Forecasting in Macroeconomics (in preparation for the Handbook of Research Methods and Applications on Empirical Macroeconomics)

Raffaella Giacomini and Barbara Rossi

UCL and Duke University

December 2011

Abstract

This chapter reviews forecasting methodologies that are useful for macroeconomists. The goal is to provide guidance to macroeconomists regarding which methods to use when facing a particular forecasting problem at hand. The chapter is divided in two parts. The first part is an overview of econometric methods available for forecast evaluation, including both traditional methods as well as new methodologies that are robust to instabilities. The second part addresses specific issues of importance in practice, including forecasting output growth and inflation as well as the use of real-time data and structural models for forecasting.

1 Introduction

This chapter offers a review of forecasting methodologies and empirical applications that are useful for macroeconomists. The chapter is divided in two parts. The first part overviews econometric methods available for forecast evaluation, including both traditional methods as well as new methodologies that are robust to instabilities. We discuss their usefulness, their assumptions as well as their implementation, to provide practical guidance to macroeconomists. The second part addresses special issues of interest to forecasters, including forecasting output growth and inflation as well as the use of real-time data and structural models for forecasting.

Part I. Econometric Methodologies for Forecasting in Macroeconomics

2 Methods for Forecast Evaluation

An important area of research over the past couple of decades has been the development of formal econometric techniques for evaluating the accuracy of forecasts. The problem can be viewed in a decision-theoretic context: if y_{t+1} is the variable of interest and f_t its forecast made at time t, the accuracy of f_t is judged by the expected loss $E\left[L(y_{t+1}, f_t)\right]$, for a choice of loss function $L(\cdot)$ that reflects the type of forecast (point-, interval- or density-) and the decision problem of the forecaster. The vast majority of empirical work has typically focused on the quadratic or absolute error loss, but there is some literature discussing different choices of loss function, e.g., Diebold and Lopez (1996), Amisano and Giacomini (2007), Giacomini and Komunjer (2006), Leitch and Tanner (1991), West, Edison and Cho (1993). See also Elliott, Komunjer and Timmermann (2005) for a method for eliciting forecasters' loss functions from survey data. Most of the methods discussed in the remainder of this chapter will be applicable to a general loss function, and we will provide some concrete examples below.

The expected loss of a forecast is in practice estimated using sample data. This can be done in a relatively straightforward manner if the data consists of a sequence of forecasts and corresponding realizations, as is the case for applications analyzing the accuracy of survey-based forecasts. The econometric methods for this case are standard, and we refer to, e.g., Diebold's (2007) textbook for further discussion. In many empirically relevant situations,

however, a forecaster is interested in assessing the accuracy of model-based forecasts using macroeconomic and financial time series data. In this case a sequence of forecasts is obtained by a so-called "pseudo-out-of-sample" forecasting exercise, which we describe formally below. Informally, this involves pretending that one could go back in time to a given date R in the sample (of total size T) and mimic what an actual forecaster would have done as time went by: estimate the model using the data up to time R, produce a forecast for time $R+1$, wait until $R+1$, observe the realization of the variable and compare it to the forecast, re-estimate the model including the data at time $R+1$, produce a forecast for $R+2$, wait until $R+2$ and compare it to the actual realization and so on. This procedure results in a sequence of $P = T - R$ forecasts $\left\{ f_t(\widehat{\theta}_t) \right\}_{t=R}^{T-1}$ and of corresponding out-of-sample losses $\left\{ L(y_{t+1}, f_t(\widehat{\theta}_t)) \right\}_{t=R}^{T-1}$ which depend on the data and on parameters estimated over a sequence of overlapping samples. The accuracy of the forecast is then estimated by the average of the out-of-sample losses

$$\widehat{E\left[L(y_{t+1}, f_t)\right]} = \frac{1}{P} \sum_{t=R}^{T-1} L(y_{t+1}, f_t(\widehat{\theta}_t)), \tag{1}$$

which, in the typical case of a quadratic loss, is the Mean Square Forecast Error (MSFE).

This estimate of the accuracy of the forecast is not in general directly interpretable, as it depends on the units of measurement of y_t. In practice therefore one typically relates the accuracy of a model to that of a benchmark model, or compares the accuracy of multiple competing models by comparing their out-of-sample average losses (1). In the remainder of this section we focus on testing the relative predictive ability of models, and separately consider the case of pairwise and multiple comparisons. Even though the technicalities are different, the basic idea of all the approaches that we discuss below is to develop statistical tests to assess whether the average out-of-sample losses of competing models are significantly different from each other in a way that takes into account their dependence on out-of-sample data, in-sample data and recursively estimated parameters.

A further econometric challenge that arises in the context of developing such tests is the fact that one needs to pay attention to whether the models compared are nested (in the sense that one model can be obtained from the other by imposing parameter restrictions) or non-nested. We will discuss this issue and the possible solutions below.

Finally, we will briefly consider the extension to conditional predictive ability testing, which goes beyond assessing the forecasting performance of models on average.

2.1 The Econometric Environment

In the following, we assume that the researcher is interested in forecasting the scalar variable y_t and that she has available a number of competing forecasting models. Out-of-sample testing involves dividing a sample of size T into an in-sample portion of size R and an out-of-sample portion of size P. The models are then first estimated using data from 1 to R and the parameters are used to produce $h-$ step ahead forecasts. We denote the first forecast from model i by $f_R^{(i)}(\widehat{\theta}_R)$. Some of the approaches that we discuss below do not impose restrictions on the type of model (e.g., linear or non-linear) or the estimation method used in-sample, whereas others are only applicable in special cases (e.g., the linear model estimated by OLS). We will make these assumptions explicit in each subsection. The forecasts at time R are then compared to the realization y_{R+h} and the corresponding loss for model i is denoted by $L^{(i)}(y_{R+h}, f_R^{(i)}(\widehat{\theta}_R))$. The second sets of $h-$ step ahead forecasts are obtained at time $R+1$ by either: keeping the parameter estimates fixed at $\widehat{\theta}_R$ (fixed scheme); re-estimating the models over data indexed $1, ..., R+1$ (recursive scheme) or re-estimating the models over data indexed $2, ..., R+1$ (rolling scheme). The loss for model i is then given by $L^{(i)}(y_{R+1+h}, f_{R+1}^{(i)}(\widehat{\theta}_{R+1}))$, where the definition of $\widehat{\theta}_{R+1}$ depends on the estimation scheme used. Iterating this procedure until all sample observations are exhausted yields a sequence of $P = T - h - R + 1$ out-of-sample losses $\left\{L^{(i)}(y_{t+h}, f_t^{(i)}(\widehat{\theta}_t))\right\}_{t=R}^{T-h}$ for each model i.

It is important to note that most of the techniques described below can be applied regardless of whether the forecasts are point-, volatility-, interval-, probability- or density-forecasts. The only difference lies in selecting the appropriate loss function for each type of forecast. Examples of loss functions for point forecasts are: (i) (quadratic) $L(y_{t+h}, f_t) = (y_{t+h} - f_t)^2$; (ii) (absolute error) $L(y_{t+h}, f_t) = |y_{t+h} - f_t|$; (iii) (lin-lin) $L(y_{t+h}, f_t) = (\alpha - 1(y_{t+h} - f_t < 0))(y_{t+h} - f_t)$ for $\alpha \in (0,1)$; (iv) (linex) $L(y_{t+h}, f_t) = \exp(a(y_{t+h} - f_t)) - a(y_{t+h} - f_t) - 1$ for $a \in \mathbb{R}$; (v) (direction-of-change) $L(y_{t+h}, f_t) = 1\{sign(y_{t+h} - y_t) \neq sign(f_t - y_t)\}$. Loss functions for conditional variance forecasts are (i) $L(y_{t+h}, f_t) = \left(\log(y_{t+h}^2) - \log(f_t)\right)^2$; (ii) $L(y_{t+h}, f_t) = \left(\frac{y_{t+h}^2}{f_t} - 1\right)^2$; (iii) $L(y_{t+h}, f_t) = \log(f_t) + \frac{y_{t+h}^2}{f_t}$. For probability forecasts, we have $L(y_{t+h}, f_t) = (f_t - I_t)^2$, where $I_t = 1$ if the event occurred and is 0 otherwise. For density forecasts one can consider $L(y_{t+h}, f_t) = \log f_t(y_{t+h})$.

2.2 Pairwise (Unconditional) Predictive Ability Testing

When there are only two models, one can compare their accuracy by computing the difference in, say, MSFEs, ask whether the difference is significantly different from zero and, if so, choose

the model with the smallest MSFE. For a general loss function, a test of equal predictive ability can be implemented by first constructing the time series of P out-of-sample loss differences $\left\{\Delta L_{t+h}\left(\widehat{\theta}_t\right)\right\}_{t=R}^{T-h}$ where[1] $\Delta L_{t+h} = L^{(1)}(y_{t+h}, f_t^{(1)}(\widehat{\theta}_t)) - L^{(2)}(y_{t+h}, f_t^{(2)}(\widehat{\theta}_t))$ and then conducting a t-test of the hypothesis $H_0 : \mu = 0$ in the regression

$$\Delta L_{t+h}\left(\widehat{\theta}_t\right) = \mu + \varepsilon_{t+h},\ t = R, ..., T-h. \tag{2}$$

The test has a standard normal asymptotic distribution, provided one uses the correct standard errors which take into account the time-series properties of ε_{t+h} and the dependence of ΔL_{t+h} in (2) on estimated in-sample parameters. The former challenge is relatively easy to deal with and has long been addressed in the literature, starting from Diebold and Mariano (1995), who suggested considering the test statistic

$$\left|\frac{\sqrt{P}\widehat{\mu}}{\widehat{\sigma}}\right| = \left|\frac{1}{\sqrt{P}}\sum_{t=R}^{T-h}\frac{\Delta L_{t+h}\left(\widehat{\theta}_t\right)}{\widehat{\sigma}}\right|, \tag{3}$$

where $\widehat{\sigma}$ is a heteroskedasticity- and autocorrelation-consistent standard error, e.g.,

$$\widehat{\sigma}^2 = \sum_{j=-q+1}^{q-1}(1-|j/q|)P^{-1}\sum_{t=R}^{T-h}\Delta L_{t+h}\Delta L_{t+h-j}, \tag{4}$$

with truncation lag $q = h-1$. The challenge of accounting for estimation uncertainty is trickier and has been the subject of a sizable body of literature. Broadly speaking, there are two strands of the literature, which correspond to two different asymptotic approximations in the derivation of a test of equal predictive ability. The two approaches are exemplified by West (1996) and Giacomini and White (2006).

2.2.1 West (1996)

The key insight of West (1996) is to acknowledge the dependence of (2) on $\widehat{\theta}_t$ and propose a test of equal predictive ability that is valid as both the in-sample size R and the out-of-sample size P grow to infinity. West (1996) considers a t-test of $H_0 : \mu = 0$ in a modification of the regression in (2) where the dependent variable is a function of the population parameter θ^* (interpretable as the probability limit of $\widehat{\theta}_t$ as the size of the estimation sample grows to infinity):

$$\Delta L_{t+h}\left(\theta^*\right) = \mu + \varepsilon_{t+h},\ t = R, ..., T-h. \tag{5}$$

[1]For ease of notation we stack the parameters of the two models in $\widehat{\theta}_t$.

The practical implication of this focus on population parameters is that one needs to take into account that the test statistic depends on in-sample parameter estimates, which may have an effect on the estimator of the asymptotic variance to be used in the test. Formally, West's (1996) test statistic is

$$\frac{1}{P}\sum_{t=R}^{T-h}\frac{\Delta L_{t+h}\left(\widehat{\theta}_t\right)}{\widehat{\sigma}},$$

where $\widehat{\sigma}$ is an asymptotically valid standard error that reflects the possible contribution of in-sample parameter estimation uncertainty. The main technical contribution of West (1996) is to show how to construct $\widehat{\sigma}$ for a fairly wide class of models and estimation procedures, as well as point out special cases in which estimation uncertainty is asymptotically irrelevant and $\widehat{\sigma}$ is the same standard error (4) as in the Diebold and Mariano (1995) test statistic (for example, this occurs in the case of MSFE comparisons of models estimated by OLS).

West's (1996) test has two main "disadvantages". The first, which is merely an issue of convenience of implementation, is that $\widehat{\sigma}$ is not as easily computed as the corresponding standard error in the Diebold and Mariano (1995) test, because in general it depends on the estimators used by the two models and on the estimation scheme. The second disadvantage is of a more fundamental nature, and has been discussed in a series of papers by Clark and McCracken (2001, 2005) and McCracken (2007). The key limitation of West's (1996) result is that it is only applicable to comparisons between non-nested models and thus rules out the empirically relevant comparison between a model and a nested benchmark such as an autoregression or a random walk. The technical reason for this is that West's (1996) result requires the probability limit of $\widehat{\sigma}$ to be positive as both R and P grow to infinity, which may be violated in the case of nested models. Clark and McCracken (2001, 2005) and McCracken (2007) show that it is nonetheless possible to derive a valid test of equal predictive ability for nested models within a more restrictive class of models and estimation procedures (i.e., linear models estimated by OLS and direct multi-step forecasting). The asymptotic distribution is however non-standard, so critical values for the t-test must be simulated in each specific case.

2.2.2 Giacomini and White (2006)

Giacomini and White (2006) propose deriving predictive ability tests in a different asymptotic environment with growing out-of-sample size P and fixed in-sample size R. Importantly, this assumption rules out the use of the recursive scheme in the construction of the out-of-sample

test, but allows for both fixed and the rolling schemes. The basic idea is to propose a test of $H_0 : \mu = 0$ in the regression

$$\Delta L_{t+h}\left(\widehat{\theta}_t\right) = \mu + \varepsilon_{t+h},\ t = R, ..., T-h, \tag{6}$$

where the dependent variable $\Delta L_{t+h}\left(\widehat{\theta}_t\right)$ is now a function of estimated - rather than population - parameters. This corresponds to taking a different "philosophical" view on what the relevant object of interest of the forecasting exercise is[2]. In practice, the test statistic considered by Giacomini and White (2006) is the same as the Diebold and Mariano (1995) test statistic, and thus the key message is that Diebold and Mariano's (1995) test is valid regardless of whether the models are nested or non-nested, as long as the estimation window does not grow with the sample size. The reason for the test being valid regardless whether the models are nested or non-nested is that, in a context with non-vanishing estimation uncertainty (due to the finite estimation window), the estimator $\widehat{\theta}_t$ do not converge to its probability limit and thus the denominator $\widehat{\sigma}$ of the Diebold and Mariano (1995) test cannot converge to zero.

The asymptotic framework with non-vanishing estimation uncertainty allows Giacomini and White (1996) to weaken many of the assumptions used by West (1996), Clark and McCracken (2001, 2005) and McCracken (2007), and as a result yields a test that is applicable to a wide class of models and estimation procedure, including any linear or non-linear model estimated by classical, Bayesian, semi-parametric or nonparametric procedures. The only restriction to keep in mind is that the forecasts cannot be obtained by using the recursive scheme (see Clark and McCracken (2009) for an example of a test of the Giacomini and White (2006) null hypothesis that permits recursive estimation, applicable in the special case of linear models estimated by OLS).

2.3 Pairwise (Conditional) Predictive Ability Testing

The central idea of conditional predictive ability testing (also in Giacomini and White, 2006) is to ask whether one could use available information - above and beyond past average performance - to predict which of the two forecasts will be more accurate in the future. Another way to look at this is to argue that more could be learned about the forecasting performance

[2]From a technical point of view, the reason why things work is that the assumption of a finite estimation window means that $\Delta L_{t+h}\left(\widehat{\theta}_t\right)$ can be viewed as a function of the finite history of the predictor and predictands, and as such it inherits their time series properties, which makes it easy to derive the test.

of models by studying the time series properties of the sequence of loss differences in its entirety, rather than limiting oneself to asking whether it has mean zero. For example, one could extend the regression (6) to

$$\Delta L_{t+h}\left(\widehat{\theta}_t\right) = \beta' X_t + \varepsilon_{t+h},\ t = R, ..., T-h, \tag{7}$$

where X_t contains elements from the information set at time t, such as a constant, lags of ΔL_t and economic indicators that could help predict the relative performance of the models under analysis. One could then test $H_0 : \beta = 0$ by a Wald test:

$$W = P\left(\widehat{\beta}\right)' \widehat{\Sigma}^{-1}\left(\widehat{\beta}\right), \tag{8}$$

where, because of the finite-estimation window asymptotic framework, $\widehat{\Sigma}$ is the standard HAC estimator computed by any regression software. The test is also applicable to both nested and non-nested models.

One useful feature of the extension to conditional predictive ability testing is that rejection of the null $H_0 : \beta = 0$ implies that the future relative performance of the models is predictable using current information, which suggests the following simple decision rule for choosing at time T a forecasting model for time $T+h$: choose the second model if $\widehat{\beta}' X_T > 0$ and the first model otherwise, where $\widehat{\beta}$ is estimated from (7).

2.4 Multiple Predictive Ability Testing

It is often the case that a forecaster is interested in comparing the performance of several models to that of a benchmark model, which can be viewed as a problem of multiple hypothesis testing. Referring back to the notation in Section 2.1, suppose there are N models and a benchmark denoted by 0, so that $\Delta L_{t+h}^{(i)} = L_{t+h}^{(0)} - L_{t+h}^{(i)}$ is the loss difference between the benchmark and model i. The null hypothesis of interest is that none of the N models outperforms the benchmark, and the key econometric challenge is to propose procedures that control the overall Type I error of the procedure, while taking into account the dependence of the forecast losses on each other and on the in-sample parameter estimates. White (2000) does so by proposing a "reality check" test of

$$\begin{aligned} H_0 &: \max_{i=1,...,N} E\left[\Delta L_{t+h}^{(i)}\right] \leq 0 \text{ against} \\ H_1 &: \max_{i=1,...,N} E\left[\Delta L_{t+h}^{(i)}\right] > 0, \end{aligned} \tag{9}$$

where the alternative states that there is at least one model outperforming the benchmark. White (2000) uses the asymptotic framework of West (1996) to derive the asymptotic distribution of the test statistic, which is the (out-of-) sample analogue of (9). The asymptotic distribution is the maximum of a Gaussian process and thus the p-values must be obtained by simulation. White (2000) suggests a bootstrap procedure for obtaining p-values that are valid under the assumption that at least one model is not nested in (and non-nesting) the benchmark and that estimation uncertainty is asymptotically irrelevant as in the special cases considered by West (2006) (e.g., MSFE comparison in linear models estimated by OLS).

Hansen (2005) modifies White's (2000) procedure to obtain a test that is less sensitive to the inclusion of poor-performing models and thus has higher power than White's (2000) test. Romano and Wolf (2005) suggest a further possible power improvement by adopting a "step-wise" multiple testing approach.

While the approaches described above are useful for identifying the best performing model relative to the benchmark, if there is one, they are silent about what to do in case the null hypothesis is not rejected (which could mean that the benchmark is more accurate than all the competing models or that it is as accurate as all or some of them). One may try to take a further step and ask whether it is possible to eliminate the worst-performing models and retain all the models that have equal performance, which is related to the notion of constructing a "model confidence set" (MCS), as described by Hansen, Lunde and Nason (2011). The procedure consists of the following steps:

1. Let M be the set of all possible models. Test $H_0 : E\left[L_{t+h}^{(i)} - L_{t+h}^{(j)}\right] = 0$ for all $i, j \in M$ using the statistic
$$T = \max_{i,j \in M} t_{i,j}, \tag{10}$$
where $t_{i,j}$ is the Diebold and Mariano (1995) test statistic in (3).

2. If fail to reject, all models in M are equally accurate. If reject, eliminate the worst model (that with the highest average loss) and repeat step 1 until no model is eliminated.

As in the case of the tests described above, the p-value for the test in step 1 is obtained by bootstrap methods as the test statistic (10) is not pivotal since it depends on the cross-sectional correlation of the $t'_{i,j}s$. The bootstrap p-values are computed by con-

sidering the bootstrap test statistic $T^{(b)} = max_{i,j \in M} \left| \frac{\sqrt{P}(\widehat{\mu}^{*(b)} - \widehat{\mu})}{\widehat{\sigma}^*} \right|$ for $b = 1, ..., B$, where $\widehat{\sigma}^* = \frac{P}{B} \sum_{b=1}^{B} \left(\widehat{\mu}^{*(b)} - \widehat{\mu} \right)^2$ and computing $p^* = \frac{1}{B} \sum_{b=1}^{B} 1_{\left\{ T^{(b)} > T \right\}}$.

2.5 Open Issues in Forecast Evaluation

An important issue that has been largely ignored by the literature so far, at least from a theoretical standpoint, is how to choose the sample split and/or rolling window size for the out-of-sample evaluation exercise. The question is in part linked to which asymptotic approximation one considers. There is limited evidence based on Monte Carlo simulations that shows that Giacomini and White's (2006) approximation works best when the in-sample size is small relatively to the out-of-sample size, as one would expect given the finite-estimation window assumption. Regarding West's (1996) approximation, instead, no clear guidelines seem to emerge from the simulations in the literature, except that it might not work very well when the in-sample size is small. Note that a direct comparison between the two approximations is not possible, as they test different null hypotheses. This issue has attracted a lot of attention and several new techniques have been proposed to help researchers reach empirical conclusions that are robust to the choice of the rolling window size and/or the split point, or where the latter are chosen optimally. Section 3.4 will review the recently proposed techniques that address this issue.

Another important issue is that the methodologies previously discussed are applicable only in stationary environments, which for example rules out unit roots or highly persistent variables. Analyses of the properties of forecast tests in the presence of high persistence are provided by Corradi, Swanson and Olivetti (2001) and Rossi (2005).

A more general question that has received no clear answer in the literature is if, why and when out-of-sample testing is preferable to in-sample testing, particularly when the null hypothesis is formulated in terms of (pseudo-) true parameters, as in (5). An argument against out-of-sample testing is for example made by Inoue and Kilian (2004), who show that out-of-sample tests may in fact have lower power than in-sample tests and not necessarily guard against data-mining, as is generally believed. An argument in favour is indirectly given by Clark and McCracken (2005), who show that out-of-sample tests may have an advantage over in-sample tests in that they are more "robust" to changes in predictive ability due to un-modeled structural change. Rossi and Sekhposyan (2011a) propose a new methodology to explain the difference between in-sample fit and out-of-sample forecasting

performance. They propose to decompose models' forecasting ability into asymptotically uncorrelated components that measure the contribution of instabilities, predictive content and over-fitting. We will discuss these contributions more in detail in Section 3.5.

The last result suggests that the link between predictive ability testing and structural change is worth exploring in greater depth, which is the subject of the research summarized in the next section.

3 Methods for Forecast Evaluation in the Presence of Instabilities

Stock and Watson (2003) and Rossi (2011) have discussed two main stylized facts existing in the forecasting literature on macroeconomic variables. The first stylized fact is that the predictive ability is unstable over time. For example, instabilities have been found when forecasting GDP growth using the term spread for both the U.S. (Giacomini and Rossi, 2006, and Bordo and Haubrich, 2008) as well as other major developed countries (Schrimpf and Wang, 2010, and Wheelock and Wohar, 2009). Instabilities have been found in a variety of predictors for forecasting inflation and output growth over time, as shown in Stock and Watson (2007) and Rossi and Sekhposyan (2010).

More in detail, Stock and Watson (2003) assess the lack of stability using parameter instability tests in Granger-causality type regressions. In-sample Granger-causality tests assess the significance of the proposed predictors in a regression of the dependent variable (say y_{t+h}) onto the lagged predictors (say, x_t), where h is the forecast horizon. That is, the Granger-causality test is a simple F-test on the parameter vector β_h, where:

$$y_{t+h} = \beta_h' x_t + \gamma_h' z_t + \varepsilon_{t,h},\ t = 1, ..., T \tag{11}$$

and z_t are other control variables (for example, lags of y: $y_t,\ y_{t-1}, ...$) and the total sample size available to the researcher is $T + h$. Stock and Watson (2003) evaluate the stability of β_h in regression (11) by using Andrews' (1993) test for structural breaks, and reject stability for most of the regressors. In addition, they evaluate the forecasting ability of predictors in sub-samples and find that the existence of predictability in a sub-sample does not necessarily imply existence of predictability in the other sub-samples.

A second stylized fact existing in the literature is that the existence of in-sample predictive ability does not necessarily imply out-of-sample forecasting ability. That is, predictors

that Granger-cause macroeconomic variables in the in-sample regression (11) do not necessarily perform well in an out-of-sample forecasting framework. A well-known example is the fact that, while exchange rate models fit well in-sample, their forecasting ability is poorer than that of a random walk (Meese and Rogoff, 1983). For other examples, see Swanson and White (1995) in forecasting interest rates, Swanson (1998) for forecasting monetary aggregates, Stock and Watson (2003) for forecasting output growth and inflation using a large dataset of predictors. That is, out-of-sample forecasting ability is harder to find than in-sample predictive ability, and therefore it is a tougher metric to be used in evaluating the performance of macroeconomic models.

How does one then assess predictive ability or estimate forecast models in the presence of instabilities? Does the widespread evidence of instabilities in macroeconomic forecasting models change our evaluation of whether it is possible to forecast macroeconomic variables? In what follows, we will review several methodologies that can be used to answer these questions; for a more detailed discussion of several test statistics as well as estimation strategies that have been proposed explicitly to address the presence of instabilities, see Rossi (2011).

In what follows, we will focus on a simplified situation where researchers are interested in predicting the h-steps ahead value of the dependent variable (say y_{t+h}) using lagged predictors (say, x_t), where h is the forecast horizon. That is,

$$y_{t+h} = \beta_h' x_t + \varepsilon_{t,h},\ t = 1, ..., T. \tag{12}$$

3.1 Granger-Causality Tests Robust To Instabilities

Traditional Granger-causality tests are invalid in the presence of instabilities. In fact, Rossi (2005) showed that traditional Granger-causality tests may have no power in the presence of instabilities. Rossi (2005) proposed a test that is robust to the presence of instabilities.

Rossi (2005) is interested in testing whether the variable x_t has no predictive content for y_t in the situation where the parameter β_t might be time-varying.[3] Her procedure is based on testing jointly the significance of the predictors and their stability over time. Among the various forms of instabilities that she considers, we focus on the case in which β_t may shift from β_1 to $\beta_2 \neq \beta_1$ at some unknown point in time, τ. That is, $\beta_t = \beta_1 \cdot 1(t < \tau) + \beta_2 \cdot 1(t \geq \tau)$.

[3] Rossi (2005) also considers the general case of testing possibly nonlinear restrictions in models estimated with Generalized Method of Moments (GMM). She also considers the case of tests on subsets of parameters, that is, in the case of Granger-causality regressions, tests on whether x_t Granger-causes y_t in the model $y_{t+h} = x_t'\beta_t + z_t'\gamma + \varepsilon_{t,h}$.

The test is implemented as follows. Let $\widehat{\beta}_{1\tau}$ and $\widehat{\beta}_{2\tau}$ denote the OLS estimators before and after the break:

$$\begin{aligned}\widehat{\beta}_{1\tau} &= \left(\frac{1}{\tau}\sum_{t=1}^{\tau-1} x_t x_t'\right)^{-1}\left(\frac{1}{\tau}\sum_{t=1}^{\tau-1} x_t y_{t+h}\right), \\ \widehat{\beta}_{2\tau} &= \left(\frac{1}{T-\tau}\sum_{t=\tau}^{T} x_t x_t'\right)^{-1}\left(\frac{1}{T-\tau}\sum_{t=\tau}^{T} x_t y_{t+h}\right).\end{aligned}$$

The test builds on two components: $\frac{\tau}{T}\widehat{\beta}_{1\tau}+\left(1-\frac{\tau}{T}\right)\widehat{\beta}_{2\tau}$ and $\widehat{\beta}_{1\tau}-\widehat{\beta}_{2\tau}$. The first is simply the full-sample estimate of the parameter, $\frac{\tau}{T}\widehat{\beta}_{1\tau}+\left(1-\frac{\tau}{T}\right)\widehat{\beta}_{2\tau} = \left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} x_t y_{t+h}\right)^{-1}$; a test on whether this component is zero is able to detect situations in which the parameter β_t is constant and different from zero. However, if the regressor Granger-causes the dependent variable in such a way that the parameter changes but the average of the estimates equals zero (as in the example previously discussed), then the first component would not be able to detect such situations. The second component is introduced to perform this task. It is the difference between the parameters estimated in the two sub-samples; a test on whether this component is zero is able to detect situations in which the parameter changes. Rossi (2005) proposes several test statistics, among which the following:

$$\begin{aligned}QLR_T^* &= \sup_{\tau=[0.15T],\ldots,[0.85T]} \Phi_T^* \\ Exp-W_T^* &= \frac{1}{T}\sum_{\tau=[0.15T]}^{[0.85T]} \frac{1}{0.7}\exp\left\{\left(\frac{1}{2}\right)\Phi_T^*\right\} \\ Mean-W_T^* &= \frac{1}{T}\sum_{\tau=[0.15T]}^{[0.85T]} \frac{1}{0.7}\Phi_T^*\end{aligned} \tag{13}$$

where $\Phi_T^* \equiv \left(\begin{array}{cc}\left(\widehat{\beta}_{1\tau}-\widehat{\beta}_{2\tau}\right)' & \left(\frac{\tau}{T}\widehat{\beta}_{1\tau}+\left(1-\frac{\tau}{T}\right)\widehat{\beta}_{2\tau}\right)'\end{array}\right)\widehat{V}^{-1}\left(\begin{array}{c}\left(\widehat{\beta}_{1\tau}-\widehat{\beta}_{2\tau}\right) \\ \left(\frac{\tau}{T}\widehat{\beta}_{1\tau}+\left(1-\frac{\tau}{T}\right)\widehat{\beta}_{2\tau}\right)\end{array}\right)$,

$$\widehat{V} = \left(\begin{array}{cc}\frac{\tau}{T}S_{xx}'\widehat{S}_1^{-1}S_{xx} & 0 \\ 0 & \frac{T-\tau}{T}S_{xx}'\widehat{S}_2^{-1}S_{xx}\end{array}\right),$$

$$
\begin{aligned}
S_{xx} &\equiv \frac{1}{T}\sum_{t=1}^{T} x_t x_t' \\
\widehat{S}_1 &= \left(\frac{1}{\tau}\sum_{t=1}^{\tau} x_t \widehat{\varepsilon}_{t+h}\widehat{\varepsilon}_{t+h} x_t^{'}\right) + \sum_{j=1}^{\tau-1}\left(1-\left|\frac{j}{\tau^{1/3}}\right|\right)\left(\frac{1}{\tau}\sum_{t=j+1}^{\tau} x_t \widehat{\varepsilon}_{t+h}\widehat{\varepsilon}_{t+h-j} x_{t-j}'\right), \qquad (14) \\
\widehat{S}_2 &= \left(\frac{1}{T-\tau}\sum_{t=\tau+1}^{T-\tau} x_{t-1}\widehat{\varepsilon}_{t+h}\widehat{\varepsilon}_{t+h} x_t^{'}\right) \qquad (15) \\
&\quad + \sum_{j=\tau+1}^{T-\tau}\left(1-\left|\frac{j}{(T-\tau)^{1/3}}\right|\right)\left(\frac{1}{T-\tau}\sum_{t=j+1}^{T-\tau} x_t \widehat{\varepsilon}_{t+h}\widehat{\varepsilon}_{t+h-j} x_{t-j}'\right),
\end{aligned}
$$

for $\widehat{\varepsilon}_{t+h} \equiv y_{t+h} - x_t'\widehat{\beta}$. If there is no serial correlation in the data, only the first component in (14) and (15) is relevant. Under the joint null hypothesis of no Granger-causality and no time-variation in the parameters ($\beta_t = \beta = 0$), QLR_T^*, $Mean - W_T^*$ and $Exp - W_T^*$ have asymptotic distributions whose critical values depend on the number of predictors, p, and are tabulated in Rossi's (2005) Table B1. For example, the 5% critical values of the QLR_T^*, $Mean - W_T^*$ and $Exp - W_T^*$ tests are, respectively: (9.826, 3.134, 5.364) in the presence of one regressor, and (14.225, 5.015, 8.743) in the presence of two regressors.

3.2 Forecast Comparisons Tests Robust To Instabilities

If researchers are interested in establishing which model forecasts the best in the presence of instabilities, they could use Giacomini and Rossi's (2010) Fluctuation test. To simplify notation, let $\Delta L_{t+h}\left(\widehat{\theta}_t\right)$, defined in eq. (2), be denoted by ΔL_{t+h}. To test the null hypothesis of equal performance at each point in time:

$$
H_0 : E\left(\Delta L_{t+h}\right) = 0 \text{ for all } t, \qquad (16)
$$

they propose computing the sequence of statistics

$$
F_t = \widehat{\sigma}^{-1} m^{-1/2} \sum_{j=t-m/2}^{t+m/2-1} \Delta L_j, \; t = R+h+m/2, ..., T-m/2+1, \qquad (17)
$$

where $m(< R)$ is a user-defined "bandwidth", $\widehat{\sigma}^2$ is a HAC estimator of the asymptotic variance of the forecast losses, e.g.,

$$
\widehat{\sigma}^2 = \sum_{j=-\widetilde{q}+1}^{\widetilde{q}-1} (1-|j/\widetilde{q}|) P^{-1} \sum_{t=R+h}^{T} \Delta L_{t+h}\Delta L_{t+h-j}, \qquad (18)
$$

and $\widetilde{q}$ is appropriately chosen (see e.g., Andrews, 1991 and Newey and West, 1987). They rely on an asymptotic approximation that assumes $\lim_{T\to\infty} \frac{m}{P} = \delta$. The null hypothesis is rejected

at the $100\alpha\%$ significance level against the two-sided alternative $E(\Delta L_{t+h}) \neq 0$ for some t when $\max_t |F_t| > k_\alpha^{GR}$, where k_α^{GR} is the appropriate critical values, which depend on δ. The critical values depend on δ, and are reported in their Table 1. For example, for values of δ equal to (.1, .2, .3, .4, .5, .6, .7, .8 and .9), the critical values are 3.393, 3.179, 3.012, 2.890, 2.779, 2.634, 2.560, 2.433, 2.248 respectively.[4]

The test statistic F_t in (17) is equivalent to Diebold and Mariano's (1995) and Giacomini and White's (2006) (unconditional) test statistic, computed over rolling out-of-sample windows of size m. Giacomini and Rossi (2010) show that their approach can be generalized to allow for any other commonly used test for out-of-sample predictive ability comparisons discussed in Section 2, as long as their asymptotic distribution is Normal. In particular, one could use the test statistics proposed by West (1996) or by Clark and West (2007), which are respectively applicable to non-nested and nested models.[5] The adoption of West's (1996) framework involves replacing $\widehat{\sigma}$ in (18) with an estimator of the asymptotic variance that reflects the contribution of estimation uncertainty (see Theorem 4.1 of West (1996)). For the nested case, the use of the Clark and West (2007) test statistic in practice amounts to replacing ΔL_{t+h} in (17) with Clark and West's (2007) corrected version.

Also note that West's (1996) approach allows the parameters to be estimated using a recursive scheme, in addition to a rolling or fixed scheme. In that case, let $\left\{W_t^{OOS}\right\}$ denote a sequence of West's (1996) test statistics for h-steps ahead forecasts calculated over recursive windows (with an initial window of size R) for $t = R + h + m/2, ..., T - m/2 + 1$. Giacomini and Rossi (2010) show that the null hypothesis of equal predictive ability is rejected when $\max_t \left|W_t^{OOS}\right| > k_\alpha^{rec}\sqrt{\frac{T-R}{t}}\left(1 + 2\frac{t-R}{T-R}\right)$, where (α, k_α^{rec}) are $(0.01, 1.143)$, $(0.05, 0.948)$ and $(0.10, 0.850)$.

Empirically, taking into account instabilities when assessing predictive ability is very important. For example, Rossi and Sekhposyan (2010) used the Fluctuation test to empirically investigate whether the relative performance of competing models for forecasting U.S. industrial production growth and consumer price inflation has changed over time. Their predictors include interest rates, measures of real activity (such as unemployment and GDP growth),

[4] They also derive critical values for one-sided tests.

[5] The fundamental difference between these approaches and Giacomini and White (2006) is that they test two different null hypotheses: the null hypothesis in West (1996) and Clark and West (2006, 2007) concerns forecast losses that are evaluated at the population parameters, whereas in Giacomini and White (2006) the losses depend on estimated in-sample parameters. This reflects the different focus of the two approaches on comparing forecasting models (West, 1996, and Clark and West, 2006, 2007) versus comparing forecasting methods (Giacomini and White, 2006).

stock prices, exchange rates and monetary aggregates. Their benchmark model is the autoregressive model. Using both fully revised and real-time data, they find sharp reversals in the relative forecasting performance. They also estimate the time of the reversal in the relative performance, which allows them to relate the changes in the relative predictive ability to economic events that might have happened simultaneously. In particular, when forecasting output growth, interest rates and the spread were useful predictors in the mid-1970s, but their performance worsened at the beginning of the 1980s. Similar results hold for money growth (M2), the index of supplier deliveries, and the index of leading indicators. When forecasting inflation, the empirical evidence in favor of predictive ability is weaker than that of output growth, and the predictive ability of most variables breaks down around 1984, which dates the beginning of the Great Moderation. Such predictors include employment and unemployment measures, among others, thus implying that the predictive power of the Phillips curve disappeared around the time of the Great Moderation.

3.3 Forecast Optimality Tests Robust To Instabilities

Rossi and Sekhposyan's (2011b) proposed robust tests of forecast optimality that can be used in case researchers are interested in assessing whether forecasts are rational. In fact, traditional tests for forecast rationality are subject to the same issues as the other tests previously discussed: they are potentially invalid in the presence of instabilities.

Consider the forecast optimality regression:

$$v_{t+h} = g_t' \cdot \alpha + \eta_{t,h}, \text{ for } t = R, ..., T, \tag{19}$$

where α is a $(p \times 1)$ parameter vector. The null hypothesis of interest is $H_0 : \alpha = \alpha_0$, where typically $\alpha_0 = 0$. For example, in forecast rationality tests (Mincer and Zarnowitz, 1969), $v_{t+h} = y_{t+h}$, $g_t = [1, y_{t+h|t}]$, $\alpha = [\alpha_1, \alpha_2]'$, and typically a researcher is interested in testing whether α_1 and α_2 are jointly zero. For forecast unbiasedness, $g_t = 1$, for forecast encompassing g_t is the forecast of the encompassed model, and for serial uncorrelation $g_t = v_t$.

To test forecast optimality, one typically uses the re-scaled Wald test:

$$\mathcal{W}_T = \widehat{\alpha}' \, \widehat{V}_\alpha^{-1} \widehat{\alpha}, \tag{20}$$

where $\widehat{V}_\alpha$ is a consistent estimate of the long run variance of the parameter vector obtained following West and McCracken (1998).[6]

[6] West and McCracken (1998) have shown that it is necessary to correct eq. (20) for parameter estimation

Rossi and Sekhposyan (2011b) propose the following procedure, inspired by Giacomini and Rossi (2010). Let $\widehat{\alpha}_t$ be the parameter estimate in regression (19) computed over centered rolling windows of size m (without loss of generality, we assume m to be an even number). That is, consider estimating regression (20) using data from $t - m/2$ up to $t + m/2 - 1$, for $t = m/2, ..., P - m/2 + 1$. Also, let the Wald test in the corresponding regressions be defined as:

$$\mathcal{W}_{t,m} = \widehat{\alpha}_t' \, \widehat{V}_{\theta,t}^{-1} \widehat{\alpha}_t, \text{ for } t = m/2, ..., P - m/2 + 1, \tag{21}$$

where $\widehat{V}_{\alpha,t}$ is a consistent estimator of the asymptotic variance of the parameter estimates in the rolling windows obtained following West and McCracken (1998). Rossi and Sekhposyan (2011b) refer to $\mathcal{W}_{t,m}$ as the Fluctuation optimality test. The test rejects the null hypothesis $H_0 : E(\widehat{\alpha}_t) = 0$ for all $t = m/2, ..., P - m/2 + 1$ if $\max_t \mathcal{W}_{t,m} > k_{\alpha,k}^{RS}$, where $k_{\alpha,k}^{RS}$ are the critical values at the $100\alpha\%$ significance level. The critical values are reported in their Table 1 for various values of $\mu = [m/P]$ and the number of restrictions, p.[7]

3.4 The Choice of the Window Size

In the presence of breaks, it might be useful to use a rolling window. But which size of the rolling window should be used? Similarly, recursive window forecasts require researchers to split the sample between an in-sample and an out-of-sample portion. Again, which split-point should be used? For simplicity, in this section we will focus on the choice of the window size, although we note that similar issues and solutions are applicable to the choice of split-point. The choice of the estimation window size has always been a concern for practitioners, and they raise several concerns. The first concern is that the use of different window sizes may lead to different empirical results in practice. In addition, arbitrary choices of window sizes have consequences about how the sample is split into in-sample and out-of-sample portions. Notwithstanding the choice of the window size is crucial, in the forecasting literature it is common to only report empirical results for one window size.

Pesaran and Timmermann (2007) study the problem of determining the optimal window size that guarantees the best forecasting performance, especially in the presence of breaks.

error in order to obtain test statistics that have good size properties in small samples, and proposed a general variance estimator as well as adjustment procedures that take into account estimation uncertainty.

[7] Rossi and Sekhposyan (2011b) also note that a simple, two-sided t-ratio test on the s-th parameter, $\alpha^{(s)}$, can be obtained as $\widehat{\alpha}_t^{(s)} \, \widehat{V}_{\alpha^{(s)},t}^{-1/2}$, where $\widehat{V}_{\alpha^{(s)},t}$ is element in the s-th row and s-th column of $\widehat{V}_{\alpha,t}$; then, reject the null hypothesis $H_0 : E\left(\widehat{\alpha}_t^{(s)}\right) = \alpha_0^{(s)}$ for all $t = m/2, ..., P - m/2 + 1$ at the $100\alpha\%$ significance level if $\max_t \left|\widehat{\alpha}_t^{(s)} \widehat{V}_{\alpha^{(s)},t}^{-1/2}\right| > k_\alpha^{GR}$, where k_α^{GR} are the critical values provided by Giacomini and Rossi (2010a).

They propose several methods in practice, among which several are available if the researcher possesses an estimate of the break, in which case, using either only the post-break window data to estimate the parameter or a combination of pre- and post-break data according to weights that trade-off bias against reduction in parameter estimation error, might improve forecasting performance. A difficulty in the latter methods is the fact that, in practice, it may be difficult to precisely estimate the time and magnitude of the break. Thus, rather than selecting a single window, they propose to combine forecasts based on several estimation windows. For example, they propose an average ("Ave") forecast:

$$y_{t+h|t}^{AVE,f} = (T - \underline{R} + 1)^{-1} \sum_{R=t-\underline{R}}^{t} y_{t+h|t}^{f}(R) \tag{22}$$

where R is the size of the rolling window, $\underline{R}$ is the researcher's minimum number of observations to be used for estimation, and the forecast for the target variable h-steps into the future made at time t based on data from the window size R (that is data from time $t-R+1$ to t) is denoted by $y_{t+h|t}^{f}(R)$.

An alternative approach is suggested by Inoue and Rossi (2010). While Pesaran and Timmermann's (2007), Inoue and Rossi (2010) are interested in assessing the robustness of conclusions of predictive ability tests to the choice of the estimation window size. Inoue and Rossi (2010) argue that the common practice of reporting empirical results for only one window size raises two types of concerns. First, it might be possible that satisfactory results (or lack thereof) were obtained simply by chance, and are not robust to other window sizes. Second, it might be possible that the data were used more than once for the purpose of selecting the best forecasting model and thus the empirical results were the result of data snooping over many different window sizes and the search process was not ultimately taken into account when reporting the empirical results. Inoue and Rossi (2010) propose new methodologies for comparing the out-of-sample forecasting performance of competing models that are robust to the choice of the estimation and evaluation window size by evaluating the models' relative forecasting performance for a variety of estimation window sizes, and then taking summary statistics. Their methodology can be applied to most of the tests of predictive ability that have been proposed in the literature, such as those discussed in Section 2.

Inoue and Rossi's (2010) proposed methodology is as follows. Let $\Delta L_T(R)$ denote the test of equal predictive ability implemented using forecasts based either on a rolling window of size R or recursive/split estimation starting at observation R. For example, for the case

of the Diebold and Mariano's (1995) and West's (2006) test, $\Delta L_T(R)$ is defined as in eq. (3). Similarly, let $\Delta L_T^{\varepsilon}(R)$ denote the Clark and McCracken's (2001) ENCNEW test for nested models comparison based either on rolling window estimation with window size R or recursive/split window estimation starting at observation R. Finally, let $\mathcal{W}_T(R)$ denote tests for forecast optimality analyzed by West and McCracken (1998), including tests of forecast encompassing (Clements and Hendry, 1993, Harvey, Leybourne and Newbold, 1998), tests for forecast rationality (Mincer and Zarnowitz, 1969) and tests of forecast uncorrelatedness (Granger and Newbold, 1986, and Diebold and Lopez, 1996) based on forecast errors obtained either by estimation on a rolling window of size R or recursive/split estimation starting at observation R.

They suggest the following statistics:

$$\mathcal{R}_T = \sup_{R\in\left\{\underline{R},\ldots\overline{R}\right\}} |\Delta L_T(R)|, \text{ and } \mathcal{A}_T = \frac{1}{\overline{R}-\underline{R}+1}\sum_{R=\underline{R}}^{\overline{R}} |\Delta L_T(R)|, \tag{23}$$

$$\mathcal{R}_T^{\varepsilon} = \sup_{R\in\left\{\underline{R},\ldots\overline{R}\right\}} \Delta L_T^{\varepsilon}(R) \text{ and } \mathcal{A}_T^{\varepsilon} = \frac{1}{\overline{R}-\underline{R}+1}\sum_{R=\underline{R}}^{\overline{R}} \Delta L_T^{\varepsilon}(R), \tag{24}$$

$$\mathcal{R}_T^{\mathcal{W}} = \sup_{R\in\left\{\underline{R},\ldots\overline{R}\right\}} \mathcal{W}_T(R), \text{ and } \mathcal{A}_T^{\mathcal{W}} = \frac{1}{\overline{R}-\underline{R}+1}\sum_{R=\underline{R}}^{\overline{R}} \mathcal{W}_T(R), \tag{25}$$

where $\underline{R}$ is the smallest window size considered by the researcher, $\overline{R}$ is the largest window size, and $\widehat{\Omega}_R$ is a consistent estimate of the long run variance matrix.[8] Inoue and Rossi's (2010) obtain asymptotic approximations to eqs. (23), (24) and (25) by letting the size of the window R be asymptotically a fixed fraction of the total sample size: $\zeta = \lim_{T\to\infty}(R/T) \in (0,1)$. The null hypothesis of equal predictive ability or forecast optimality at each window size for the $\mathcal{R}_T$ test is rejected when the test statistics are larger than the critical values reported in the tables in Inoue and Rossi (2010). For example, at the 5% significance level and for $\underline{R} = [0.15T]$ and $\overline{R} = [0.85T]$, the critical values for the $\mathcal{R}_T$ and $\mathcal{A}_T$ test are, respectively, 2.7231 and 1.7292. Inoue and Rossi (2010) also consider cases where the window size is fixed – we refer interested readers to their paper for more details. Hansen and Timmermann (2011) propose a similar approach; the difference is that focus on nested models' comparisons based on recursive window estimation procedure. The advantage of their method is to provide analytic power calculations for the test statistic under very general assumptions. Unlike

[8] See West (1996) for consistent variance estimates in eq. (23), Clark and McCracken (2001) for eq. (24) and West and McCracken (1998) for eq. (25).

Inoue and Rossi (2011), however, they do not consider rolling window estimation, nor the effects of time-varying predictive ability on the power of the test.

3.5 Empirical Evidence on Forecasting in the Presence of Instabilities

In an empirical analysis focusing on the large dataset of macroeconomic predictors used in Stock and Watson (2003), Rossi (2011) finds that the Granger-causality test robust to instability proposed by Rossi (2005) is capable to overturn existing stylized facts about macroeconomic predictability and identifies more empirical evidence in favor of macroeconomic predictability, due to the fact that, in several cases, predictability only appears in sub-samples of the data. She also finds similar results when evaluating the out-of-sample forecasting ability of macroeconomic predictors: using tests of predictive ability that are robust to instabilities (such as Giacomini and Rossi, 2010) is key to uncover more predictive ability than previously found. On the other hand, tests of forecast rationality that are robust to instability (such as Rossi and Sekhposyan 2011b) find instead more evidence that typical macroeconomic predictors of inflation and output growth lead to forecasts that are not optimal.

Finally, in the presence of instabilities, as discussed in Inoue and Rossi (2010), traditional tests may encounter two problems due to the fact that they are performed conditional on a given estimation window: they might either find spurious predictability (if the researcher had performed data-mining over several window sizes) or may find too little predictive ability (if the window chosen for estimation was not the optimal one given the instability in the data). Inoue and Rossi (2010) and Hansen and Timmermann (2010) propose methods to assess forecasting ability in a way that is robust to the choice of the estimation window size.

Rossi (2011) also notes that there are several estimation procedures that have been proposed to improve models' estimation in the presence of instabilities. One should note that, as shown in Elliott and Muller (2007), it is very difficult to estimate break dates in the data, which complicates estimation in the presence of instabilities; in addition, Pesaran and Timmermann (2002) have shown that, unlike what one might suspect, it is not necessarily optimal to use only observations after a break to forecast. Estimation methods that perform well in forecasting are therefore a bit more sophisticated than models in sub-samples estimated according to possible break-dates. For example, Pesaran and Timmermann (2002, 2007) propose to adapt the estimation window to the latest break in a more sophisticated

manner; Pesaran, Pettenuzzo and Timmermann (2006) and Koop and Potter (2007) propose time-varying parameter models where the size and duration of the process is modeled explicitly, and Clemens and Hendry (1996) propose intercept corrections. Alternative methods include forecast combinations (Timmermann, 2009) and Bayesian model averaging (Wright 2008, 2009). In her empirical analysis on the large dataset of macroeconomic predictors for inflation and output growth, Rossi (2011) finds that, among the estimation and forecasting methodologies robust to instabilities discussed above, forecast combinations with equal weights are the best.

Should one rely on in-sample measures of fit or out-of-sample measures of forecast performance when evaluating models? The short answer is that the two provide very different assessments of models' validity. Clark and McCracken (2005) show that out-of-sample forecasting procedures have more power in finding predictive ability than traditional in-sample Granger-causality tests in the presence of instabilities since they re-estimate the models' parameters over time. On the other hand, Inoue and Kilian (2005) argue that in-sample tests are based on a larger sample size than out-of-sample forecast tests, and thus may be better when designed appropriately. In fact, Clark and McCracken (2005) also show that the Granger-causality tests designed to be robust to instabilities (Rossi, 2005) performs even better. However, instabilities are only one of the sources of the difference between in-sample fit and out-of-sample forecasting performance. Giacomini and Rossi (2009) show that the difference depends on parameter instabilities, instabilities in other aspects of the forecasting model, as well as estimation uncertainty (including over-fitting). They also propose a "Forecast Breakdown" test to determine whether, empirically, models' in-sample fit differs from out-of-sample forecasting ability. How does one determine empirically why in-sample fit differs from out-of-sample forecasting ability? Rossi and Sekhposyan (2011a) provide a methodology to decompose the models' forecasting ability into asymptotically uncorrelated components that measure the contribution of instabilities, predictive content and over-fit in explaining the differences between in-sample fit and out-of-sample forecasting performance. Using their method, one can uncover what is the source of the difference between the two. In an empirical analysis on a large dataset of macroeconomic predictors, Rossi (2011) finds that most predictors for output growth and inflation experienced a forecast breakdown based on Giacomini and Rossi's (2009) test. She investigates the reasons for the breakdown using Rossi and Sekhposyan's (2011a) decomposition, and finds that, when forecasting inflation, instabilities are a major determinant when using interest rates as predictors, whereas when using real measures of activity (such as unemployment) not only there are instabilities but

the predictive content is misleading (that is, out-of-sample forecasting ability goes in the opposite direction relative to in-sample fit). When forecasting output growth, overfitting drives a wedge between in-sample and out-of-sample measures of performance even for predictors that have significant predictive content.

Part II. Special Empirical Issues in Forecasting in Macroeconomics

In the second part of the chapter we will focus on special issues that arise in practice when forecasting with macroeconomic data. Given the space constraints we will focus only on four issues that are especially important in practice, in particular: forecasting real activity with leading indicators, forecasting inflation, forecasting with real-time data, and including economic theory in forecasting.

4 Forecasting Real Economic Activity with Leading Indicators

An important goal of forecasting is to identify and evaluate leading indicators of real economic activity. Typically, the target variable for the leading indicator is either Gross Domestic Product (GDP) or industrial production or a composite index. For example, Burns and Mitchell define business cycles as co-movements, happening at the same time, in a large number of economic variables, which fluctuate from expansions and recessions and whose duration can last between 1 to 8 years (see Stock and Watson, 1999a). Since typically most measures of economic activity are highly correlated with GDP, one can use the latter as the measure of the business cycle, or an index (weighted average of several real economic variables) summarizing the joint co-movements among the real variables. An example of the latter is the Stock and Watson (1989) coincident index of economic activity based on industrial production, real disposable income, hours and sales.

The objective of the leading indicators literature is to predict the future values of such target variables, and successful leading indicators either: (i) successfully predict turning points while at the same time maintaining good predictive power across the various stages of the business cycle; for example, a good leading indicator should systematically anticipate the target variable with a stable lead time and be capable of predicting peaks and troughs with sufficient lead times; (ii) are economically and statistically significant predictors; for example, one would expect that good leading indicators have significant marginal predictive content and Granger-cause the target variable. In order for a leading indicator to have the aforementioned properties, it is often necessary to transform (or filter) the leading indicator to remove high frequency fluctuations and very long-run components that do not contain useful information on the business cycle. Typically, filtering the data is done by using Baxter

and King's (1999) bandpass filter, which allows research to focus on the frequencies of interest (see Stock and Watson, 1999a); note that Hodrick-Prescott filters, while removing very long frequencies, do keep very high frequency movements and therefore are not ideal.

Widely used leading indicators include model-free composite indexes as well as model-based indexes. The former apply statistical methods such as detrending, seasonal adjustment and removal of outliers to the candidate leading indicator series. An example is the composite coincident index (CCI) by the Conference Board. A major problem of model-free composite indexes is that it is not possible to construct measures of uncertainty around them, since they are not estimated models. Model-based leading indicators instead rely on either dynamic factor models or Markov-switching models to estimate the index, and the estimation procedure does provide a measure of uncertainty around the point forecast. The difference between the two is that the underlying unobservable state of the economy is modeled as a continuous variable in the former and as a discrete variable in the latter. Examples of the former include the dynamic factor models of Geweke (1977), Sargent and Sims (1977), Stock and Watson (1991, 1993) and Forni, Hallin, Lippi and Reichlin (2000); examples of the latter include Hamilton (1989), Diebold and Rudebusch (1989), Chauvet (1998) and Kim and Nelson (1998), among others. For a detailed technical description of these models, see Marcellino (2006). It is also possible to model directly the state of the business cycle (that is, the expansions/recessions) using probit or logit models, as in Stock and Watson (1991) or Estrella and Mishkin (1998), for example.

Marcellino (2006) provides an extensive empirical analysis of the success of leading indicators in practice as well as an excellent overview of the theoretical literature. He notes that most CCI indicators behave similarly for the U.S., and their peaks and throughs coincide with the recession dates identified by the NBER.

To evaluate the success of a leading indicator, it is common practice to compare its out-of-sample predictions with the realized values of the target variable, either the business cycle indicator (expansion/recession) or the state of the business cycle. In the latter case, the tests for forecast comparisons listed in Section 2 can be used; in the former case, one often constructs probability scores. For example, Diebold and Rudebusch (1989) have proposed using the quadratic probability score:

$$QPS = \frac{2}{P} \sum_{t=R+1}^{P} \left(P_{t+h|t} - R_{t+h} \right),$$

where R_{t+h} is a binary indicator indicating whether the economy is in a recession or expansion at time $t+h$, and $P_{t+h|t}$ is the probability of recession/expansion at time $t+h$ based on the

leading indicator using information up to time t. The lower the quadratic probability score, the better the forecast; a value close to zero indicates perfect forecasts.

Marcellino (2006) compares the success of several leading indicators at the one and six months ahead forecast horizon in an out-of-sample forecast exercise over the period 1989 to 2003, which includes two recessions.

Stock and Watson (1999a) examine comovements across many series and real GDP, which they think of as a proxy for the overall business cycle. They find large correlations between several variables and real GDP growth at a variety of leads and lags. Variables that Granger-cause output can be thought of as leading indicators for the business cycle, although the predictive ability of several such indicators is unstable over time, according to parameter stability tests in the Granger-causality regressions. For example, housing starts and new orders lead output growth.

Rossi and Sekhposyan (2011) evaluate various economic models' relative performance in forecasting future US output growth. They show that the models' relative performance has, in fact, changed dramatically over time, both for revised and real-time data. In addition, they find that most predictors for output growth lost their predictive ability in the mid-1970s, and became essentially useless in the last two decades.

More recent developments focus on developing better methods to handle data irregularities and improve nowcasts of macroeconomic variables in real time. Nowcasts are the current period forecasts of unobserved macroeconomic variables which will be revealed or revised subsequently. Giannone, Reichlin and Small (2008) develop a formal methodology to evaluate the information content of intra-monthly data releases for nowcasts of GDP growth. They show that their method can handle large data sets with staggered data-release dates and successfully tracks the information in real time.

5 Forecasting Inflation

In a classic paper, Stock and Watson (1999b) investigated one-year ahead forecasts of U.S. inflation. They focused on predicting inflation using the unemployment rate, according to the Phillips curve. In a sample of monthly data from 1959 to 1997, they found that the latter produces more accurate forecasts than other macroeconomic variables, including commodity prices and monetary aggregates. They also found statistical evidence of instabilities in the parameters of the Phillips curve. In addition, they show that, by including index measures of real activity, it is possible to improve inflation forecasts beyond those based on

unemployment.

Rossi and Sekhposyan (2011) evaluate various economic models' relative performance in forecasting inflation by taking into account the possibility that the models' relative performance can be varying over time. They show that fewer predictors are significant for forecasting inflation than forecasting output growth, and their predictive ability significantly worsened around the time of the Great Moderation.

Faust and Wright (2011) investigate subjective forecasts, which empirically appear to have an advantage over traditional model-based forecasts. They attempt to incorporate subjective forecast's information into model-based forecasts. They argue that, by exploiting boundary values provided by subjective forecasts (e.g. where inflation will be in the medium and long run), it might be possible to improve model-based forecasts. However, they find that, given good boundary values, models cannot improve much on trivial paths between the boundaries and, overall, perform equally well.

6 Forecasting with Real-Time Data

When conducting a forecasting exercise, typically researchers utilize data that they have collected at the time of the analysis for the macroeconomic variables of interest from the beginning of the sample up to the most recent data available. Then, using these data, they mimic what a forecaster would have done in the past to obtain a series of pseudo out-of-sample forecasts over time. However, these data are not necessarily the same as the data that were available at the time forecasters were actually producing a forecast. In fact, data are constantly subject to data revisions, changes, updates, which not only change the contemporaneous value of the variables but also their past values. To avoid this problem, Croushore and Stark (2001, 2003) introduced a database (the "Real-Time Data Set for Macroeconomists") that is available for free at the Federal Reserve Bank of Philadelphia. The database consists in a series of datasets of macroeconomic variables collected at each point in time (vintage); at each time, the dataset contains data of macroeconomic variables as they existed at that point in time, starting from the first datapoint up to the time of collection. Each dataset is a snapshot of the data that a forecaster would have been able to observe and use at each point in time. Using real-time data effectively allows to evaluate the actual forecasting ability of models or predictors.

Using real-time data is important in practice. Orphanides (2001) has shown that implications of macroeconomic models for studying the effects of monetary policy in-sample

might change if one uses real-time as opposed to revised data. Similarly, the empirical results of forecasting exercises might differ depending on whether the researcher uses real-time as opposed to revised data. In fact, Orphanides and Van Norden (2005) show that, although ex-post measures of the output gap are useful for predicting inflation, in real time the predictive content disappears. Edge, Laubach and Williams (2007) also find the same result when forecasting long-run productivity growth. Similarly, Faust, Rogers and Wright (2003) show that exchange rates are much more difficult to forecast using real-time data. Swanson (1996) finds that Granger causality test results change depending on whether one uses the first release of the data or the latest available data. Finally, Amato and Swanson (2001) show that money supply has predictive content for output only when using fully revised data rather than real-time data.

There are three main reasons why forecasts may be affected by revisions (Croushore, 2006). First, the data are different: real time databases provide vintages of data; thus, the data to be forecasted are different. In contrast, typical forecasting exercises are implemented and evaluated using the last revised data available at the time the data are collected. Secondly, the estimated coefficients change. In fact, the forecasting exercise can be implemented by either using the data available in the latest vintage of data (that is, what the forecaster would have had available at that point in time) or by using for each time the data that were immediately released at that time. Again, this is different from estimating coefficients using data that are available at the time the data are collected (fully revised data). Koenig, Dolmas and Piger (2003) find that it is best to use the first release of the data in forecasting rather than real-time data. Third, the model used for forecasting may be different as well (e.g., the number of lags estimated using real-time data might differ from that estimated in fully revised data). See Croushore (2006) for more details.

Finally, the fact that data are revised might be exploited to improve forecasts as well. For example, one might optimally take into account data revisions by using a Kalman filter or a state-space model. See Howrey (1978) for how to do so in practice.

7 Economic Theory and Forecasting

Can economic theory help us produce better forecasts? This is a fundamental question that has received little attention in the literature. In fact, a general picture that emerges from the recent literature on forecasting methodology is the almost exclusive focus on "a-theoretical" econometric models. This may be partly due to the fact that some of these methods have

proven to be quite successful, in particular those that provide a way to extract the information contained in large datasets while at the same time controlling the dimensionality of the problem, such as factor models (Stock and Watson, 2002, Forni, Hallin, Lippi, and Reichlin, 2000), Bayesian VARs (BVARs, e.g., Litterman, 1986; Giannone, Lenza, and Primiceri, 2010) and forecast combination methods such as Bayesian model averaging (Raftery, Madigan, and Hoeting, 1997; Aiolfi, Capistrán, and Timmermann, 2010) and bagging (Inoue and Kilian, 2008). On the other hand, there has been some call in the literature (particularly from researchers at central banks and policy institutions) for forecasts that are based on models that can "tell a story" (Edge, Kiley and Laforte, 2008). As a response, a small literature has investigated the forecasting performance of the new generation of dynamic stochastic general equilibrium (DSGE) models that are large scale theoretical models built on microfoundations with optimizing agents (e.g., Smets and Wouters (2003). See, e.g., Adolfson, Linde and Villani (2007), Wang (2009), Lees, Matheson and Smith (2010) and Edge, Kiley and Laforte (2010). The evidence from this literature is however still limited and the conclusions should be taken with caution as they are typically based on short evaluation samples that moreover do not include the most recent periods of recession. A more thorough evaluation of the forecasting performance of DSGE models is clearly needed.

In particular, Gurkaynak and Edge (2010) empirically assess the forecasting performance of the Smets and Wouters DSGE model. They explore how this model would have forecasted, from 1 to 8 quarters ahead, movements in inflation, output growth, and interest rates between 1997 and 2006 and evaluate how good forecasts based on DSGE models are using real-time data. They find that their forecasts are not worse than those based on several competing alternatives, including official forecasts such as the Greenbook and Blue Chip Consensus forecasts. Greenbook forecasts are judgemental forecasts produced by the Board of Governors of the Federal Reserve System; they are produced before each FOMC meeting, approximately eight times a year, and are made available to the public with a 5-year delay. Importantly, Greenbook forecasts are produced conditional on expected paths for financial variables such as the policy interest rate. The Blue Chip Consensus forecasts are forecasts of several important macroeconomic variables (such as output growth, inflation and interest rates) made monthly by a sample of approximately 50 banks and consulting firms; the average forecast across the sample is called the Consensus forecast. However, their absolute performance is very poor, especially during the Great Moderation period, since there was basically nothing to be forecasted. Similarly, Edge, Kiley and Laforte (2010) compared the forecasts from the Federal Reserve Board's DSGE model with alternative forecasts based on

time series models as well as Greenbook forecasts.

A further branch of the literature has looked for a middle ground and proposed "hybrid" approaches. One example in the context of model estimation is the use of theoretical models to construct priors for the parameters of econometric models (An and Schorfheide, 2007; Schorfheide, 2000), or the idea of constructing an optimal combination of the theoretical and econometric models (Del Negro and Schorfheide, 2004).

We will next discuss two different hybrid approaches applied to the specific case of out-of-sample forecasting.

7.1 Carriero and Giacomini (2010)

The idea of optimally combining the theoretical and the econometric model can be easily extended to the context of out-of-sample forecasting, as shown by Carriero and Giacomini (2010). The basic idea is to first acknowledge that the theoretical model can often be viewed as an econometric model with theory-based parameter restrictions. This is the case of the DSGE models considered in the literature mentioned above, since they are linearized DSGE models that can therefore be written as vector ARMA models subject to cross-equation restrictions implied by theory. The problem is therefore that of combining two forecasts in a non-standard framework, in which there is only one model but the forecaster has the option of either imposing the parameter restrictions implied by the theoretical model or of forecasting with the unrestricted model. The forecast combination problem is non-standard because the combination is between forecasts based on the same model but that use different estimators, which may yield perfectly correlated forecasts in large samples. This problem can be overcome by adopting the asymptotic framework of Giacomini and White (2006), where the estimation uncertainty is non-vanishing.

Carriero and Giacomini (2010) propose estimating the optimal combination weight out-of-sample and constructing an out-of-sample encompassing test. Let the forecast combination be $f_t^* = f_t^R + (1-\lambda)(f_t^U - f_t^R)$, and define the optimal weight λ^* as the one that minimizes a general expected out-of-sample loss

$$\begin{aligned} \lambda^* &= \arg\min_{\lambda \in R} E\left[\frac{1}{P}\sum_{t=R}^{T-h} L\left(y_{t+h}, f_t^*\right)\right] \\ &= \arg\min_{\lambda \in R} E\left[Q_P\left(\lambda\right)\right], \end{aligned} \tag{26}$$

which can be estimated by

$$\begin{aligned} \widehat{\lambda} &= \arg\min_{\lambda \in R} \frac{1}{P} \sum_{t=R}^{T-h} L\left(y_{t+h}, f_t^*\right) \\ &= \arg\min_{\lambda \in R} Q_P\left(\lambda\right). \end{aligned} \tag{27}$$

Under suitable assumptions, Carriero and Giacomini (2010) show that a test of the "usefulness" of the parameter restrictions for out-of-sample forecasting can be obtained by first constructing

$$\begin{aligned} t^U &= \frac{\sqrt{n}\left(\widehat{\lambda}-1\right)}{\widehat{\sigma}} \text{ and} \\ t^R &= \frac{\sqrt{n}\widehat{\lambda}}{\widehat{\sigma}}, \end{aligned} \tag{28}$$

where $\widehat{\sigma}$ is given by

$$\begin{aligned} \widehat{\sigma} &= \sqrt{\widehat{H}^{-1}\widehat{\Omega}\widehat{H}^{-1}}; \\ \widehat{H} &= \nabla_{\lambda\lambda} Q_n\left(\widehat{\lambda}\right); \\ \widehat{\Omega} &= \sum_{j=-p+1}^{p-1} \left(1-|\frac{j}{p}|\right) n^{-1} \sum_{t=R+j}^{T-h} s_t\left(\widehat{\lambda}\right) s_{t-j}\left(\widehat{\lambda}\right); \\ s_t\left(\widehat{\lambda}\right) &= \nabla_{\lambda} L\left(y_{t+h}, f_t^R + (1-\widehat{\lambda})(f_t^U - f_t^R)\right), \end{aligned} \tag{29}$$

where p is a bandwidth that increases with the sample size (Newey and West, 1987). Then the hypotheses $H_0^U : \lambda^* = 1$ (the unrestricted forecast is useless) and $H_0^R : \lambda^* = 0$ (the restricted forecast is useless) are rejected at a significance level α respectively when $\left|t^U\right| > c_{\alpha/2}$ and $\left|t^R\right| > c_{\alpha/2}$, with $c_{\alpha/2}$ indicating the $1-\alpha/2$ quantile of a $N(0,1)$ distribution. If both hypotheses are rejected, the estimated weight $\widehat{\lambda}$ yields the forecast combination that optimally exploits the theoretical restrictions, given the user-defined loss function.

Note that the same test can be used to combine forecasts based on any two competing estimators, and it is not necessary that the forecast f_t^U be based on the unrestricted models (in other words, f_t^U could be a forecast based on any other estimator, e.g., a-theoretical restrictions such as a BVAR or a random walk).

7.2 Giacomini and Ragusa (2011)

The approach discussed in the previous section requires one being able to construct forecasts based on the theoretical model. A model that fully specifies a likelihood for all the variables

of interest (e.g., in the multivariate case) is not however always available, and there might be a concern that a full-fledged DSGE is misspecified. One may for example be interested in asking whether the restrictions embedded in, say, a Euler equation are useful for forecasting, which is a difficult question to answer as the Euler equation does not provide a conditional likelihood that can be used for forecasting.

Giacomini and Ragusa (2011) propose adopting a hybrid approach to forecasting that starts from a forecast based on the econometric model (e.g., a BVAR or a factor model) and modifies it in a way that results in a forecast that satisfies theoretical restrictions written in the form of nonlinear moment conditions, such as, e.g., Euler equations or moment conditions implied by Taylor rules.

This is obtained by projection methods as follows. Suppose the theory-based moment restrictions for the vector y_{t+h} are

$$E_t[g(y_{t+h}, \theta_0)] = 0, \tag{30}$$

where the subscript t indicates conditioning on the information set at time t and θ_0 is assumed to be known, calibrated, or estimated on a different data set than the one used for forecasting (note that the moment conditions could possibly only involve a subset of y_{t+h}). One proceeds as follows:

1. Produce a sequence of h-step ahead density forecasts from an econometric model, $f_t(y_{t+h})$ for $t = R, ..., T-h$.

2. Project each $f_t(y_{t+h})$ onto the space of distributions that satisfy the moment condition $E_t[g(y_{t+h}, \theta_0)] = 0$. This yields a new density $\widetilde{f}_t(y_{t+h})$ given by:

$$\widetilde{f}_t(y_{t+h}) = f_t(y_{t+h}) \exp\left\{\eta_t + \lambda_t' g(y_{t+h}, \theta_0)\right\}. \tag{31}$$

   The new density by construction satisfies the moment condition (30).

3. For each t, estimate η_t and λ_t by (numerically) solving:

$$\begin{aligned} \lambda_t &= \min_{\lambda} \int f_t(x) \exp\left\{\lambda' g(x, \theta_0)\right\} dx \\ \eta_t &= \log\left\{\int f_t(x) \exp\left\{\lambda_t' g(x, \theta_0)\right\} dx\right\}^{-1}. \end{aligned} \tag{32}$$

The new forecast $\widetilde{f}_t(y_{t+h})$ can be interpreted as the density which, out of all the densities that satisfy the moment condition, is the closest to the initial density $f_t(y_{t+h})$ according

to a Kullback-Leibler measure of divergence. The paper shows that the theory-coherent density forecast $\widetilde{f}_t(y_{t+h})$ is weakly more accurate than the initial, a-theoretical forecast, when accuracy is measured by a logarithmic scoring rule, provided the moment restrictions are true at all time periods.

The method is an alternative to forecasting with full-fledged DSGE models and can be used to investigate the role of theory in forecasting in a variety of empirical applications. Because of the possibility of accommodating non-linearity in the moment conditions (a task that may be difficult to accomplish in a likelihood-based context) the method can also be used to ask whether incorporating the nonlinearity suggested by theory into forecasts can lead to improvements in accuracy in practice.

8 Conclusions

This chapter provides an overview of forecast methodologies and empirical results that are useful for macroeconomists and practitioners interested in forecasting using macroeconomic databases. A more detailed exposition of these techniques as well as other available techniques that we did not include due to space constraints is provided in Granger, Elliott and Timmermann (2006) and Elliott and Timmermann (2011).

References

[1] Amato, J.D. and N.R. Swanson (2001), "The Real Time Predictive Content of Money for Output," *Journal of Monetary Economics* 48, 3-24.

[2] An, S. and F. Schorfheide (2007): "Bayesian analysis of DSGE models," *Econometric Reviews*, 26, 113–172.

[3] Andrews, D.W. (1993), "Tests for Parameter Instability and Structural Change With Unknown Change Point," *Econometrica* 61(4), 821-856.

[4] Adolfson, M., J. Linde, and M. Villani (2007), "Forecasting performance of an open economy DSGE model", *Econometric Reviews*, 26, 289-328.

[5] Aiolfi, M., C. Capistrán and A. Timmermann (2010): "Forecast Combinations," CREATES Research Paper No. 2010-21.

[6] Amisano, G. and R. Giacomini (2007), "Comparing density forecasts via weighted likelihood ratio tests", *Journal of Business and Economic Statistics,* 25, 177-190.

[7] Andrews, D.W.K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation", *Econometrica,* 59, 817-858.

[8] Baxter, M. and R. King (1999), "Measuring the Business Cycle: Approximate Band-Pass filters for Economic Time Series", *Review of Economics and Statistics.*

[9] Bordo, M.D. and J.G. Haubrich (2008), "The Yield Curve as a Predictor of Growth: Long-Run Evidence, 1875-1997," *Review of Economics and Statistics* 90(1), 182-185.

[10] Burns, A.F. and W.C. Mitchell (1946), *Measuring Business Cycles*, NBER.

[11] Carriero, A. and R. Giacomini (2011): "How useful are no-arbitrage restrictions for forecasting the term structure of interest rates?", *Journal of Econometrics*, 164, 21-34.

[12] Chauvet, M. (1998), "An Econometric Characterization of Business Cycle Dynamics with Factor Structure and Regime Switching", *International Economic Review* 39(4), 969-996.

[13] Clark, T. and M. McCracken (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models", *Journal of Econometrics,* 105(1), 85-110.

[14] 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, 1-31.

[15] Clark, T. and M. McCracken (2009), "Nested forecast model comparisons: a new approach to testing equal accuracy", *mimeo.*

[16] Clark, T.E. and K.D. West (2007), "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics* 138, 291-311.

[17] Clements, M.P. and D.F. Hendry (1993), "On the Limitations of Comparing Mean Square Forecast Errors," *Journal of Forecasting* 12, 617-637.

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

[19] Corradi, V., N. Swanson and C. Olivetti (2001), "Predictive Ability with Cointegrated Variables", *Journal of Econometrics* 104(2), 315-358.

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

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

[22] Croushore, D. (2006), "Forecasting with Real-Time Macroeconomic Data", in: G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Elsevier, Amsterdam: North-Holland.

[23] Del Negro, M. and F. Schorfheide (2004): "Priors from general equilibrium models for VARs," *International Economic Review*, 643–673.

[24] Diebold, F. X. and J. Lopez (1996), "Forecast Evaluation and Combination" in G.S. Maddala and C.R. Rao (eds.), *Handbook of Statistics,* Amsterdam: North-Holland, 241-268.

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

[26] Diebold, F. X. and G. D. Rudebusch (1989), "Scoring the Leading Indicators", *The Journal of Business* 62(3), 369-391.

[27] Diebold, F. X. (2007), Elements of forecasting (Fourth Edition), South-Western College Publishing.

[28] Edge, R.M., M.T. Kiley, and J.-P. Laforte (2010), "A Comparison of Forecast Performance between Federal Reserve Staff Forecasts, Simple Reduced-Form Models, and a DSGE Model", *Journal of Applied Econometrics* 25, 720-54.

[29] Edge, R.M., T. Laubach and J.C. Williams, (2007), "Learning and Shifts in Long-Run Productivity Growth," *Journal of Monetary Economics* 54, 2421-2438.

[30] Elliott, G., I. Komunjer and A. Timmermann (2005), "Estimation and Testing of Forecast Rationality under Flexible Loss", *Review of Economic Studies*, 72, 1107-1125.

[31] Elliott, G. and U. Muller (2007), "Confidence Sets for the Date of a Single Break in Linear Time Series Regressions," *Journal of Econometrics* 141, 1196-1218.

[32] Elliott, G., C. Granger and A. Timmermann (2006), *Handbook of Economic Forecasting Vol. 1*, North Holland: Elsevier.

[33] Elliott, G. and A. Timmermann (2011), *Handbook of Economic Forecasting*, Volume 2, Elsevier-North Holland Publications.

[34] Estrella, A. and F. S. Mishkin (1998), "Predicting US Recessions: Financial Variables as Leading Indicators", *The Review of Economics and Statistics* 80(1), 45-61.

[35] Faust, J., J.H. Rogers and J.H. Wright (2003), "Exchange Rate Forecasting: the Errors We've Really Made," *Journal of International Economics* 60, 35-59.

[36] Faust, J. and J. Wright (2011), "Forecasting Inflation," in: G. Elliott and A. Timmermann, *Handbook of Economic Forecasting Vol. 2*, North Holland: Elsevier.

[37] Forni, M., Hallin, M., Lippi, M. and L. Reichlin (2000), "The Generalized Factor Model: Identification and Estimation", *The Review of Economics and Statistics* 82(4), 540-554.

[38] Geweke, J. (1977), "The Dynamic Factor Analysis of Economic Time Series", in: Aigner, D. J., and A. S. Goldberger (eds.), *Latent Variables in Socio-economic Models*, North Holland Publishing, ch. 19.

[39] Giannone, D., M. Lenza, and G. Primiceri (2010): "Prior selection for vector autoregressions" mimeo.

[40] Giannone, D., L. Reichlin and D. Small (2008), "Nowcasting: The real-time informational content of macroeconomic data," *Journal of Monetary Economics* 55(4), 665-676.

[41] Giacomini, R. and I. Komunjer (2005), "Evaluation and combination of conditional quantile forecasts", *Journal of Business and Economic Statistics*, 23, 416-431.

[42] 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.

[43] Giacomini, R. and B. Rossi (2009), "Detecting and Predicting Forecast Breakdowns," *Review of Economic Studies* 76(2), 2009.

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

[45] Giacomini, R. and B. Rossi (2011), "Model Comparisons in Unstable Environments", *mimeo*, Duke University.

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

[47] Giacomini, R. and G. Ragusa (2011), "Incorporating theoretical restirctions into forecasting by projection methods", mimeo.

[48] Granger, C.W.J. and P. Newbold (1986), *Forecasting Economic Time Series* (2nd ed.), New York: Academic Press.

[49] Gurkaynak, R. and R. Edge (2010), "How Useful Are Estimated DSGE Model Forecasts for Central Bankers?", *Brookings Papers on Economic Activity*, 209-259.

[50] Hamilton, J. D. (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle", *Econometrica* 57, 357-384.

[51] Hansen, P. R. (2005), "A test for superior predictive ability", *Journal of Business and Economic Statistics,* 23, 365-380.

[52] Hansen, P. R., A. Lunde, and J. M. Nason (2011), "The model confidence set", *Econometrica*, 79, 452-497.

[53] Hansen, P.R. and A. Timmermann (2011), "Choice of Sample Split in Out-of-Sample Forecast Evaluation," *mimeo*.

[54] Harvey, D.I., S.J. Leybourne and P. Newbold (1998), "Tests for Forecast Encompassing," *Journal of Business and Economic Statistics* 16 (2), 254-259.

[55] Howrey, E.P. (1978), "The Use of the Preliminary Data in Econometric Forecasting," *Review of Economics and Statistics* 60, 193-200.

[56] Inoue, A. and L. Kilian (2006): "On the Selection of Forecasting Models", *Journal of Econometrica*, 130, 273-306.

[57] Inoue, A. and L. Kilian (2008): "How Useful is Bagging in Forecasting Economic Time Series? A Case Study of US Consumer Price Inflation", *Journal of the American Statistical Association*, 103, 511–522.

[58] Inoue, A. and B. Rossi (2010), "Out of Sample Forecast Tests Robust to the Window Size Choice," *CEPR Working Paper* No. 8542 and *Philadelphia Fed Working Paper* No. 11-31.

[59] Kim, C.J. and C. R. Nelson (1998), "Business Cycle Turning Points, a New Coincident Index, and Tests of Duration Dependence Based on a Dynamic Factor Model with Regime Switching", *The Review of Economics and Statistics* 80, 188-201.

[60] Koenig, E., S. Dolmas and J. Piger, J. (2003), "The Use and Abuse of 'Real-Time' Data in Economic Forecasting," *Review of Economics and Statistics* 85, 618-628.

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

[62] Lees, K, T. Matheson, and C. Smith (2011): "Open economy forecasting with a DSGE-VAR: head to head with the RBNZ published forecasts", *International Journal of Forecasting*, 27, 512-528.

[63] Leitch, G. and E. J. Tanner (1991): "Economic forecast evaluation: profits versus the conventional error measures", *American Economic Review*, 81(3), 580 - 90.

[64] Litterman, R. (1986): "A statistical approach to economic forecasting," *Journal of Business & Economic Statistics*, 1–4.

[65] Marcellino, M. (2009), "Leading Indicators," in: G. Elliott, C. Granger and A. Timmermann, *Handbook of Economic Forecasting Vol. 1*, North Holland: Elsevier.

[66] McCracken, M. (2007), "Asymptotics for out-of-sample tests of Granger causality", *Journal of Econometrics,* 140, 719-752.

[67] Meese, R. and K. S. Rogoff (1983), "Exchange Rate Models of the Seventies. Do They Fit Out of Sample?," *Journal of International Economics* 14, 3-24.

[68] Mincer, J. and V. Zarnowitz (1969), "The Evaluation of Economic Forecasts," in: J. Mincer (ed.), *Economic Forecasts and Expectations*, New York: National Bureau of Economic Research, pp. 81–111.

[69] Newey, W. and K. West (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica,* 55, 703-708.

[70] Orphanides, A. and S. van Norden (2005). "The Reliability of Inflation Forecasts Based on Output Gaps in Real Time." *Journal of Money, Credit, and Banking* 37, 583–601.

[71] Orphanides, A. (2001). "Monetary Policy Rules Based on Real-Time Data," *American Economic Review* 91, 964–985.

[72] Pesaran, M.H. and A. Timmermann (2002), "Market Timing and Return Prediction Under Model Instability," *Journal of Empirical Finance* 9(5), 495–510.

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

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

[75] Raftery, A., D. Madigan, and J. Hoeting (1997): "Bayesian model averaging for linear regression models" *Journal of the American Statistical Association*, 179–191.

[76] Romano, J. P. and M. Wolf (2005), "Stepwise multiple testing as formalized data snooping", *Econometrica*, 73, 1237-1282.

[77] Rossi, B. (2005b), "Testing Long-Horizon Predictive Ability with High Persistence and the Meese-Rogoff Puzzle", *International Economic Review* 46(1), 61-92.

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

[79] Rossi, B. (2011), "Advances in Forecasting under Instabilities", in G. Elliott and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Volume 2, Elsevier-North Holland Publications.

[80] Rossi, B. and T. Sekhposyan (2010), "Have Models' Forecasting Performance Changed Over Time, and When?," *International Journal of Forecasting* 26(4).

[81] Rossi, B. and T. Sekhposyan (2011a), "Understanding Models' Forecasting Performance," *Journal of Econometrics* 164, 158-172.

[82] Rossi, B. and T. Sekhposyan (2011b), "Forecast Optimality Tests in the Presence of Instabilities," *mimeo*, Duke University.

[83] Sargent, T.J. and C.A. Sims (1977), "Business Cycle Modeling Without Pretending to Have Too Much a Priori Economic Theory", in: C. Sims et al. (eds.), *New Methods in Business Cycle Research*, Federal Reserve Bank of Minneapolis.

[84] Schorfheide, F. (2000): "Loss function-based evaluation of DSGE models," *Journal of Applied Econometrics*, 15, 645–670.

[85] Schrimpf, A. and Q.W. Wang (2010), "A Reappraisal of the Leading Indicator Properties of the Yield Curve Under Structural Instability," *International Journal of Forecasting* 26(4), 836-857.

[86] Smets, F. and R. Wouters (2003): "An estimated dynamic stochastic general equilibrium model of the euro area", *Journal of the European Economic Association*, 1, 1123–1175.

[87] Stock, J.H. and M.W. Watson (1991), "A Probability Model of the Coicident Indicators", in K. Lahir and G. H. Moore (eds.), *Leading Economic Indicators: New Approaches and Forecasting Records*, Cambridge University Press.

[88] Stock, J.H. and M.W. Watson (1993), "A Procedure for Predicting Recessions with Leading Indicators: Econometric Issues and Recent Experience", in: J.H. Stock and M.W. Watson (eds.), *Business Cycles, Indicators, and Forecasting*, The University of Chicago Press, 95-153.

[89] Stock, J.H. and M.W. Watson (1999a), "Business Cycle Fluctuations in US Macroeconomic Time Series", in: J.B. Taylor and M. Woodford (eds.), *Handbook of Macroeconomics* Vol. 1A, Elsevier Science, North-Holland.

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

[91] Stock, J. and M. Watson (2002): "Forecasting using principal components from a large number of predictors," *Journal of the American Statistical Association*, 97, 1167–1179.

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

[93] Stock, J.H. and M.W. Watson (2007), "Has Inflation Become Harder to Forecast?," *Journal of Money, Credit and Banking* 39 (1), 3–34.

[94] Swanson, N.R. (1996), "Forecasting Using First Available Versus Fully Revised Economic Time Series Data," *Studies in Nonlinear Dynamics and Econometrics* 1, 47-64.

[95] Swanson, N.R. (1998), "Money and Output Viewed Through a Rolling Window," *Journal of Monetary Economics* 41, 455-473.

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

[97] Timmermann, A. (2006), "Forecast Combinations," in: G. Elliott, C. Granger and A. Timmermann, *Handbook of Economic Forecasting Vol. 1*, North Holland: Elsevier.

[98] Wang, M.C. (2009), "Comparing the DSGE model with the factor model: an out-of-sample forecasting experiment", *Journal of Forecasting,* 28, 167-182.

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

[100] West, K. D., H. J. Edison, and D. Cho (1993): "A Utility-Based Comparison of Some Models of Exchange Rate Volatility," *Journal of International Economics*, 35, 23–45.

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

[102] Wheelock, D.C. and M.E. Wohar (2009), "Can the Term Spread Predict Output Growth and Recessions? A Survey of the Literature," *Federal Reserve Bank of St. Louis Review* 91(5), 419-440.

[103] White, H. (2000), "A reality check for data snooping", *Econometrica,* 68, 1097-1127.

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

[105] Wright, J.H. (2009), "Forecasting US Inflation by Bayesian Model Averaging," *Journal of Forecasting* 28(2), 131-144.