plots three-quarter-ahead forecasts of U.S. GDP growth made by the Federal Reserve (the dashed line, labeled "Greenbook") and the Survey of Professional Forecasters (the dotted line, labeled "SPF") together with the actual realization of GDP growth (the continuous line, labeled "actual").

Figure 2. Forecasting Inflation



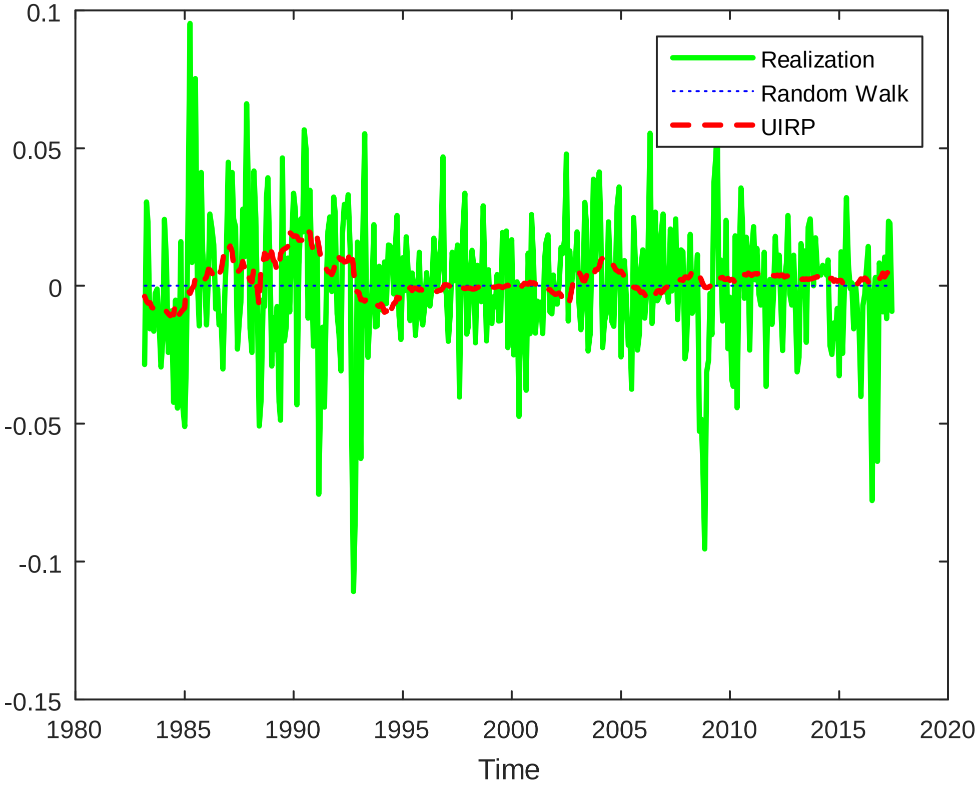
Note. The figure plots two-quarter-ahead forecasts of U.S. inflation made by the Federal Reserve (the dashed line, labeled "Greenbook") and the Survey of Professional Forecasters (the dotted line, labeled "SPF") together with the actual realization of inflation (the continuous line, labeled "actual").

Figure 3. Forecasting the Equity Premium



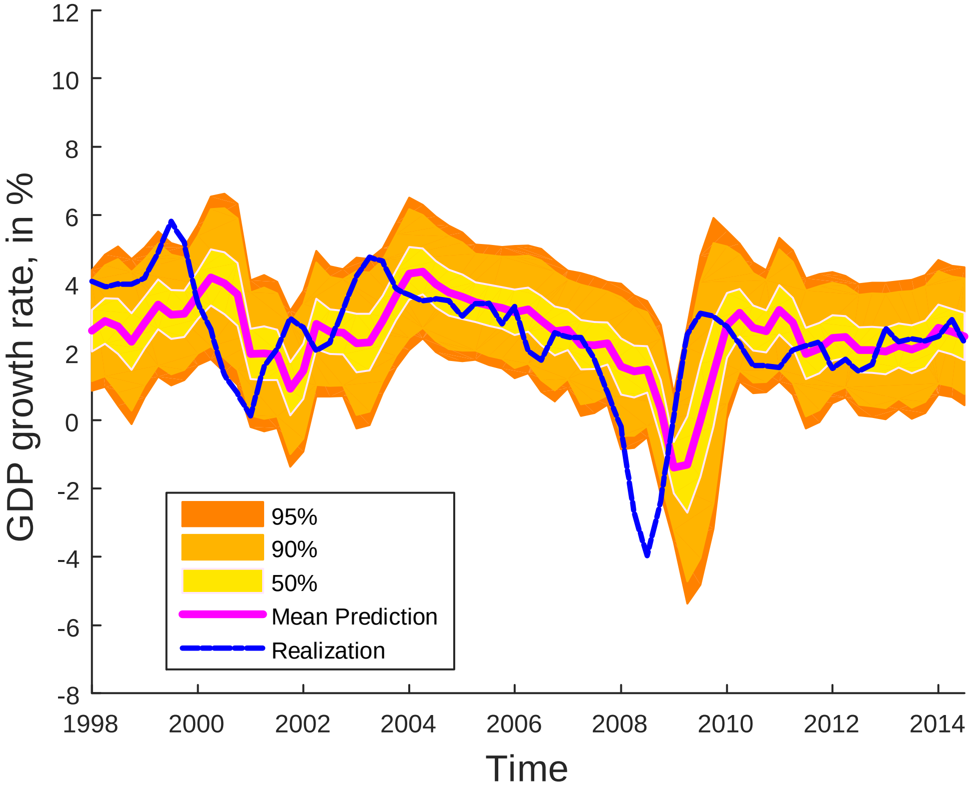
Notes. Realized values of U.S. equity premia (continuous line, labeled "Realization") together with forecasts based on the historical mean (dotted line) as well as models with predictors, depicted by the dashed line, including: book to market ratio ("BookToMkt"); the default yield spread ("DFY"); the investment capital ratio ("Inv/K"); the consumption, wealth and income ratio ("CAY"); the long term yield ("LongYield"); and the term spread ("Spread").

Figure 4. Exchange Rate Forecasts



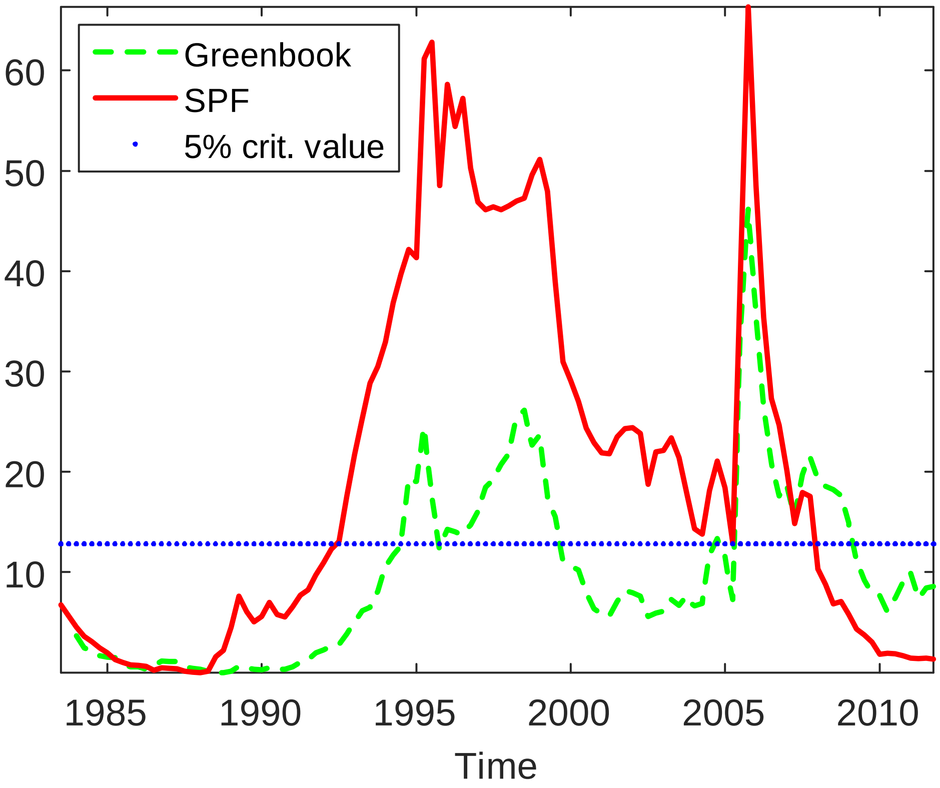
Note. The figure plots realizations of the rate of growth of the U.K. pound/U.S. dollar exchange rate (solid line, labeled "Realization") as well as the one-month-ahead forecasts of the random walk (dotted line, labeled "Random Walk") and Uncovered Interest Rate Parity (dashed line, labeled "UIRP").

Figure 5. Quantiles of Survey-based Predictive Densities of GDP Growth



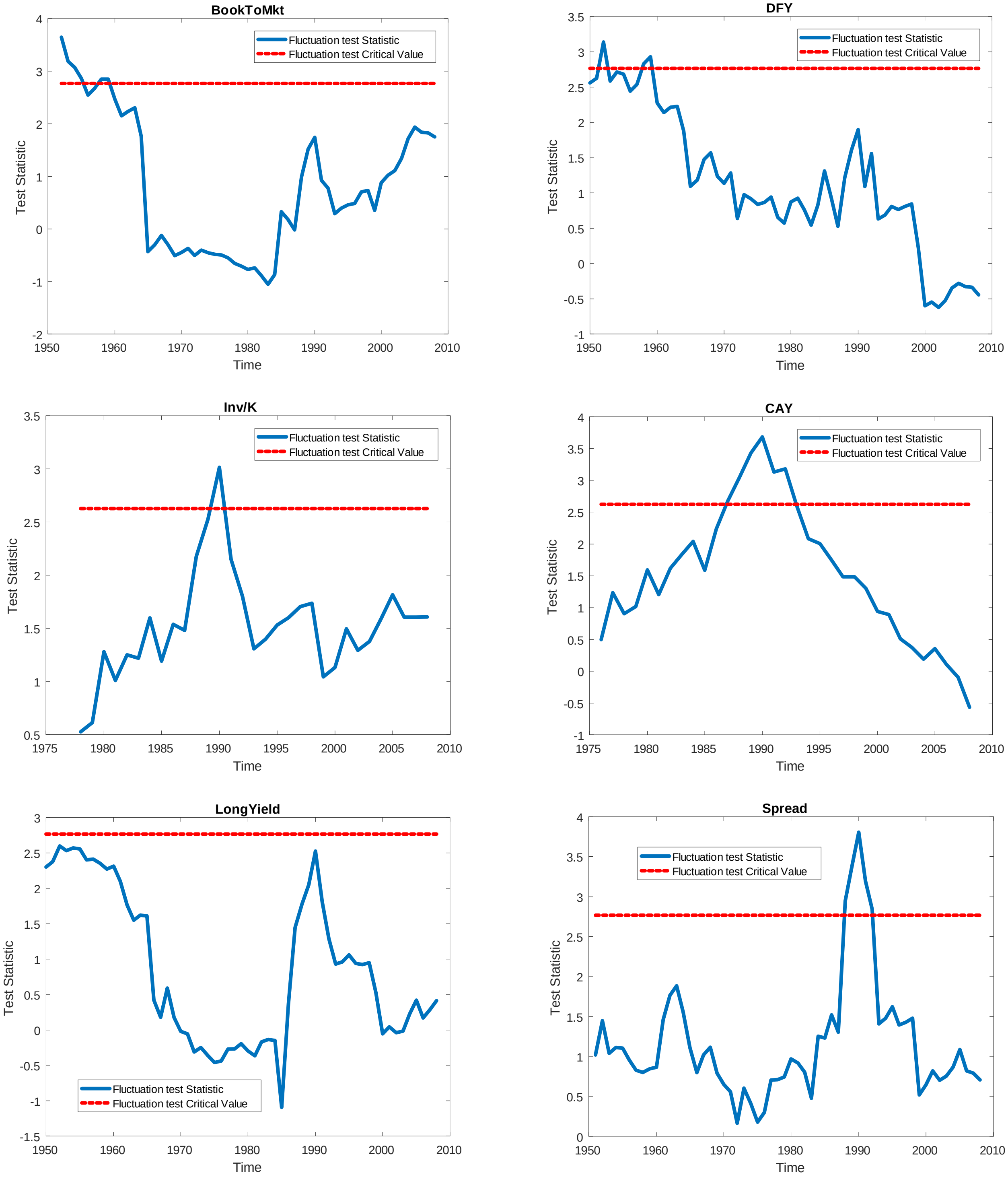
Note. The figure plots quantiles of the four-quarter-ahead, fixed-horizon Survey of Professional Forecasters density forecasts, together with the mean prediction (the central line) and the realized value (the darkest continuous line).

Figure 6. Forecast Rationality Tests
Robust to the Presence of Instabilities



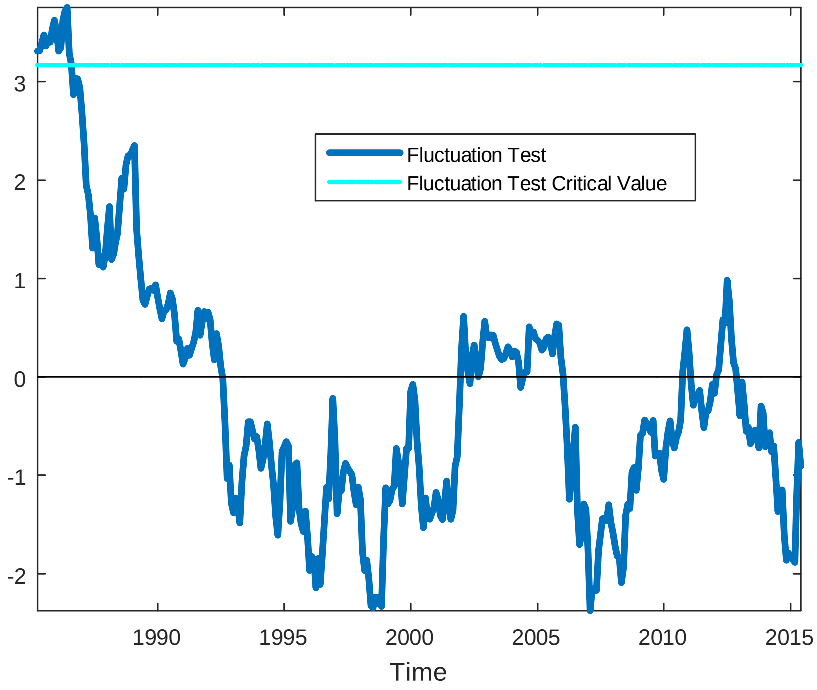
Note. The figure shows the results of the Fluctuation Rationality test for two-quarter-ahead inflation forecasts from the Federal Reserve (labeled "Greenbook") and the Survey of Professional Forecasters (labeled "SPF"). The figure depicts $W_{t,m}$; the Fluctuation Rationality test statistic is $sup_t W_{t,m}$; when the latter is above the critical value line (labeled "5% crit. value"), the forecasts are not rational.

Figure 7. Forecasting Equity Premia:
Economic Predictors vs. the Historical Average



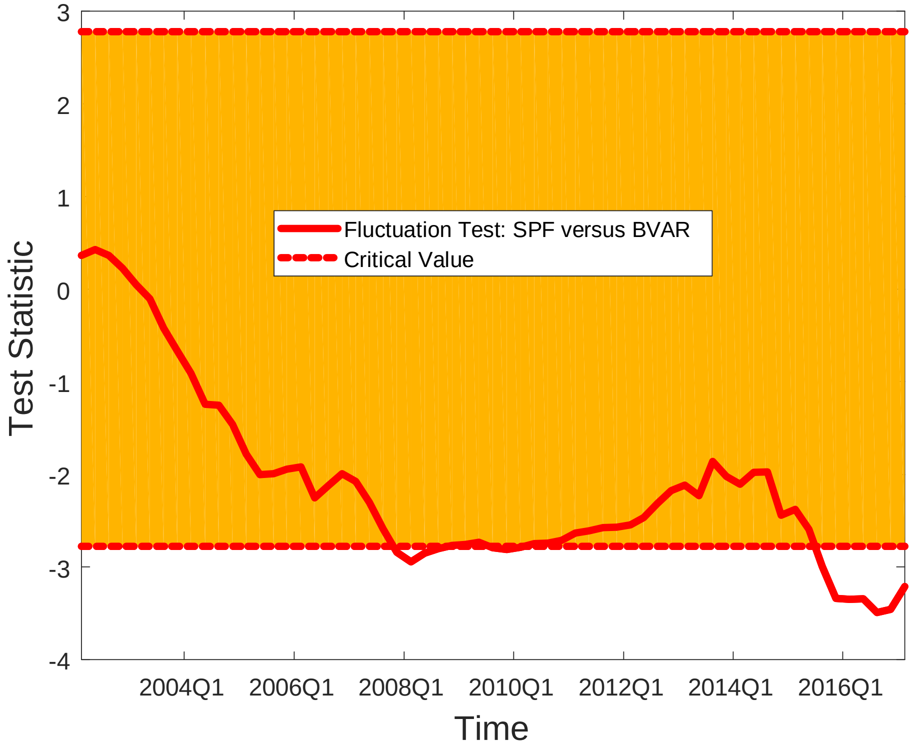
Notes. The figure depicts $F_{t,m}$ (labeled "Fluctuation test Statistic") for the following equity premium predictors relative to the historical mean: book to market ratio ("BookToMkt"); the default yield spread ("DFY"); the investment capital ratio ("Inv/K"); the consumption, wealth and income ratio ("CAY"); the long term yield ("LongYield"); and the term spread ("Spread"). Positive, significant values indicate that the model with the economic predictor forecasts better than the historical mean. The Fluctuation test statistic is $sup_t|F_{t,m}|$; when the latter is above the critical value line (labeled "Fluctuation test Critical Value"), the forecast obtained by using the economic predictor is better than the historical mean.

Figure 8. Forecast Comparison Tests
Robust to the Presence of Instabilities



Note. The figure depicts $F_{t,m}$ (labeled "Fluctuation Test") for comparing the Uncovered Interest Rate Parity model's forecasts to those of a random walk. The Fluctuation test statistic is $sup_t|F_{t,m}|$; when the latter is above the critical value line (labeled "Fluctuation Test Critical Value"), the model forecasts better than the random walk.

Figure 9. Forecast Density Comparisons



Note. The figure depicts $F_{t,m}$ (labeled "Fluctuation test: SPF versus BVAR") for comparing the Bayesian VAR and the Survey of Professional Forecasters predictive densities using the CRPS loss function. The Fluctuation test statistic is $sup_t|F_{t,m}|$; when the latter is above (below) the critical value line, the Bayesian VAR forecasts are better (worse) than the SPF. The forecast densities are four-quarter-ahead. The realizations are from 1998:Q3 to 2018:Q1 and the window size is $m = 30$ quarters.