Implementing Tests For Forecast Evaluation in the Presence of Instabilities

Barbara Rossi[*] Matthieu Soupre[†]

February 10, 2017

Abstract

In this paper, we review methodologies to fix the size distortions of tests for forecast evaluation when the forecasts display instabilities by implementing forecast evaluation tests that are robust to instabilities. We discuss tests for relative as well as absolute forecast evaluation, and describe two Stata procedures that implement the tests.

Keywords: Forecasting, instabilities, structural change.
J.E.L. Codes: C22, C52, C53.[1]

[*] ICREA Professor, Univ. Pompeu Fabra, Barcelona GSE and CREI. Address: Universitat Pompeu Fabra, C/ Ramon Trias Fargas 25-27, 08005 Barcelona, Spain. E-mail: barbara.rossi@upf.edu

[†] Univ. Pompeu Fabra. Address: Universitat Pompeu Fabra, C/ Ramon Trias Fargas 25-27, 08005 Barcelona, Spain. E-mail: matthieu.soupre@upf.edu

[1] *Acknowledgements:* We are grateful to Greg Ganics. Barbara Rossi gratefully acknowledges financial support from the European Research Council under Grant No. 615608.

1 Introduction

It is often of interest to test models' forecasting ability. In particular, researchers are often interested in determining which of two competing forecasting models predicts the best; such tests are known as "tests of relative forecast comparisons". Examples of such tests include: Diebold and Mariano (1995), West (1996) and Clark and McCracken (2001), among others. Another typical but very different type of forecasting ability test involves evaluating whether forecasts fulfill some minimal requirements, such as being unbiased or producing forecast errors that are unpredictable using any information available at the time the forecast is made; such tests are typically referred to as "tests of absolute forecasting performance". Examples of such tests include, among others: Mincer and Zarnowitz (1969) and West and McCracken (1998). While both tests of relative and absolute forecast performance are tests of forecasting ability, they differ substantially in their theoretical properties as well as their purpose: in fact, the former are used to compare forecasting models, the latter are used to evaluate one specific forecasting model.

When applying tests of forecasting ability to macroeconomic time series data, researchers face an important practical problem. It is well-known that economic time series data are prone to instabilities. A recent example is the Great Recession of 2007-2009, when several macroeconomic relationships changed drastically. For example, interest rates lost their ability to predict output growth during that time, while credit spreads became useful predictors (Ng and Wright, 2013). Similarly, Rossi (2013b) finds severe instabilities in exchange rate forecasting models. More generally, Stock and Watson (1996) investigated the presence of instabilities in several different forecasting models in a large dataset of key macroeconomic variables (76 representative U.S. monthly postwar macroeconomic series) using formal testing procedures. The tests for structural breaks that they used include the Quandt (1960) and Andrews (1993) QLR test, the Mean and Exponential Wald test statistics by Andrews and Ploberger (1994), the Ploberger and Kramer (1992) maximal OLS CUSUM statistic, and Nyblom's (1989) test. Their analysis uncovered substantial and widespread instabilities in many economic time series. Thus, when testing models' forecasting ability, it is potentially important to allow their forecasting ability to change over time. In fact, traditional tests of forecast evaluation are not reliable in the presence of instabilities, which may lead to incorrect inference. The problem arises because traditional tests assume stationarity, an assumption that is violated in the presence of instabilities.

The contribution of this paper is to propose and discuss Stata commands that illustrate how to test forecast unbiasedness/rationality and how to test competing models' forecasting performance in a way that is robust to the presence of instabilities. The tests are based on the methodologies developed by Giacomini and Rossi (2010) and Rossi and Sekhposyan (2016), and discussed thoroughly by Rossi (2013a). The Stata commands we present implement both Rossi and Sekhposyan's

(2016) Fluctuation Rationality test as well as Giacomini and Rossi's (2010) Fluctuation test. The tests are separately presented, as they address very different concerns. For instance, the Rossi and Sekhposyan's (2016) Fluctuation Rationality test allows researchers to evaluate whether the forecasts fulfill some minimal requirements (such as being unbiased and being highly correlated with the ex-post realized value) in environments characterized by instabilities; hence, such tests are "tests of absolute forecasting performance robust to instabilities". Giacomini and Rossi's (2010) Fluctuation test instead allows researchers to detect which model forecasts the best in unstable environments, and hence it is a "test of relative forecasting performance robust to instabilities". In the presence of instabilities, the latter tests are more powerful than traditional tests and provide a visual illustration of when predictive ability appears or breaks down in the data. For each test, first we introduce the test, then present the Stata commands that implement it, and finally discuss a simple empirical exercise to illustrate the output of the tests and how to interpret the results. The codes are freely available in Stata in the zipped file FOREC_INSTAB from the Stata SSC archive.[2]

In Section 2, we establish the notation and definitions. Section 3 discusses Rossi and Sekhposyan's (2016) Fluctuation rationality test and Section 4 discusses Giacomini and Rossi's (2010) Fluctuation test; sub-sections explain the syntax of the Stata commands and demonstrate their usage.

2 Notation and Definitions

We first introduce the notation and discuss the assumptions about the data, the models and the estimation procedures. We are interested in evaluating h-step ahead forecasts for the variable y_t, which we assume to be a scalar for simplicity. The evaluation can be relative (i.e. comparing the relative forecasting performance of competing models) or absolute (i.e. evaluating the forecasting performance of a model in isolation).

We assume that the researcher has a sequence of P h-step-ahead out-of-sample forecasts for two models, denoted respectively by $y_{t,h}^{(1)}$ and $y_{t,h}^{(2)}$, made at time t, where $t = 1, ..., P$.[3] Finally, let the forecast error associated with the h-step-ahead forecast made at time t by the first model be denoted by $v_{t,h}$.[4]

[2] The codes are also available at the website: barbararossi.eu

[3] The models' parameters are estimated either using a fixed or a rolling scheme, where the size of the sample used to estimate the parameters is fixed. This rules out recursive estimation schemes.

[4] For example, in a simple linear regression model with h-period lagged $(k \times 1)$ vector of regressors x_t, where $E_t y_{t+h} = x_t' \gamma$, the forecast at time t is: $y_{t,h} = x_t' \widehat{\gamma}_{t,R}$ and the forecast error is: $v_{t,h} = y_{t+h} - x_t' \widehat{\gamma}_{t,R}$, where $\widehat{\gamma}_{t,R}$ is the estimated vector of coefficients.

3 Tests of Relative Forecast Comparisons Robust to Instabilities

3.1 Giacomini and Rossi's (2010) Fluctuation Test

The Fluctuation test compares the relative forecasting performance of competing models over time, where the performance is judged based on a loss function chosen by the forecaster. For a general loss function $L(.)$, the researcher has available a sequence of P out-of-sample forecast loss differences, $\{\Delta L_{t,h}\}_{t=1}^{P}$,where $\Delta L_{t,h} \equiv L_{t,h}^{(1)} - L_{t,h}^{(2)}$, which depend on the realizations of the variable, y_{t+h}. For example, for the traditional quadratic loss associated with Mean Squared Forecast Error (MSFE) measures, $L_{t,h}^{(1)} = v_{t+h}^{2}$ and $\Delta L_{t,h}$ is the difference between the squared forecast errors of the two competing models.[5] As the square loss function is the most widely used loss function in practice, this is the one we implement in the Stata procedure described below.

Giacomini and Rossi (2010) define the local relative loss for the two models as the sequence of out-of-sample loss differences computed **over rolling windows of size** m:

$$m^{-1} \sum_{j=t-m+1}^{t} \Delta L_{j,h}, \ t = m, m+1, ..., P. \tag{1}$$

They are interested in testing the null hypothesis of equal predictive ability at each point in time:

$$H_0 : E\left[\Delta L_{t,h}\right] = 0 \text{ for all } t,$$

and the alternative can be either $E\left[\Delta L_{t,h}\right] \neq 0$ (two-sided alternative) or $E\left[\Delta L_{t,h}\right] > 0$ (one-sided alternative). Their Fluctuation test statistic is the largest value over the sequence of the relative forecast error losses defined in eq. (1):

$$\max_t \mathcal{F}_{t,m}^{OOS}, \tag{2}$$

where

$$\mathcal{F}_{t,m}^{OOS} = \widehat{\sigma}^{-1} m^{-1/2} \sum_{j=t-m+1}^{t} \Delta L_{j,h}, \quad t = m, m+1, ..., P, \tag{3}$$

where $\widehat{\sigma}^2$ is a heteroskedasticity and autocorrelation consistent (HAC) estimator of the long run variance of the loss differences (Newey and West, 1987). The null hypothesis is rejected against the two-sided alternative $E\left[\Delta L_{t,h}\left(\widehat{\gamma}_{t,R}, \widehat{\beta}_{t,R}\right)\right] \neq 0$ when $\max_t \left|\mathcal{F}_{t,m}^{OOS}\right| > k_{\alpha,\mu}$, where the critical value $k_{\alpha,\mu}$ depends on the choice of μ, which is the size of the rolling window relative to the number of out-of-sample loss differences P, formally $m = [\mu P]$. Note also that $\mathcal{F}_{t,m}^{OOS}$ is simply a traditional test of equal predictive ability computed over a sequence of rolling out-of-sample windows of size m.

[5] In fact, $P^{-1} \sum_{t=1}^{P} \Delta L_{t,h}$ is exactly the MSFE.

3.2 The giacross Command

The `giacross` command is the Stata equivalent to the Matlab command written by Giacomini and Rossi (2010). The general syntax of the command giacross is:

`giacross` *data forecast1 forecast2*, `window`(*size*) `alpha`(*level*) [`nw`(*bandwidth*) `side`(#)].

data contains the realizations of the target variable (the realized values against which each forecast is compared), that is y_{t+h} as per the notation in Section 3.1, $t = 1, 2, ..., P$, where P is the number of forecasts available.

forecast1 and *forecast2* each contain the forecasts from the competing models that are tested, that is $y_{t,h}^{(1)}$ and $y_{t,h}^{(2)}$. Note that the inputs of the function are simply the forecasts: there is no need to input the models' parameter estimates in the procedure. In fact, as explained in Giacomini and Rossi (2010), the test can also be implemented if the researcher does not know the models that generated the forecasts (as, for example, in the case of survey forecasts).

size corresponds to the size of the window in the implementation **of the test, that is** m**.**

level equals the significance level of the test, either 0.05 for a 5% level or 0.10 for 10%.

bandwidth is an option allowing the user to choose the truncation lag used in the estimation of the variance $\widehat{\sigma}^2$. If no bandwidth is specified, the truncation lag is automatically determined using the Schwert (1987) criterion.

The *side* option takes the value 1 or 2 and specifies if the null is compared to a one- or two-sided alternative, respectively. If the alternative is one-sided, the alternative hypothesis is that the first model forecasts worse than the second model. If the alternative is two-sided, models' forecasts are significantly different from each other under the alternative.

The `giacross` command returns the following items:

`r(tstat_sup)` The maximum absolute value of the (rolling) test statistic over the sample, i.e. the value of eq. (2).

`r(cv)` The critical value of the test.

`r(level)` The significance level of the test specified by the user.

`r(cmd)` The name of the command used, namely "giacross."

`r(cmdline)` The whole command line input by the user `r(testtype)` Whether the test is one- or two-sided.

`r(RollStat)` The whole history of the rolling test statistic. A variable `FlucTest` is also created in the dataset if needed.

Finally, the `giacross` command automatically produces a graph plotting the test statistic against time with the critical value(s) implied by the level specified. We show such a graph in the example in the next section.

3.3 Example of Practical Implementation in Stata

The following sample code runs a sample application of the Giacomini-Rossi test using the `giacross` command.

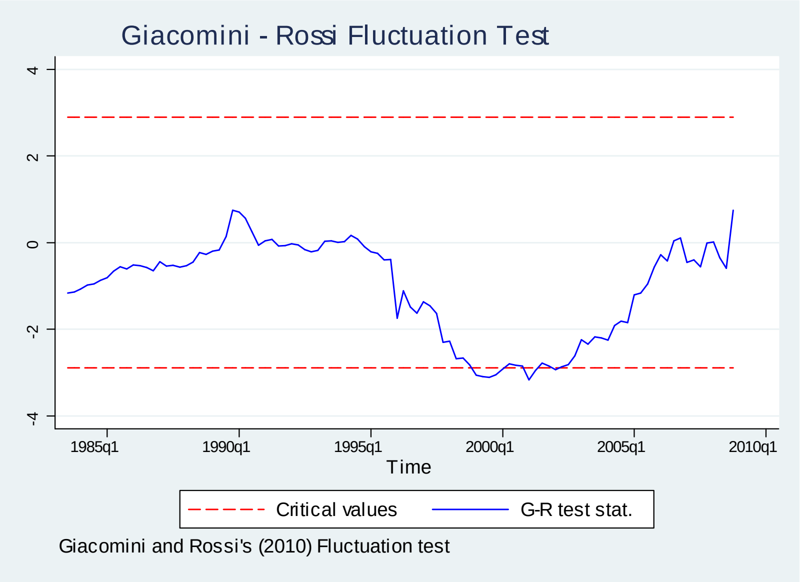
```
set more off
insheet using rosssekh_test_data1new.csv, clear

generate year = int(pdate)
generate quarter = (pdate - int(pdate))*4 + 1
generate tq = yq(year, quarter)
format tq %tq
tsset tq

* lag length set to 3, default 2-sided test
giacross realiz forc spf, window(60) alpha(0.05) nw(3)
dis "The value of the test statistic is " r(tstat_sup)
dis "The critical value is " r(cv) " at significance level " r(level)
```

The dataset in test_data1new.csv includes quarterly realizations of inflation for the US starting in 1968Q4 until 2008Q4, as well as the Greenbook (labeled "forc") and the Survey of Professional Forecast (SPF) nowcasts (labeled "spf) for the same variable. The first few lines of the code inputs the data and construct the time series of dates.

- The code returns the following output (with a graph showing the history of the rolling test statistic): the value of the test statistic ($\max_t \left|\mathcal{F}_{t,m}^{OOS}\right|$) is 3.1647 and is larger than the critical value at the 5% significance level, equal to 2.89. Therefore, we reject the null hypothesis that the models' forecasting performance is the same in favor of the alternative that the first model forecasts significantly better.

- Figure 1 provides a visual interpretation. The figure plots the sequence of $\mathcal{F}_{t,m}^{OOS}$ over time (depicted by a continuous line), and shows that it is clearly outside the critical value lines (± 2.89, depicted by the dashed lines). The strongest evidence against the null appears to be around the beginning of 2000s: this is when the empirical evidence in favor of the first model is the strongest.

Figure 1. Giacomini and Rossi's Test



Notes to the figure. The figure depicts $\mathcal{F}_{t,m}^{OOS}$ from eq. (3) as a function of time (t) for the first example in Section 3.

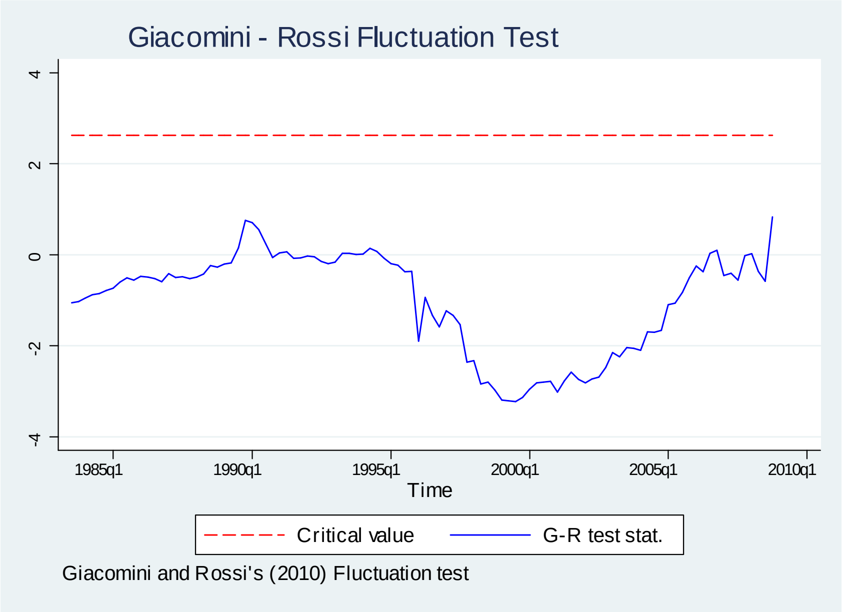
We also include an example of the one-sided version of the test using the following sample code:

```
* automatic lag length selection based on Schwert criterion, one-sided test
giacross realiz forc spf, window(60) alpha(0.05) side(1)
dis "The value of the test statistic is " r(tstat_sup)
dis "The critical value is " r(cv) " at significance level " r(level)
```

- The code returns the following output: the value of the test statistic is 0.8267, and the critical value is 2.624 at significance level 0.05. The test does not reject the null hypothesis that the two models forecast performance is the same against the alternative that the first model forecasts worse than the second model.

- The output also includes a plot of the models' relative forecasting performance over time, depicted in Figure 2.

Figure 2. Giacomini and Rossi's Test



Notes to the figure. The figure depicts $\mathcal{F}_{t,m}^{OOS}$ from eq. (3) as a function of time (t) for the second example in Section 3.

3.4 A Comparison With Traditional Tests

A common test used to compare models' forecasting performance is the Diebold and Mariano (1995) and West (1996) test. The Diebold and Mariano (1995) and West (1996) test statistic is:

$$\mathcal{DMW}_P = \widehat{\sigma}^{-1} P^{-1/2} \sum_{t=1}^{P} \Delta L_{j,h},$$

where $\widehat{\sigma}^2$ is a HAC estimator of the long run variance of the loss differences. The test is designed to test the (unconditional) null hypothesis $H_0 : E[\Delta L_{t,h}] = 0$ and, under the null, has an asymptotic standard normal distribution.

The $\mathcal{DMW}_P$ test can be obtained in Stata using the following command:

```
* Diebold Mariano comparison of forecast accuracy
dmariano realiz forc spf, max(3)
```

- The command returns the following output: the p-value is 0.2177 so the test does not reject the null of equal forecast accuracy of the two forecasts at the 0.05 significance level.

- Importantly, note that the empirical conclusions are very different from those that a researcher would obtain by using the Fluctuation test. In fact, the $\mathcal{DMW}_P$ test ignores the time variation in the relative forecasting performance over time, visible in Figure 1: instead, it averages across all the out-of-sample observations, thus losing power to detect differences in the models' forecasting performance.

4 Tests of Absolute Forecasting Performance Robust to Instabilities

4.1 Rossi and Sekhposyan's (2016) Fluctuation Rationality Test

Tests for forecast rationality evaluate whether forecasts satisfy some "minimal" requirements, such as being an unbiased predictor or being uncorrelated with any additional information available at the time the forecast was made. Thus, traditional tests of forecast rationality (such as Mincer and Zarnowitz, 1969, and West and McCracken, 1998) verify that forecast errors have zero mean or that they are uncorrelated with any other variable known at the time the forecast was made. However, they assume stationarity and are thus invalid in the presence of instabilities.

In order to make the tests robust to instabilities, Rossi and Sekhposyan (2016) propose to estimate the following forecast rationality regressions in rolling windows (of size m):

$$v_{j,h} = g_j' \cdot \theta + \eta_{j,h}, \; j = t-m+1, ..., t \tag{4}$$

where the forecast errors denoted by $v_{j,h}$ refer to an h-step ahead out-of-sample forecast made at time j using data available up to that point in time and may depend on parameter estimates; g_j is an $(\ell \times 1)$ vector function of period j data (which can also possibly be a function of the models' parameter estimates), θ is an $(\ell \times 1)$ parameter vector, and $\eta_{j,h}$ is the residual in the regression. The regression in eq. (4) is thus estimated in rolling windows of size m: at time t, the researcher uses data from $t-m+1$ to t to obtain the parameter estimate, $\widehat{\theta}_t$; by repeating the procedure at times $t = m, m+1, ..., P$, the researcher obtains a sequence of parameter estimates over time.

Rossi and Sekhposyan's (2016) main interest is testing forecast rationality in the presence of instabilities. In fact, in the presence of instabilities, tests that focus on the average out-of-sample performance of a model may be misleading, as they may average out instabilities. Thus, the hypothesis to be tested is:

$$H_0 : \theta_t = \theta_0 \text{ vs. } H_A : \theta_t \neq \theta_0, \; \forall \; t, \tag{5}$$

where $\theta_0 = 0$ and θ_t is the time-varying parameter value.

The framework in equation (4) is quite general; here we focus on tests of forecast unbiasedness ($g_t = 1$); tests of forecast efficiency (g_t is the forecast itself); and tests of forecast rationality (g_t includes both the forecast and 1).[6] We refer to all these tests under the maintained assumption

[6] In general, g_t may also contain any other variable known at time t which was not included in the forecasting model); the framework in equation (4) also potentially includes tests of forecast encompassing (g_t is the forecast of the encompassed model) and serial uncorrelation tests (g_t is the lagged forecast error).

that $\theta_0 = 0$ as "tests for forecast rationality." The zero restriction on the parameter under the null hypothesis ensures that the forecast errors are truly unpredictable given the information set available at the time the forecast is made.

Rossi and Sekhposyan (2016) propose the following "Fluctuation Rationality" test:

$$\max_t \mathcal{W}_{t,m}, \tag{6}$$

where

$$\mathcal{W}_{t,m} = m\widehat{\theta}_t' \; \widehat{V}_\theta^{-1}\widehat{\theta}_t, \text{ for } t = m, m+1, ..., P, \tag{7}$$

is the Wald test in regressions computed at time t over rolling windows of size m and based on the parameter estimate $\widehat{\theta}_t$, which is sequentially estimated in regression (4) and $\widehat{V}_\theta$ is a HAC estimator of the asymptotic variance of the parameter estimates in the same rolling windows.

Here we focus on the version of the Rossi and Sekhposyan (2016) test where either parameter estimation error is irrelevant, or the forecasts are model free, or the models' parameters are rollingly re-estimated in a finite window of data, although their test is valid in more general situations as well (see Rossi and Sekhposyan, 2016).

The null hypothesis is rejected if $\max_t \mathcal{W}_{t,m} > \kappa_{\alpha,\ell}$, where $\kappa_{\alpha,\ell}$ is the critical value at the $100\alpha\%$ significance level with the number of restrictions equal to ℓ.

4.2 The rosssekh Command

The `rosssekh` command is the Stata equivalent to the Matlab command written by Rossi and Sekhposyan (2016). The general syntax of the command `rosssekh` is:

`rosssekh` *data forecast*, `window`(*size*) `alpha`(*level*) [`nw`(*bandwidth*)].

data contains the realizations of the target variable (the realized values against which each forecast is compared), that is y_{t+h} in the notation of Section 3.1, $t = 1, 2, ..., P$, where P is the number of forecasts available.

forecast is g_t in our notation.

size corresponds to the size of the window in the implementation of the test, that is, m.

level equals the significance level of the test, either 0.05 for a 5% level or 0.10 for 10%.

bandwidth is an option allowing the user to choose the truncation lag used in the HAC variance estimation. If no bandwidth is specified, the truncation lag is automatically determined using the Schwert (1987) criterion.

The `rosssekh` command returns the following items:

r(tstat_sup) contains the maximum value attained by the (rolling) test statistic over the sample, eq. (6).

r(cv) is the matrix of critical values of the test at the level specified by the user.

r(level) is the level of the test specified by the user.

r(cmd) is the name of the command used, namely “rosssekh.”

r(cmdline) is the whole input command line.

r(RollStat) is the whole time series of the rolling test statistic. A variable RossSekhTest is also created in the dataset.

r(CV) is the critical values of the test.

4.3 Example of Practical Implementation in Stata

The following sample code runs a sample application of the Rossi-Sekhposyan test using the rosssekh command. We focus on evaluating forecast rationality of Greenbook forecasts (labeled "forc").

```
set more off

* change path according to the location of the file test_data1new.csv
insheet using rosssekh_test_data1_new.csv, clear
generate year = int(pdate)
generate quarter = (pdate - int(pdate))*4 + 1
generate tq = yq(year, quarter)
format tq %tq
tsset tq

* window size is 60 observations long, significance level is 0.05, lag length set to 3
rosssekh realiz forc, window(60) alpha(0.05) nw(3)

dis "The value of the test statistic is " r(tstat_sup)
dis "The critical value is " r(cv) " at significance level " r(level)
```

- The code returns the following output (with a graph showing the history of the rolling test statistic): the test statistic ($\max_t \mathcal{W}_{t,m}$) reaches a maximum of 38.90 for a critical value of 10.9084. The test does reject the null hypothesis of forecast rationality.

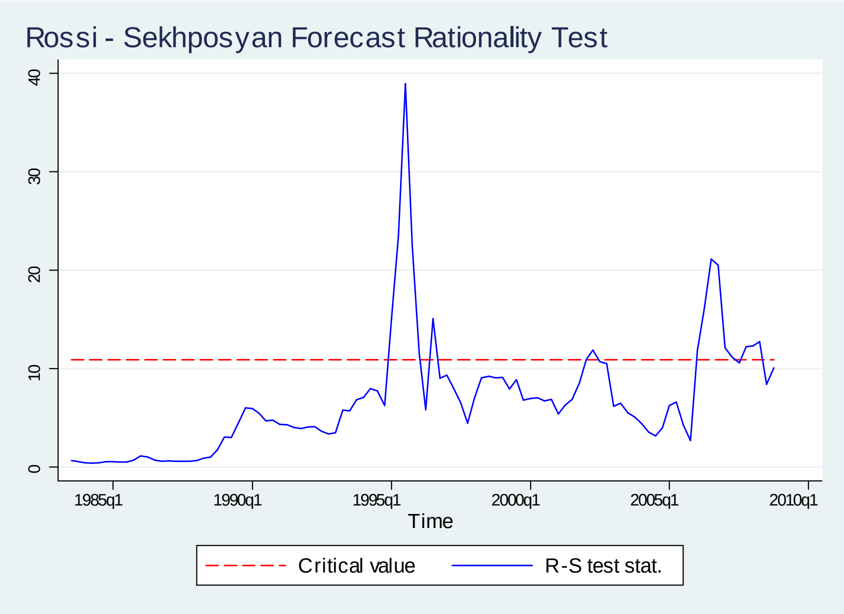
- Figure 3 provides a visual interpretation. The figure plots the sequence of $\mathcal{W}_{t,m}$ over time (depicted by a continuous line), and shows that it is clearly outside the critical value line (depicted by the dashed line). The strongest evidence against the forecast rationality appears to be around the beginning of 1995.

A similar result can be obtained by using an automatic lag length selection using the following sample code:

```
* automatic lag length selection, integer part of window^0.25
rosssekh realiz forc, window(60) alpha(0.05) nw(0)

dis "The value of the test statistic is " r(tstat_sup)
dis "The critical value is " r(cv) " at significance level " r(level)
```

- The code returns the following output: the test statistic reaches a maximum of 27.14 for a critical value of 10.9084. The test does reject the null hypothesis of forecast rationality. In this case, the plot is qualitatively similar to that in Figure 3, therefore we do not report it to save space.

Figure 3. Rossi and Sekhposyan's Test



Notes to the figure. The figure plots Rossi and Sekhposyan's sequence of test statistics ($\mathcal{W}_{t,m}$) over time.

4.4 A Comparison With Traditional Tests

A commonly used test to evaluate the forecasting performance of a model is the Mincer and Zarnowitz (1969) test. The Mincer and Zarnowitz (1969) test statistic $\mathcal{MZ}_P$ is a simple F-test in the regression: $v_{j,h} = g_j' \cdot \theta + \eta_t$, $j = 1, ..., P$:

$$\mathcal{MZ}_P = P\widehat{\theta}_P'\widehat{V}_\theta^{-1}\widehat{\theta}_P,$$

where $\widehat{V}_\theta$ is a HAC estimator of the asymptotic variance of the parameter estimates.

The test is designed to test the (unconditional) null hypothesis that $H_0 : \theta = 0$ and, under the null, has an asymptotic chi-square distribution. Again, notice that it, unlike $\max_t \mathcal{W}_{t,m}$, it is not robust to instabilities.

The $\mathcal{MZ}_P$ test can be obtained in Stata from a simple F-test using the following command:[7]

```
* Mincer Zarnowitz regression for systematic forecast bias
generate fcsterror=realiz-forc
newey fcsterror forc, lag(3)
```

- The command returns the following output: the $\mathcal{MZ}_P$ test statistic is 0.60 and its p-value is 0.4386, so the test does not reject the null at the 0.05 significance level.

- Again, in this case as well, the empirical conclusions are very different from those that a researcher would obtain by using the Fluctuation Rationality test. In fact, the $\mathcal{MZ}_P$ test ignores the time variation in the relative forecasting performance over time, visible in Figure 2: by averaging across all the out-of-sample observations, it misses the lack of forecast rationality that appears sporadically in time.

5 References

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

Andrews, D.W. and W. Ploberger (1994) "Optimal Tests When a Nuisance Parameter is Present Only Under the Alternative," *Econometrica* 62, 1383-1414.

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

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

[7] We used a lag length equal to 3 to compare the results with those in the previous example.

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

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

Newey, W.K. and K.D. West (1987) "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica* 55(3), 703-708.

Ng, S. and J.H. Wright (2013) "Facts and Challenges from the Great Recession for Forecasting and Macroeconomic Modeling". *Journal of Economic Literature* 51(4), 1120-54.

Nyblom, J. (1989), "Testing for the Constancy of Parameters Over Time", *Journal of the American Statistical Association* 84, 223-230.

Ploberger, W. and W. Kramer (1992) "The CUSUM Test With OLS Residuals", *Econometrica* 60, 271-286.

Quandt, R.E. (1960) "Tests of the Hypothesis That a Linear Regression System Obeys Two Separate Regimes", *Journal of the American Statistical Association* 55, 324-330.

Rossi, B. (2013a) "Advances in Forecasting Under Instabilities". In: Elliott G, Timmermann A. (eds.), *Handbook of Economic Forecasting*, Vol. 2B, Elsevier-North-Holland: Amsterdam.

Rossi, B. (2013b), "Exchange Rate Predictability". *Journal of Economic Literature* 51(4), 1063-1119.

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

Schwert, G. W. (1987) "Effects of Model Specification on Tests for Unit Roots in Macroeconomic Data, " *Journal of Monetary Economics* 20, 73-103.

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

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

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

About the Authors

Barbara Rossi is an ICREA professor of Economics at Univ. Pompeu Fabra. She is a CEPR Fellow, is a member of the CEPR Business Cycle Dating Committee and a Director of the International Association of Applied Econometrics. Funding from the ERC through Grant 615608 is gratefully acknowledged.

Matthieu Soupre is a PhD student of Economics at Univ. Pompeu Fabra.