



The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

The Stata Journal (2017)
17, Number 4, pp. 850–865

Implementing tests for forecast evaluation in the presence of instabilities

Barbara Rossi
ICREA Professor at University of Pompeu Fabra
Barcelona Graduate School of Economics, and
CREI
Barcelona, Spain
barbara.rossi@upf.edu

Matthieu Soupre
University of Pompeu Fabra
Barcelona, Spain
matthieu.soupre@upf.edu

Abstract. In this article, we review methodologies to fix the size distortions of tests for forecast evaluation in the presence of instabilities. The methodologies implement tests for relative and absolute forecast evaluation that are robust to instabilities. We also introduce the `giacross` and `rosssekh` commands, which implement these procedures in Stata.

Keywords: st0501, `giacross`, `rosssekh`, forecasting, instabilities, structural change

1 Introduction

Researchers often test models' forecasting ability and are often particularly 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 \(2001\)](#). Another typical but different type of forecasting ability test involves evaluating whether forecasts fulfill some minimal requirements, such as being unbiased or producing unpredictable forecast errors using any information available when a forecast is made; such tests are typically referred to as “tests of absolute forecasting performance”. Examples of such tests include [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 and purpose; in fact, the former are used to compare forecasting models, while 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: 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 instabilities in 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 [Stock and Watson](#)

(1996) used include the Quandt (1960) and Andrews (1993) quaslikelihood-ratio test, the mean and exponential Wald test statistics by Andrews and Ploberger (1994), the Ploberger and Krämer (1992) cumulative sum (CUSUM) of squares statistic, and Nyblom's (1989) test. Their analyses uncovered substantial and widespread instabilities in many economic time series. Thus, when researchers test 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, which is violated in the presence of instabilities.

In this article, we present the `giacross` and `rosssekh` commands, which illustrate how to test forecast unbiasedness and rationality as well as how to compare competing models' forecasting performance in a way that is robust to the presence of instabilities. The tests are based on methodologies developed by Giacomini and Rossi (2010) and Rossi and Sekhposyan (2016) and discussed thoroughly by Rossi (2013a). The commands we present implement both Rossi and Sekhposyan's (2016) Fluctuation Rationality Test and Giacomini and Rossi's (2010) Fluctuation Test. The tests are separately presented because they address different concerns. For instance, 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. 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 illustrate when predictive ability appears or breaks down in the data. For each test, we first introduce the test, present the commands that implement it, and then discuss a simple empirical exercise to illustrate the test output and show how to interpret the results.

In section 2, we establish the notation and definitions. Section 3 discusses Rossi and Sekhposyan's (2016) Fluctuation Rationality Test. Section 4 discusses Giacomini and Rossi's (2010) Fluctuation Test. In both sections 3 and 4, we explain the syntax of the 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 (that is, comparing the relative forecasting performance of competing models) or absolute (that is, 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

$t = 1, \dots, P$.¹ Finally, let the forecast error associated with the h -step-ahead forecast made at time t by the first model be denoted by $v_{t,h}$.²

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 performances of competing models over time, where the performance is judged based on a loss function chosen by the forecaster. Let $L(\cdot)$ denote the (general) loss function chosen by the researcher and let $L^{(j)}(\cdot)$ denote the loss corresponding to model j , $j = 1, 2$. The researcher can use 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 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.³ Because the square loss function is the most widely used loss function in practice, we implement it in the 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 :

$$m^{-1} \sum_{j=t-m+1}^t \Delta L_{j,h} \quad t = m, m+1, \dots, P \quad (1)$$

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

$$H_0: E(\Delta L_{t,h}) = 0, \forall t$$

and the alternative can be either $E(\Delta L_{t,h}) \neq 0$ (two-sided alternative) or $E(\Delta L_{t,h}) > 0$ (one-sided alternative).

When one considers the two-sided alternative, their Fluctuation Test Statistic is the largest value over the sequence of the (rescaled) relative forecast error losses defined in (1),

$$\max_t |\mathcal{F}_{t,m}^{\text{OOS}}| \quad (2)$$

where

$$\mathcal{F}_{t,m}^{\text{OOS}} = \hat{\sigma}^{-1} m^{-1/2} \sum_{j=t-m+1}^t \Delta L_{j,h} \quad t = m, m+1, \dots, P \quad (3)$$

-
1. The models' parameters are estimated using either 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.
 2. For example, in a simple linear regression model with h -period lagged ($k \times 1$) vector of regressors \mathbf{x}_t , where $E_t y_{t+h} = \mathbf{x}_t' \boldsymbol{\gamma}$, the forecast at time t is $y_{t,h} = \mathbf{x}_t' \hat{\boldsymbol{\gamma}}_{t,R}$ and the forecast error is $v_{t,h} = y_{t+h} - \mathbf{x}_t' \hat{\boldsymbol{\gamma}}_{t,R}$, where $\hat{\boldsymbol{\gamma}}_{t,R}$ is the estimated vector of coefficients.
 3. In fact, $P^{-1} \sum_{t=1}^P \Delta L_{t,h}$ is exactly the mean squared forecast error.

and $\hat{\sigma}^2$ 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 hypothesis $E(\Delta L_{t,h}) \neq 0$ when $\max_t |\mathcal{F}_{t,m}^{\text{OOS}}| > k_{\alpha,\mu}$, 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 , or formally, $m = \lfloor \mu P \rfloor$.

Similarly, when one considers the one-sided alternative, the Fluctuation Test Statistic is

$$\max_t \mathcal{F}_{t,m}^{\text{OOS}} \quad (4)$$

and the null hypothesis is rejected in favor of the alternative that model 2 forecasts better at some point in time when $\max_t \mathcal{F}_{t,m}^{\text{OOS}}$ is larger than the one-sided critical value.⁴

Note also that $\mathcal{F}_{t,m}^{\text{OOS}}$ is simply a traditional test of equal predictive ability computed over a sequence of rolling out-of-sample windows of size m .

3.2 The `giacross` command

Syntax

The `giacross` command is the equivalent to the MATLAB command written by Giacomini and Rossi (2010). The `dmariano` command (Baum 2003) is required. To install the `dmariano` command, type `ssc install dmariano` in the Command window. The general syntax of the `giacross` command is

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

`realized_value` 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, $t = 1, 2, \dots, P$, where P is the number of forecasts available.

`forecast1` and `forecast2` each contain the forecasts from the competing tested models, 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 (for example, in the case of survey forecasts).

Options

`window(size)` controls the size of the rolling window used for the test, that is, m . `window()` is required.

4. The critical value for the one-sided test differs from that of the two-sided one.

854 *Implementing tests for forecast evaluation in the presence of instabilities*

`alpha(level)` equals the significance level of the test, either 0.05 for a 5% level or 0.10 for 10%. `alpha()` is required.

`nw(bandwidth)` allows the user to choose the truncation lag used in the estimation of the variance $\hat{\sigma}^2$. If no bandwidth is specified, the truncation lag is automatically determined using the [Schwert \(1987\)](#) criterion.

`side(#)` takes the value 1 or 2 and specifies if the null is compared with 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 significantly differ from each other under the alternative.

Stored results

`giacross` stores the following in `r()`:

Scalars

<code>r(tstat_sup)</code>	maximum absolute value of the (rolling) test statistic over the sample, that is, the value of (2)
<code>r(cv)</code>	critical value of the test
<code>r(level)</code>	significance level of the test specified by the user

Macros

<code>r(cmd)</code>	<code>giacross</code>
<code>r(cmdline)</code>	command as typed
<code>r(testtype)</code>	whether the test is one or two sided

Matrices

<code>r(RollStat)</code>	whole temporal sequence of $\mathcal{F}_{t,m}^{\text{OOS}}$, which is also saved as a variable called <code>FlucTest</code> ; note that <code>FlucTest</code> is not the Fluctuation Test Statistic, which is either (2) or (4), depending on whether the test is two sided or one sided
--------------------------	---

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

3.3 Example of practical implementation in Stata

In what follows, we illustrate how to use the `giacross` command to implement the two-sided test. The comma-separated file we use (`giacross_test_data.csv`, provided with the article files) includes quarterly realizations of inflation for the United States starting in 1968Q4 until 2008Q4 as well as the Greenbook (labeled `forc`) and the Survey of Professional Forecast nowcasts (labeled `spf`) for the same variable.

```
. insheet using giacross_test_data.csv, clear
(5 vars, 161 obs)
. generate year = int(pdate)
. generate quarter = (pdate - int(pdate))*4 + 1
. generate tq = yq(year, quarter)
. format tq %tq
. tsset tq
      time variable:  tq, 1968q4 to 2008q4
              delta:  1 quarter
. * lag length set to 3, default 2-sided test
. giacross realiz forc spf, window(60) alpha(0.05) nw(3)
```

Running the Giacomini - Rossi (2010) test for forecast comparison...

REMINDER

First forecast: forc

Second forecast: spf

Actual series: realiz

Newey - West HAC estimator bandwidth: 3

NOTE: the program generates the following variables for plotting:

2 sided alternative: cvlo and cvhi, which contain the lower and upper critical

> values

1 sided alternative: cvone, which contains the one-sided critical value

FlucTest, which contains the sequence of rolling Giacomini - Rossi test

> statistics

(output omitted)

```
. display "The value of the test statistic is " r(tstat_sup)
```

The value of the test statistic is 3.1646998

```
. display "The critical value is " r(cv) " at significance level " r(level)
```

The critical value is 2.89 at significance level .05

```
. graph save GR_demo_1, asis replace
```

(output omitted)

Here is how to interpret the results. The value of the test statistic ($\max_t |\mathcal{F}_{t,m}^{\text{OOS}}|$) is 3.1647, which is larger than the critical value at the 5% significance level, or 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.

The code also returns a graph showing the whole sequence of the Giacomini and Rossi rolling statistic $\mathcal{F}_{t,m}^{\text{OOS}}$, reported in figure 1. The sequence of $\mathcal{F}_{t,m}^{\text{OOS}}$ over time (depicted by a continuous line) 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 the 2000s; this is when the empirical evidence in favor of the first model is the strongest. The figure is saved as `GR_demo_1`.

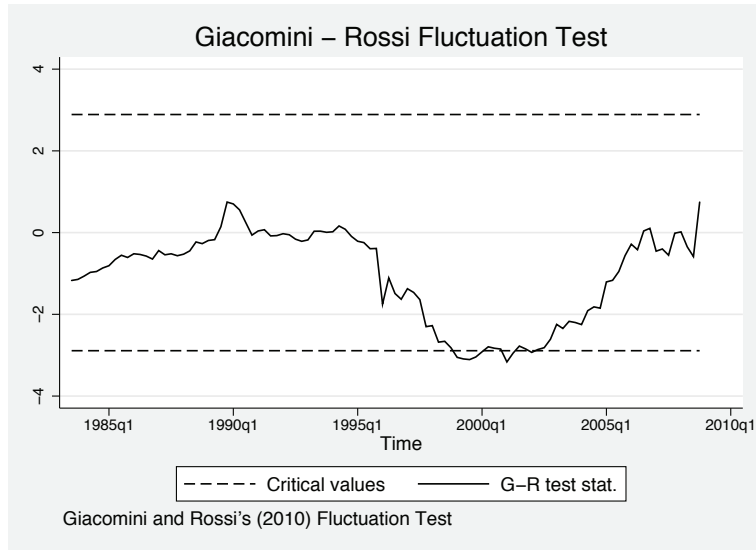


Figure 1. Giacomini and Rossi's test (two sided). The figure depicts $\mathcal{F}_{t,m}^{\text{OOS}}$ from (3) as a function of time (t) for the first example in section 3. The time on the x axis corresponds to the endpoint of the rolling window.

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)

Running the Giacomini - Rossi (2010) test for forecast comparison...

REMINDER
First forecast: forc
Second forecast: spf
Actual series: realiz

Newey - West HAC estimator bandwidth chosen automatically with the Schwert
> criterion.

NOTE: the program generates the following variables for plotting:
2 sided alternative: cvlo and cvhi, which contain the lower and upper critical
> values
1 sided alternative: cvone, which contains the one-sided critical value
FlucTest, which contains the sequence of rolling Giacomini - Rossi test
> statistics
(output omitted)
. display "The value of the test statistic is " r(tstat_sup)
The value of the test statistic is .8266927
```



```
. display "The critical value is " r(cv) " at significance level " r(level)
The critical value is 2.624 at significance level .05
. graph save GR_demo_2, asis replace
(output omitted)
```

Here is how to interpret the results. 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 second model forecasts better than the first model.

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

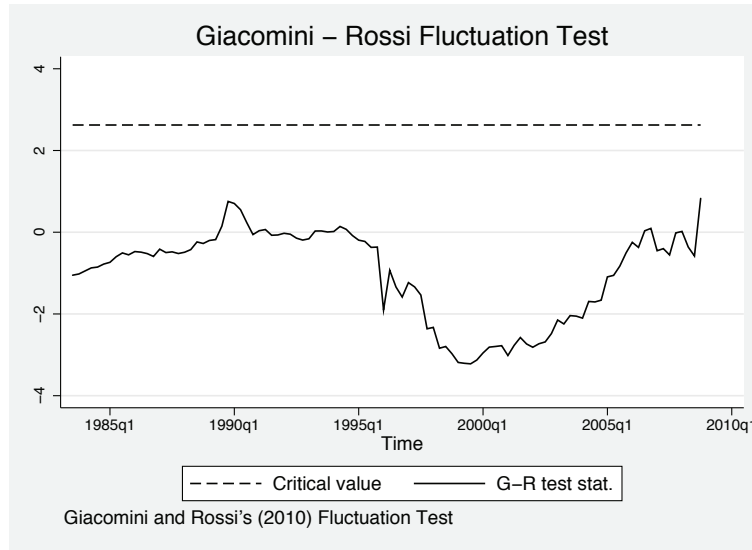


Figure 2. Giacomini and Rossi's test (one sided). The figure depicts $\mathcal{F}_{t,m}^{\text{OOS}}$ from (3) as a function of time (t) for the second example in section 3. The time on the x axis corresponds to the endpoint of the rolling window.

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, Mariano, and West (DMW) test statistic is

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

where $\hat{\sigma}^2$ 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 DMW_P test can be obtained in Stata using the following code:⁵

```
. * Diebold Mariano comparison of forecast accuracy (to compare with GR test)
. dmariano realiz forc spf, max(3)

Diebold-Mariano forecast comparison test for actual : realiz
Competing forecasts: forc versus spf
Criterion: MSE over 161 observations
Maxlag = 3   Kernel : uniform

Series              MSE
-----
forc                1.145
spf                 1.338
Difference          -.1935

By this criterion, forc is the better forecast
H0: Forecast accuracy is equal.
S(1) =    -1.233   p-value = 0.2177
```

Here is how to interpret the results. The command yields a p -value of 0.2177, so the test does not reject the null of equal-forecast accuracy of the two forecasts at the 0.05 significance level. Note that the empirical conclusions are very different from those a researcher would obtain with the Fluctuation Test. In fact, the 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 of the forecast. 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 of the forecast. However, they assume stationarity and are thus invalid in the presence of instabilities.

To make the tests robust to instabilities, [Rossi and Sekhposyan \(2016\)](#) propose estimating the following forecast rationality regressions in rolling windows (of size m),

$$v_{j,h} = \mathbf{g}'_j \boldsymbol{\theta} + \eta_{j,h} \quad j = t - m + 1, \dots, t \quad (5)$$

5. The DMW_P test statistic is also the same as Giacomini and White's (2006).

where the forecast error denoted by $v_{j,h}$ refers to an h -step-ahead out-of-sample forecast made at time j using data available up to that point in time, \mathbf{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), $\boldsymbol{\theta}$ is an $(\ell \times 1)$ parameter vector, and $\eta_{j,h}$ is the residual in the regression. The regression in (5) 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, $\hat{\boldsymbol{\theta}}_t$; by repeating the procedure at times $t = m, m + 1, \dots, P$, 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 because they may average out instabilities. Thus, the hypothesis to be tested is

$$H_0: \boldsymbol{\theta}_t = \boldsymbol{\theta}_0 \text{ versus } H_A: \boldsymbol{\theta}_t \neq \boldsymbol{\theta}_0, \forall t$$

where $\boldsymbol{\theta}_0 = \mathbf{0}$ and $\boldsymbol{\theta}_t$ is the time-varying parameter value.

In (5), we focus on tests of forecast unbiasedness ($\mathbf{g}_t = \mathbf{1}$), forecast efficiency (\mathbf{g}_t is the forecast itself), and forecast rationality (\mathbf{g}_t includes both the forecast and 1).⁶ We refer to tests under the maintained assumption that $\boldsymbol{\theta}_0 = \mathbf{0}$ as "tests for forecast rationality". The zero restriction 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,

$$\max_t \mathcal{W}_{t,m} \quad (6)$$

where

$$\mathcal{W}_{t,m} = m \hat{\boldsymbol{\theta}}_t' \hat{\mathbf{V}}_\theta^{-1} \hat{\boldsymbol{\theta}}_t \quad \text{for } t = m, m + 1, \dots, P$$

is the Wald test in regressions computed at time t over rolling windows of size m and is based on the parameter estimate $\hat{\boldsymbol{\theta}}_t$, which is sequentially estimated in regression (5), and $\hat{\mathbf{V}}_\theta$ is a HAC estimator of the asymptotic variance of the parameter estimate $\hat{\boldsymbol{\theta}}_t$ in the same rolling window.

Here we focus on the version of the Rossi and Sekhposyan (2016) test where parameter estimation error is irrelevant, the forecasts are model free, or the models' parameters are rollingly reestimated 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 .

6. In general, \mathbf{g}_t may also contain any other variable known at time t that was not included in the forecasting model; the framework in (5) also potentially includes tests of forecast encompassing (\mathbf{g}_t is the forecast of the encompassed model) and serial uncorrelation tests (\mathbf{g}_t is the lagged forecast error). See Rossi and Sekhposyan (2016) for details on the implementation in the general case.

4.2 The `rosssekh` command

Syntax

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

```
rosssekh realized_value forecast, window(size) alpha(level) [nw(bandwidth)]
```

realized_value 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, $t = 1, 2, \dots, P$, where P is the number of forecasts available. *forecast* is g_t in our notation.

Options

`window(size)` corresponds to the size of the window in the implementation of the test, that is, m . `window()` is required.

`alpha(level)` equals the significance level of the test, either 0.05 for a 5% level or 0.10 for 10%. `alpha()` is required.

`nw(bandwidth)` allows 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.

Stored results

`rosssekh` stores the following in `r()`:

Scalars

- `r(tstat_sup)` contains the maximum value attained by the (rolling) test statistic $\mathcal{W}_{t,m}$ over the sample, that is, (6)
- `r(cv)` matrix of critical values of the test at the level specified by the user
- `r(level)` level of the test specified by the user

Macros

- `r(cmd)` `rosssekh`
- `r(cmdline)` command as typed

Matrices

- `r(RollStat)` whole time series of the rolling statistic $\mathcal{W}_{t,m}$, which is also saved in the variable `RossSekhTest`; note that the Fluctuation Rationality Test Statistic (6) is the largest value over the sequence
- `r(CV)` contains the critical values of the test

4.3 Example of practical implementation in Stata

In what follows, we illustrate how to use `rosssekh`. `rosssekh_test_data.csv` is the same as in section 3.3. We focus on evaluating forecast rationality of Greenbook forecasts (labeled `forc`).

```

. rosssekh realiz forc, window(60) alpha(0.05) nw(3)

Running the Rossi - Sekhposyan (2016) forecast rationality test...

Critical value for the test: 10.9084
NOTE: the program generates two variables for plotting:
      - cvrs, which contains the critical values
      - RossSekhTest, which contains the sequence of rolling Rossi -
> Sekhposyan test statistics
  (output omitted)
. display "The value of the test statistic is " r(tstat_sup)
The value of the test statistic is 38.899807
. display "The critical value is " r(cv) " at significance level " r(level)
The critical value is 10.9084 at significance level .05
. graph save RS_demo_1, asis replace
  (output omitted)

```

Here is how to interpret the results. The test statistic ($\max_t \mathcal{W}_{t,m}$) reaches a maximum of 38.90, and the critical value is 10.9084. Thus, the test rejects the null hypothesis of forecast rationality.

The code also produces a graph showing Rossi and Sekhposyan's (2016) sequence of test statistics $\mathcal{W}_{t,m}$ over time [defined in (6)], reported in figure 3. The sequence of $\mathcal{W}_{t,m}$ (depicted by a continuous line) is clearly outside the critical value line (depicted by the dashed line). The strongest evidence against forecast rationality appears to be around the beginning of 1995. The figure is saved as RS_demo_1.

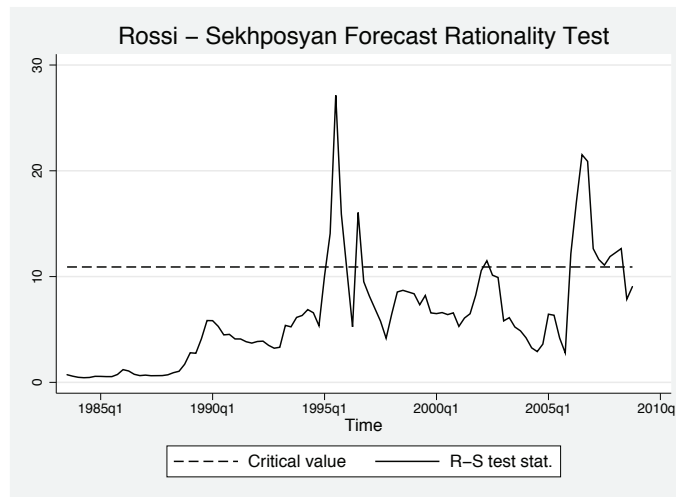


Figure 3. Rossi and Sekhposyan's (2016) test. The figure plots Rossi and Sekhposyan's (2016) sequence of test statistics ($\mathcal{W}_{t,m}$) over time. The time on the x axis corresponds to the endpoint of the rolling window.

862 *Implementing tests for forecast evaluation in the presence of instabilities*

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

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

Running the Rossi - Sekhposyan (2016) forecast rationality test...

Critical value for the test: 10.9084
NOTE: the program generates two variables for plotting:
      - cvrs, which contains the critical values
      - RossSekhTest, which contains the sequence of rolling Rossi -
> Sekhposyan test statistics
(output omitted)

. display "The value of the test statistic is " r(tstat_sup)
The value of the test statistic is 27.139633

. display "The critical value is " r(cv) " at significance level " r(level)
The critical value is 10.9084 at significance level .05

. graph save RS_demo_2, asis replace
(output omitted)
```

Here is how to interpret the results. 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.

4.4 A comparison with traditional tests

A common test to evaluate the forecasting performance of a model is the Mincer and Zarnowitz (1969) test. The Mincer and Zarnowitz (MZ) (1969) test statistic, MZ_P , is a simple F test in the regression $v_{j,h} = \mathbf{g}_j' \boldsymbol{\theta} + \eta_t$, $j = 1, \dots, P$,

$$MZ_P = P \hat{\boldsymbol{\theta}}_P' \hat{\mathbf{V}}_\theta^{-1} \hat{\boldsymbol{\theta}}_P$$

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

The test is designed to test the (unconditional) null hypothesis that $H_0: \boldsymbol{\theta} = \mathbf{0}$ and, under the null, has an asymptotic chi-squared distribution with ℓ degrees of freedom. Again, note that, unlike $\max_t \mathcal{W}_{t,m}$, it is not robust to instabilities.

The MZ_P test can be obtained in Stata from a simple F test as follows:⁷

```
. * Mincer Zarnowitz regression for systematic forecast bias
> (to compare with RS test)
. generate fcsterror=realiz-forc
. newey fcsterror forc, lag(3)
Regression with Newey-West standard errors      Number of obs      =      161
maximum lag: 3                                F( 1,      159) =      0.60
                                              Prob > F          =      0.4386
```

fcsterror	Newey-West		t	P> t	[95% Conf. Interval]	
	Coef.	Std. Err.				
forc	-.0421532	.0542834	-0.78	0.439	-.1493628	.0650564
_cons	.0745427	.1953463	0.38	0.703	-.3112655	.460351

```
. newey fcsterror spf, lag(3)
Regression with Newey-West standard errors      Number of obs      =      161
maximum lag: 3                                F( 1,      159) =      0.03
                                              Prob > F          =      0.8565
```

fcsterror	Newey-West		t	P> t	[95% Conf. Interval]	
	Coef.	Std. Err.				
spf	-.0100134	.0552681	-0.18	0.856	-.1191676	.0991409
_cons	-.0540947	.203149	-0.27	0.790	-.4553132	.3471238

Here is how to interpret the results. The 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, the empirical conclusions differ from those that a researcher would obtain by using the Fluctuation Rationality Test. In fact, the 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. K. 1993. Tests for parameter instability and structural change with unknown change point. *Econometrica* 61: 821–856.
- Andrews, D. W. K., and W. Ploberger. 1994. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* 62: 1383–1414.
- Baum, C. 2003. dmario: Stata module to calculate Diebold–Mariano comparison of forecast accuracy. Statistical Software Components S433001, Department of Economics, Boston College. <http://econpapers.repec.org/software/bocbocode/s433001.htm>.

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

- Clark, T. E., and M. W. McCracken. 2001. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics* 105: 85–110.
- Diebold, F. X., and R. S. Mariano. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13: 253–263.
- Giacomini, R., and B. Rossi. 2010. Forecast comparisons in unstable environments. *Journal of Applied Econometrics* 25: 595–620.
- Giacomini, R., and H. White. 2006. Tests of conditional predictive ability. *Econometrica* 74: 1545–1578.
- Mincer, J., and V. Zarnowitz. 1969. The evaluation of economic forecasts. In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, ed. J. Mincer, 3–46. New York: National Bureau of Economic Research.
- Newey, W. K., and K. D. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55: 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: 1120–1154.
- Nyblom, J. 1989. Testing for the constancy of parameters over time. *Journal of the American Statistical Association* 84: 223–230.
- Ploberger, W., and W. Krämer. 1992. The CUSUM test with OLS residuals. *Econometrica* 60: 271–285.
- 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 instability. In *Handbook of Economic Forecasting*, vol. 2B, ed. G. Elliott and A. Timmermann, 1203–1324. Amsterdam: Elsevier.
- . 2013b. Exchange rate predictability. *Journal of Economic Literature* 51: 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: 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. H., and M. W. Watson. 1996. Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics* 14: 11–30.
- West, K. D. 1996. Asymptotic inference about predictive ability. *Econometrica* 64: 1067–1084.

West, K. D., and M. W. McCracken. 1998. Regression-based tests of predictive ability. *International Economic Review* 39: 817–840.

About the authors

Barbara Rossi is an ICREA professor of economics at the University of Pompeu Fabra. She is a CEPR Fellow, 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 the University of Pompeu Fabra.