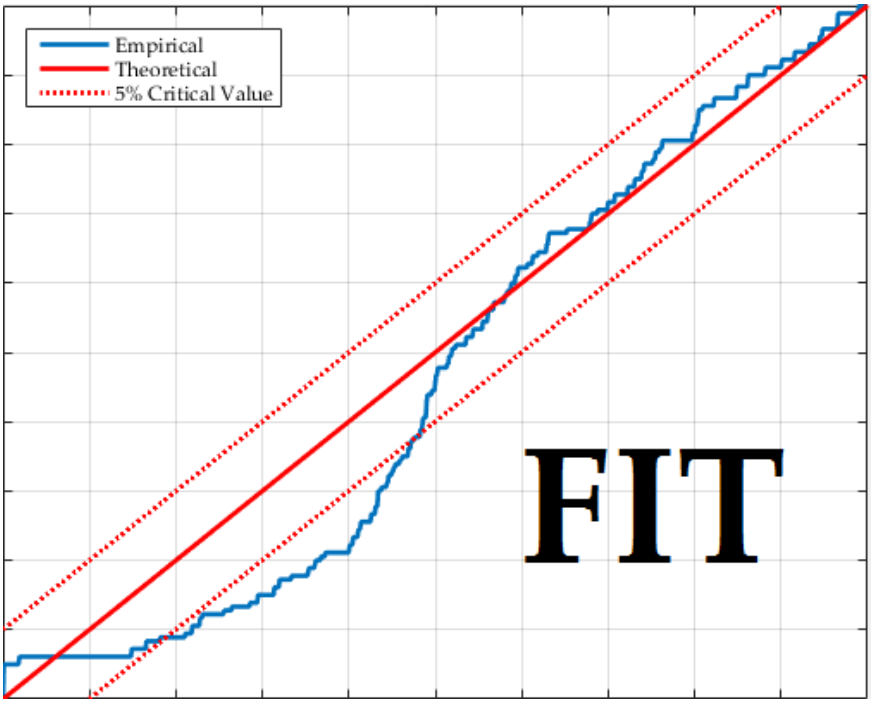
Forecast Evaluation Tests in the Presence of Instabilities*

Technical Guide

Barbara Rossi (barbararossi.work@gmail.com)
Francesca Loria (francescaloria.work@gmail.com)

April 6, 2018



*We thank Tatevik Sekhposyan for comments. The forecast instability tests are also available in the F.I.T. toolbox (see http://francescaloria.wixsite.com/francescaloria/fit-toolbox). This project has received funding from the MINECO grant ECO2012-33247 "The Evolving Transmission of Cyclical Shocks: Methods and Empirical Analyses"and from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 615608).

Forecast Evaluation Tests in BEAR

The **BEAR** toolbox implements tests of forecasting performance for point and density forecasts. The Rossi and Sekhposyan (2016) Fluctuation Rationality test is a test absolute forecasting performance of point forecasts. The Rossi and Sekhposyan (2017) test for correct calibration of forecast densities is a test of correct specification of a model's predictive density.

1 Test on Point Forecasts

It is often of interest to test models' forecasting ability. 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, Rossi (2013b) and Stock and Watson (1996) find severe instabilities in exchange rate forecasting models. Their analysis uncovered substantial and widespread instabilities in many macroeconomic 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.

We illustrate how to test forecast unbiasedness/rationality in a way that is robust to the presence of instabilities. The test is based on the methodology developed by Rossi and Sekhposyan (2016), and discussed thoroughly by Rossi (2013a). The **BEAR** toolbox implements the Rossi and Sekhposyan (2016) Fluctuation Rationality test. The test is also available in Stata (see Rossi and Soupre, 2017). The Rossi and Sekhposyan (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".

1.1 Notations and Definitions

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 $t = 1, \ldots, P$ and realizations are denoted by y_t. 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. Finally, let the forecast error associated with the h-step-ahead forecast made at time t by the first model be denoted by $v_{t,h}$[1].

[1]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.

1.2 The Giacomini and Rossi (2010) Fluctuation Test Test of Relative Forecasting Performance

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 $L(\cdot)$; 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. In BEAR, the Giacomini and Rossi (2010) test is implemented against an AR(p) benchmark, where p is chosen by BIC. Thus, a negative value of $\Delta L_{t,h}$ indicates that the model has a lower forecast error than the autoregressive benchmark. As the square loss function is the most widely used loss function in practice, this is the one we implement in the code. 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:

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

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

$$H_0 : E\left[\Delta L_{t,h}\right] = 0,\ \forall t, \tag{2}$$

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

$$\max_t \mathcal{F}_{t,m}^{OOS}, \tag{3}$$

where

$$\mathcal{F}_{t,m}^{OOS} = \frac{1}{\widehat{\sigma}\sqrt{m}}\sum_{j=t-m+1}^{t}\Delta L_{j,h},\ t=m,m+1,\ldots,P, \tag{4}$$

where $\widehat{\sigma}^2$ is a heteroskedasticity and autocorrelation consistent (HAC) estimator of the long run variance of the loss differences (see 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 $\max_t |\mathcal{F}_{t,m}^{OOS}| > k_{\alpha,\mu}$, 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, 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. Alternatively, one can test whether the forecasts of the second model are significantly better than those of the first model. In this case, since the loss difference $\Delta L_{t,h}$ equals the squared forecast error of the model minus the squared forecast error of the autoregressive benchmark, the null-hypothesis is tested against the one-sided alternative that the benchmark AR(p) forecasts are better than the BEAR model forecasts $\left(\text{i.e., } E\left[\Delta L_{t,h}\left(\widehat{\gamma}_{t,R},\widehat{\beta}_{t,R}\right)\right] > 0\right)$ and rejected when $\max_t \mathcal{F}_{t,m}^{OOS} > k_{\alpha,\mu}$. Thus, for one-sided alternatives, if the test statistic is above the critical value line then it means that the AR(p) forecasts are significantly better.

1.3 The Rossi and Sekhposyan (2016) Fluctuation Rationality Test Test of Absolute Forecasting Performance

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):

$$v_{j,h} = g_j' \cdot \theta + \eta_{j,h}, \ j = t - m + 1, \ldots, t, \ t = m, m + 1, \ldots, P \tag{5}$$

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 $(l \times 1)$ vector function of period j data (which can also possibly be a function of the models' parameter estimates), θ is an $l \times 1$ parameter vector, and $\eta_{j,h}$ is the residual in the regression. The regression in eq. (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, $\widehat{\theta}_t$; by repeating the procedure at times $t = m, m + 1, \ldots, P$, the researcher obtains a sequence of parameter estimates over time.

Rossi and Sekhposyan (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:

$$H_0 : \theta_t = \theta_0 \text{ vs. } H_A : \theta_t \neq \theta_0, \ \forall t, \tag{6}$$

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

The framework in equation (5) 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). All these tests under the maintained assumption that $\theta_0 = 0$ are referred to 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:

$$\max_t \mathcal{W}_{t,m}, \tag{7}$$

where

$$\mathcal{W}_{t,m} = m\widehat{\theta}_t' \widehat{V}_\theta^{-1} \widehat{\theta}_t, \text{ for } t = m, m + 1, \ldots, P, \tag{8}$$

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 (5) 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 $\max_t \mathcal{W}_{t,m} > \kappa_{\alpha,l}$, where $\kappa_{\alpha,l}$ is the critical value at the $100\alpha\%$ significance level with the number of restrictions equal to l.

2 Test on Density Forecasts

Recently, central banks and policy institutions have realized that it is important to complement point forecasts with the uncertainty around these forecasts. Indeed, policy makers are

not only interested in the central tendency of a target variable, such as inflation, but also in the precision of its forecasted value. Density forecasts provide this information through the estimated forecast distribution. Given the important role of density forecasts in summarizing the uncertainty around point forecasts, it is of outmost importance to evaluate whether they are well specified. The test by Rossi and Sekhposyan (2017) allows researchers to test for correct calibration of forecast densities and is a measure of their absolute performance.

The test evaluates whether predictive densities are correctly specified by focusing on evaluating the actual forecasting ability of the model at its estimated parameter values. In this sense, this test is especially appealing for practitioners, as it allows to measure a model's actual forecasting ability in finite samples. This implies that both the parametric model and the estimation technique employed by the researcher are being assessed. The usefulness of this approach is most evident when evaluating density forecasts in the Survey of Professional Forecasters (SPF). SPF panelists use a mixture of (undisclosed) estimated models and expert judgement to produce forecasts. Thus, SPF density forecasts, as well as central banks' fan charts, feature parameter estimation error, which is impossible to correct for. The only feasible approach is to maintain the parameter estimation error under the null hypothesis, as Rossi and Sekhposyan (2017) do.

2.1 Notations and Definitions

Let y_t be variable of interest and $X_t : \Omega \to \mathbb{R}$ a vector of predictors. Let $1 \leq h < \infty$. They are interested in the true, but unknown, h-step-ahead conditional predictive density for the scalar variable y_{t+h} based on $\mathcal{F}_t = \sigma\left(Z_1', \ldots, Z_t'\right)'$, which is the true information set available at time t. They denote this density by $\phi_0\left(\cdot\right)$.

They assume that the researcher has divided the available sample of size $T + h$ into an in-sample portion of size R and an out-of-sample portion of size P and obtained a sequence of h-step-ahead out-of-sample density forecasts of the variable of interest y_t using the information set $\Im_t$, such that $R + P - 1 + h = T + h$ and $\Im_t \subseteq \mathcal{F}_t$. Note that this implies that the researcher observes a subset of the true information set. They also let $\Im_{t-R+1}^t$ denote the truncated information set between time $(t - R + 1)$ and time t used by the researcher.

Let the sequence of P out-of-sample estimates of conditional predictive densities evaluated at the ex-post realizations be denoted by $\left\{\phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^t\right)\right\}_{t=R}^T$. The dependence on the information set is a result of the assumptions imposed on the in-sample parameter estimates, $\widehat{\theta}_{t,R}$. They assume that the parameters are re-estimated at each $t = R, \ldots, T$ over a window of R data indexed from $t - R + 1$ to t (rolling scheme).

Consider the probability integral transform (PIT), which is the cumulative density function (CDF) corresponding to $\phi_{t+h}(\cdot)$ evaluated at the realized value y_{t+h}:

$$z_{t+h} = \int_{-\infty}^{y_{t+h}} \phi_{t+h}\left(y|\Im_{t-R+1}^t\right) dy \equiv \Phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^t\right).$$

Let us also denote the empirical cumulative probability distribution function of the PIT by

$$\varphi_P(r) \equiv \frac{1}{P}\sum_{t=R}^{T} 1\left\{\Phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^t\right) \leq r\right\}.$$

Further, let

$$\xi_{t+h}(r) \equiv \left(1\left\{\Phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^t\right) \leq r\right\} - r\right),$$

where $1\{\cdot\}$ is the indicator function and $r \in [0, 1]$. Consider $\Psi(r) = \Pr\{z_{t+h} \leq r\} - r$ and its (rescaled) out-of-sample counterpart:

$$\Psi_P(r) \equiv \frac{1}{\sqrt{P}}\sum_{t=R}^{T} \xi_{t+h}(r).$$

2.2 Rossi and Sekhposyan (2017) Test for Correct Calibration of Forecast Densities

The Rossi and Sekhposyan (2017) test focuses on testing $\phi_{t+h}\left(y|\Im_{t-R+1}^{t}\right) = \phi_0\left(y|\mathcal{F}^t\right)$, that is, the correct specification of the density forecast of a model estimated with a given window size, R, as well as the parameter estimation method chosen by the researcher. The null hypothesis thus reads

$$H_0 : \phi_{t+h}\left(y|\Im_{t-R+1}^{t}\right) = \phi_0\left(y|\mathcal{F}^t\right) \text{ for all } t = R, \ldots, T,$$

where $\phi_0(y|\mathcal{F}_t) \equiv \Pr\left(y_{t+h} \leq y|\mathcal{F}_t\right)$ denotes the distribution specified under the null hypothesis[2]. The alternative hypothesis, H_A, is the negation of H_0.

2.3 Details on the Critical Values

One-Step Ahead Forecasts The test statistic is

$$\kappa_P = \sup_{r\in[0,1]} |\Psi_P(r)|.$$

The tests rejects H_0 at the $\alpha \cdot 100\%$ significance level if $\kappa_P > \kappa_\alpha$.The **BEAR** toolbox reports the 5% critical value $\kappa_{\alpha=0.05} = 1.34$. Please refer to Table 1 of the original paper for all critical values.

Multi-Step Ahead Forecasts When considering h-step-ahead forecasts, $h > 1$ and finite, the PITs become serially correlated by construction. Thus, the procedure allows the forecasts to be serially correlated under the null hypothesis. In this case, the limiting distribution of the correct specification test is different and a new test statistic has to be constructed. They follow Inoue (2001) in using $\eta_t \sim N\left(1, 1/l\right)$ (where $l = P^{\frac{1}{3}}$ in **BEAR**) when constructing the bootstrap statistics. Let ω be a particular bootstrap sample. Let $z_{t+h}(\omega)$ be a realization of the PIT in a particular bootstrap sample. Note that the bootstrap requires resampling the PITs, z_{t+h}, not the original data Z_t. Define the bootstrap test statistics as

$$\Psi_P^*(r;\omega) = \frac{1}{\sqrt{P}} \sum_{j=R}^{T-l+1} \eta_j \sum_{i=j}^{j+l-1} \left(1\left\{z_{i+h}(\omega) \leq r\right\} - r\right)$$

and

$$\kappa_P^*(\omega) = \sup_{r\in[0,1]} |\Psi_P^*(r;\omega)|$$

In this case, the tests rejects H_0 at the $\alpha \cdot 100\%$ significance level if $\kappa_P^* > \kappa_\alpha^*$.

The bootstrap is implemented using the following step-by-step procedure:

(i) Construct the test statistics as for the one-step ahead case;

(ii) Let S be the maximum number of bootstrap replications. For $s = 1, 2, \ldots, S$, generate $\{\kappa_{P;s}^*\}_{s=1}^{S}$, where $\kappa_{P;s}^*$ is based on $\left\{\eta_t^{(s)}\right\}_{t=R}^{T-l+1}$;

(iii) Estimate the level-α critical value $\widehat{c}_{\kappa,\alpha}^{S}$ from $\{\kappa_{P;s}^*\}_{s=1}^{S}$.

[2]Notice that the test evaluates the model's performance for a given estimation sample size R. It might thus be possible that correct specification is rejected for a model for some values of R and not rejected for the same model for some other choices of R.

Bibliography

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

Inoue, A. (2001). Testing For Distributional Change In Time Series. *Econometric Theory 17*(01), 156–187.

Mincer, J. and V. Zarnowitz (1969). The Evaluation of Economic Forecasts. pp. 3–46.

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

Rossi, B. (2013a). *Advances in Forecasting under Instability*. Elsevier.

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).

Rossi, B. and T. Sekhposyan (2017). Alternative Tests for Correct Specification of Conditional Predictive Densities.

Rossi, B. and M. Soupre (2017). Implementing Tests For Forecast Evaluation in the Presence of Instabilities. *Stata Journal 17*(4), 1–16.

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

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