Improved Tests for Forecast Comparisons in the Presence of Instabilities

Luis Filipe Martins[*] **Pierre Perron**[†]

Lisbon University Institute **Boston University**

October 6, 2015

Abstract

Of interest is comparing the out-of-sample forecasting performance of two competing models in the presence of possible instabilities. To that effect, we suggest using simple structural change tests, sup-Wald and $UDmax$ as proposed by Andrews (1993) and Bai and Perron (1998), for changes in the mean of the loss-differences. Giacomini and Rossi (2010) proposed a fluctuations test and a one-time reversal test also applied to the loss-differences. When properly constructed to account for potential serial correlation under the null hypothesis to have a pivotal limit distribution, it is shown that their tests have undesirable power properties, power that can be low and non-increasing as the alternative gets further from the null hypothesis. The good power properties they reported is simply an artifact of imposing a priori that the loss differentials are serially uncorrelated and using the simple sample variance to scale the tests. On the contrary, our statistics are shown to have higher monotonic power, especially the UDmax version. We use their empirical examples to show the practical relevance of the issues raised.

Keywords: non-monotonic power, structural change, forecasts, long-run variance.
JEL Classification: C22, C53

[*] Department of Quantitative Methods, ISCTE - Lisbon University Institute, Business School, Av. das Forças Armadas, 1649-026 Lisboa, Portugal (luis.martins@iscte.pt)

[†] Department of Economics, Boston University, 270 Bay State Rd., Boston, MA, 02215 (perron@bu.edu).

1 Introduction

Testing for the relative forecasting performance of two, or more, competing models has been the subject of substantial research. Important contributions include Diebold and Mariano (1995), West (1996), Clark and West (2006) and Giacomini and White (2006). These are based on assessing whether the out-of-sample loss differentials are significantly different from zero. They differ with respect to the exact specification of the null hypothesis (loss functions evaluated at the population values of the parameters or the in-sample estimates), having nested or non-nested models, using an unconditional perspective or one that conditions on some covariates. Being based on averages of the loss differentials, these tests may have little power when the relative forecasting performance is changing over time.

Of interest is comparing the out-of-sample forecasting performance of two competing models in the presence of possible instabilities. To that effect, we suggest using simple structural change tests, sup-Wald and $UDmax$ as proposed by Andrews (1993) and Bai and Perron (1998), for changes in the mean of the loss-differences. The tests effectively look at the entire time path of the models' relative performance, which may contain useful information not available when using tests that focus on the average relative performance.

Giacomini and Rossi (2010), henceforth GR, proposed a fluctuations test and a one-time reversal test also applied to the loss-differences. When properly constructed to account for potential serial correlation under the null hypothesis to have a pivotal limit distribution, it is shown that the tests proposed by GR have undesirable power properties, power that can be low and non-increasing as the alternative gets further from the null hypothesis. In the terminology of Perron (2006), these tests belong to the so-called "partial sums" type tests. These have repeatedly been shown to be inadequate for structural change problems. The good power properties reported in GR is simply an artifact of imposing a priori that the loss differentials are serially uncorrelated and using the simple sample variance to scale the tests.

We replicate the power properties of their tests with the appropriate Heteroskedasticity and Autocorrelation (HAC) correction using exactly the same design they used. In the case of a one-time change in the relative forecasting performance of two models, the power functions of the tests are substantially lower than what they report. More importantly, the power functions are non-monotonic. The power does not tend to one as the magnitude of the difference between the models' relative forecasting performance increases and may even decline. These are clearly undesirable features of test statistics, which makes their usage in practice unreliable. On the contrary, the test statistics we now propose are shown to have higher monotonic power, especially the $UDmax$ version.

We also revisit their empirical results related to assess the forecasting performance of the UIRP (Uncovered Interest Rate Parity) model relative to a simple random walk model for the UK pound and German Deutsche Mark exchange rate relative to the US$. We show that their tests have little power to discriminate between the models they considered, while the sup-Wald and $UDmax$ provide a strong rejection in the case of the UK Pound. However, there is no evidence that the UIRP model performed significantly better than a simple random walk model in any part of the sample. This illustrates the practical relevance of the power problems of the tests proposed by GR and the fact that the sup-Wald and $UDmax$ tests for changes in the mean of the loss-differences yield more powerful procedures.

This note is structured as follows. Section 2 reviews the framework considered by GR, our suggested tests and those proposed by GR. Section 3 reevaluates the power functions of the tests when a HAC correction is applied. Section 4 does the same for the empirical applications. Section 5 provides brief concluding remarks.

2 The framework and the tests

The interest is in comparing h-step-ahead forecasts from two competing models characterized by parameters θ and γ, respectively. There is a sample size of T observations available, which is divided into an in-sample portion of size R and an out-of-sample portion of size P. The two models yield two competing sequences of h-step-ahead out-of-sample forecasts and, for a given loss function L, these yield a sequence of P out-of-sample forecast loss differences $\{\triangle L_t(\hat{\theta}_{t-h,R}, \hat{\gamma}_{t-h,R}) = \{L^{(1)}(y_t, \hat{\theta}_{t-h,R}) - L^{(2)}\left(y_t, \hat{\gamma}_{t-h,R}\right)\}_{t=R+h}^{T}$, where $\hat{\theta}$ and $\hat{\gamma}$ are the in-sample parameter estimates. A rolling scheme method of estimation is used whereby the parameters are re-estimated at each $t = R+h, ..., T$ over a window of length R including data indexed $t-h-R+1, ..., t-h$. The local relative loss for the two models is the sequence of out-of-sample loss differences computed over centered rolling windows of size m given by (for m even): $m^{-1}\sum_{j=t-m/2}^{t+m/2-1} \triangle L_j(\hat{\theta}_{j-h,R}, \hat{\gamma}_{j-h,R})$ for $t = R+h+m/2, ..., T-m/2+1$. The simulations and applications are restricted to the case with a quadratic loss function $L_t = (y_t - f_t)^2$, where f_t is the forecast and to the case of a one-step-ahead forecast.

The null hypothesis is constant forecast accuracy:

$$H_0 : E[\triangle L_t(\hat{\theta}_{t-h,R}, \hat{\gamma}_{t-h,R})] = c \text{ for all } t = R+h, ..., T, \tag{1}$$

for some c versus the alternative hypothesis of changing relative forecast accuracy. The tests considered are 1) the simple sup-Wald test for a single change (e.g., Andrews, 1993, denoted $\sup W$) and 2) the $UD\max$ test of Bai and Perron (1998) which allows up to 5 breaks. These are applied to test for changes in the mean of the loss-differences sequence. Let SSR be the sum of the squared demeaned loss differences over the full sample and $SSR(i, j)$ be the sum of the squared demeaned loss differences over a sample involving the observation i to j. The $\sup W$ and $UD\max$ tests take the form

$$\sup W = \sup_{t\in\Lambda_1} W_1(t)$$

where $W_1(t) = \hat{\sigma}_W^{-2}\left(SSR - SSR(1,t) - SSR(t+1,P)\right)$, $\Lambda_1 = \{[\epsilon P],...[(1-\epsilon)P]\}$ and we use $\epsilon = 0.15$, $\hat{\sigma}_W^2$ is the HAC estimator of the demeaned loss differences under the alternative hypothesis. The $UD\max$ test which allows for up to 5 breaks is

$$UD\max = \max_{1\leq b\leq 5} \sup_{(t_1,...,t_b)\in\Lambda_b} W_b(t_1,...,t_b),$$

where $W_b(t_1,...,t_b) = b^{-1}\hat{\sigma}_b^{-2}(SSR - \sum_{i=1}^{b+1} SSR(t_{i-1}+1,t_i))$, where we use the convention that $t_0 = 0$ and $t_{b+1} = P$. Also,

$$\Lambda_b = \{(t_1,...,t_b) : |t_{k+1} - t_k| \geq [\epsilon P], k = 1,...,b-1, t_1 \geq [\epsilon P], t_b \leq [(1-\epsilon)P]\}$$

and $\hat{\sigma}_b^2$ is the HAC estimator of the demeaned loss differences under the alternative hypothesis. See Bai and Perron (1998) for further details. It is straightforward to show that the tests have the same limit distributions as in Andrews (1993) and Bai and Perron (1998) under the same assumptions used in GR. As we shall show, these tests have much higher power and, in particular, the $UD\max$ version always has a monotonically increasing power function.

Our tests will not have power against alternatives with unequal but constant forecast accuracy (since we do not set $c = 0$ under the null hypothesis) but in such cases the original test of Giacomini and White (2006) or that of Clark and West (2006) will have higher power than the tests proposed by GR. The way to use the tests together is as follows. First use the sup-Wald or $UDmax$ that we propose. If there is a rejection, conclude that there is a change in forecast accuracy between the models. If there is no rejection, apply the statistic of Giacomini and White (2006) or that of Clark and West (2006) to test if there is non equal but constant relative forecasting performance. When 5% size tests are used, under the null of equal forecast accuracy this strategy will have a nominal size slightly less than 5% (.95×.05). So there is no size problem related to the use of multiple tests. Second, the power of the sup-Wald or $UDmax$ will be the same as reported since it is used first. The power of the Giacomini and White (2006) or that of Clark and West (2006) will also nearly be the same as when used individually for the alternative hypothesis it is intended to detect, though in 5% of the cases a constant non-equal relative forecasting performance will be classified as a time-varying one.

The null hypothesis adopted by GR is that of equal forecast accuracy

$$H_0 : E[\triangle L_t(\hat{\theta}_{t-h,R}, \hat{\gamma}_{t-h,R})] = 0 \text{ for all } t = R+h,...,T,$$

versus the alternative hypothesis that one model provides better forecasts, i.e.,

$$H_1 : E[\triangle L_t(\hat{\theta}_{t-h,R}, \hat{\gamma}_{t-h,R})] \neq 0.$$

Tests for this null hypothesis were provided by Diebold and Mariano (1995) and the unconditional version of the statistics proposed by Giacomini and White (2006). The first test proposed by GR is the out-of-sample fluctuations test defined by $\max_t |F_{t,m}^{OOS}|$ where

$$F_{t,m}^{OOS} = \hat{\sigma}^{-1} m^{-1/2} \sum_{j=t-m/2}^{t+m/2-1} \triangle L_j(\hat{\theta}_{j-h,R}, \hat{\gamma}_{j-h,R}) \tag{2}$$

for $t = R + h + m/2, ..., T - m/2 + 1$, with $\hat{\sigma}^2$ a HAC estimate of the long-run variance $\sigma^2 = \lim_{P\to\infty} E(P^{-1/2} \sum_{t=R+h}^{T} \triangle L_t(\hat{\theta}_{t-h,R}, \hat{\gamma}_{t-h,R}))^2$. They suggest the use of a kernel-based method using the Bartlett window, i.e.,

$$\hat{\sigma}^2 = \sum_{i=-q(P)+1}^{q(P)-1} (1 - |i/q(P)|) P^{-1} \sum_{j=R+h}^{T} \triangle L_j^*(\hat{\theta}_{j-h,R}, \hat{\gamma}_{j-h,R}) \triangle L_{j-i}^*(\hat{\theta}_{j-i-h,R}, \hat{\gamma}_{j-i-h,R}) \tag{3}$$

where $q(P)$ is a bandwidth that grows with P. GR make no recommendation about how to select $q(P)$. Following state-of-the art good practice, in the simulations and applications we use a data-dependent method, specifically the one advocated by Andrews' (1991) based on an AR(1) approximation. Also, correcting for an omission in GR, the demeaned loss functions are

$$\triangle L_j^*(\hat{\theta}_{j-h,R}, \hat{\gamma}_{j-h,R}) = \triangle L_j(\hat{\theta}_{j-h,R}, \hat{\gamma}_{j-h,R}) - P^{-1} \sum_{j=R+h}^{T} \triangle L_j(\hat{\theta}_{j-h,R}, \hat{\gamma}_{j-h,R}).$$

This statistic is referred to as the GW-fluctuations test since it is based on the maximum (over some range) of the sequence of tests $F_{t,m}^{OOS}$, which are equivalent to the test of Diebold and Mariano (1996) and the unconditional version of the Giacomini and White (2006) test. The second test they propose is the one-time reversal (OTR) test defined by $QLR_P^* = \sup_t \Phi_P^*(t)$, $t \in \{[0.15P], ..., [0.85P]\}$, with $\Phi_P^*(t) = LM_1 + LM_2(t)$ where

$$LM_1 = \hat{\sigma}^{-2} P^{-1} [\sum_{j=R+h}^{T} \triangle L_j(\hat{\theta}_{j-h,R}, \hat{\gamma}_{j-h,R})]^2$$

$$LM_2(t) = \hat{\sigma}^{-2} \frac{1}{P} (\frac{t}{P})^{-1} (1 - \frac{t}{P})^{-1} [\sum_{j=R+h}^{t} \triangle L_j(\hat{\theta}_{j-h,R}, \hat{\gamma}_{j-h,R}) - (\frac{t}{P}) \sum_{j=R+h}^{T} \triangle L_j(\hat{\theta}_{j-h,R}, \hat{\gamma}_{j-h,R})]^2$$

and $\hat{\sigma}^2$ is again defined by (3).

For all tests, the framework can be adapted to a different null hypothesis in which the concern is about the forecast losses evaluated at the population parameters as considered in Clark and West (2006). In this case, one simply apply an adjustment to the forecast losses. For example, when one model specifies y_t to be a martingale difference sequence and the other is a linear regression model of the form $y_t = \beta X_{t+1} + e_t$, the adjusted mean-squared loss-differences are

$$\Delta L_t = y_t^2 - [(y_t - f_t)^2 - f_t^2]$$

where f_t is the forecast from the regression model. GR refer to the fluctuations test applied to such corrected loss functions as the CW-fluctuations test.

For both tests, the use of a HAC estimator for the long-run variance is essential. To illustrate, we generated loss differences as an AR(1) process with coefficient 0.75. Such type of serial correlation can arise as the result of serial correlation in the second order moments of the residuals and/or the regressors, including, but not restricted to, GARCH processes. In this case, the size of all tests with a fixed number of lags $q\,(P) = 2$ is near 70%. When using Andrews's data dependent method to select $q\,(P)$, the size of the GR tests (OTR and Fluctuations) is between 5 and 10%. Hence, it is important to appropriately correct for potential serial correlation in the loss differences. Also, if instabilities are present under the alternative hypothesis, a situation that indeed motivates the tests proposed, the loss differentials will exhibit features akin to serial correlation in the sense that a test for serial correlation would tend to reject the absence of correlation. This is simply a consequence of the results in Perron (1989, 1990) that a change in the mean (or slope) of a time series biases the sum of the autoregressive coefficients upwards when such changes are not explicitly modeled. Yet, GR impose a priori that the loss differentials are serially uncorrelated and use the simple sample variance as the estimate of σ^2, namely $\hat{\sigma}^2 = P^{-1} \sum_{j=R+h}^{T} \triangle L_j^*(\hat{\theta}_{j-h,R}, \hat{\gamma}_{j-h,R})^2$. They do so for both the simulations reported and the applications. As we document in the next sections, the properties of their tests are very different when the test is properly constructed with a HAC estimate and the conclusions of their empirical applications are also different.

3 The simulations

We adopt the same simulation setup as in GR in order to avoid any potential biases due to the selection of particular DGPs. We also used their code available at the Journal of Applied Econometrics website in order to correct some inaccurate reporting or typos in their paper. The results obtained and the documented power reversal of the tests could be much more severe using other DGPs. Two forecasting models are considered. For the first, there is a covariate X_t that potentially helps to forecast Y_t so that $f_{t,R}^{(1)} = \hat{\beta}_{t,R} X_{t+1}$ (assuming that X_{t+1} is known when constructing the forecast) where $\hat{\beta}_{t,R}$ is the in-sample parameter estimate from a regression of Y_t on X_t based on a rolling window of size R. For the second model, Y_t is assumed to be a zero-mean white noise process so that $f_{t,R}^{(2)} = 0$. Hence, under the GW framework the loss differentials are

$$\triangle L_{R,t+1} = Y_{t+1}^2 - (Y_{t+1} - \hat{\beta}_{t,R} X_{t+1})^2,$$

while under the CW framework, they are

$$\triangle L_{R,t+1} = Y_{t+1}^2 - [(Y_{t+1} - \hat{\beta}_{t,R} X_{t+1})^2 - (\hat{\beta}_{t,R} X_{t+1})^2].$$

We consider simulations pertaining to assess the performance of the tests when the forecasting performance of the models is time varying such that there is a one-time break in the relative performance during the out-of-sample period induced by a break in the DGP. Under the GW framework this is achieved by setting (with a proper correction for an error in GR)

$$Y_t = (-\delta + 1/\sqrt{R}) X_t I\,(t \leq R + \tau P) + (\delta + 1/\sqrt{R}) X_t I\,(t > R + \tau P) + \varepsilon_t,$$

where $X_t = 0.5 X_{t-1} + v_t$ with $v_t \sim i.i.d.\ N(0,1)$ and $\varepsilon_t \sim i.i.d.\ N(0,1)$ uncorrelated with v_t. Hence, the relative performance changes at $t = R + \tau P$. We use the parameters $\tau = 1/3$ or $\tau = 2/3$ and $\mu = m/P = 0.3,\ 0.7$. The results with a HAC correction are presented in Figures 1 ($\tau = 1/3$) and 2 ($\tau = 2/3$). The left panel considers the same values of δ as in GR (0 to 1), while the right panel shows the power functions for values of δ up to 10. In all cases, we consider 5% two-sided tests. Consider first the case with $\tau = 1/3$. When $\mu = 0.3$, the GW fluctuations test has more power than the OTR test, as in GR, but the power is much lower than they reported. More importantly, both tests suffer from non-monotonic power, none have power 100% no matter how large δ is. The power of the fluctuations test reaches a maximum value of about 0.90 when δ is near 1, while the OTR test reaches a maximum power of about 0.6 when δ gets large. The $\sup W$ does not have monotonic power either with a power function in between that of the GW fluctuations and the OTR test. The $UD\max$, on the other hand, has monotonic power that approaches 1 quickly and is the most powerful overall. When $\mu = 0.7$, the OTR test has more power than the fluctuations test, as in GR. But here with the HAC correction, the power decrease is even more pronounced. The power of the fluctuations test reaches a maximum value of about 0.37, while the OTR test reaches a maximum power of about 0.55. The $\sup W$ does not have monotonic power either but its power function is now higher than the GW fluctuations and the OTR tests. The $UD\max$ test again has monotonic power that approaches 1 quickly and is the most powerful overall. Consider now the case with $\tau = 2/3$ presented in Figure 2. For both $\mu = 0.3$ or

0.7, the $\sup W$ has highest power followed closely by the $UD\max$, both having monotonic power functions. As in GR, the OTR test has high power whether $\mu = 0.3$ or $\mu = 0.7$. In all cases, the OTR and GW fluctuations tests suffer from non-monotonic power which does not reaches 1 even for very large values of δ. When $\mu = 0.3$, the power of the OTR test achieves a maximum near but below one when δ is near 0.8 and the power remains the same as δ increases. The power of the GW fluctuations test reaches a maximum near 0.85 when δ is near 1 but it decreases to about 0.70 as δ increases further. When $\mu = 0.7$, the power function of the OTR test is similar but that of the GW fluctuations test is considerably reduced when a HAC correction is applied reaching a maximal value near 0.15.

Under the CW framework the model used is:

$$Y_t = -\delta X_t I\left(t \leq R + \tau P\right) + \delta X_t I\left(t > R + \tau P\right) + \varepsilon_t.$$

We again set $\tau = 1/3$ or $\tau = 2/3$ and $\mu = m/P = 0.3$, 0.7, but also present results for the case $\tau = 1/2$ as GR report results for this case only. The results with a HAC correction are presented in Figures S1 ($\tau = 1/3$), S2 ($\tau = 1/2$) and S3 ($\tau = 2/3$) in a separate appendix to this paper, Martins and Perron (2015). The first thing to note is that in all cases, the $\sup W$ and $UD\max$ tests have nearly identical monotonic power functions that approach one quickly. On the other hand, the power of the CW fluctuations test never increases to one no matter how large the change is. The maximal power achieved depends highly on the exact specifications. When $\mu = 0.3$, it is between .85 and .90 for the three values of τ considered. However, when $\mu = 0.7$, it is near one when $\tau = 2/3$ but not above .25 when $\tau = 1/2$ and essentially zero when $\tau = 1/3$.

In summary, the simulations show important problems of non-monotonic power for the GW or CW fluctuations and the OTR tests. The $UD\max$ test always has power functions approaching one quickly. In most cases, the power of the $\sup W$ is comparable to that of the $UD\max$ though it can also be subject to power functions flattening below one as the alternative gets large. Hence, in the presence of unequal time-varying forecast accuracy, the $UD\max$ test for changes in the mean of the loss-differences is clearly the preferred test.

A comment about the bandwidth selection is in order. It is well known that the reason for the non-monotonic power is the fact that a relatively large bandwidth is selected via Andrews' method under the alternative (see, e.g., Kim and Perron, 2009). It may be argued that with large breaks, the bandwidth selected is "too high". This is not the case. The average (over all replications) value of the bandwidth $q\left(P\right)$ selected by Andrews' method ranges between 4 and 6 when δ varies between .5 and 1 for which the power reversal is present. These values are near the default value of Stata, say, which is 5 when T=100 (Newey-West option).

While a data-dependent method is highly preferable over a fixed rule for the selection of $q\,(P)$ to ensure the proper size (asymptotically and in finite samples), one may have a strong prior that the loss differentials are weakly correlated under the null hypothesis and therefore use a fixed rule to select $q\,(P)$. Figures S4.a and S4.b in Martins and Perron (2015) present the results corresponding to Figures 1 and 2 when setting the popular rule of thumb of $q\,(P) = 5$ to construct the statistics. The results show that some of the power functions of the tests are no longer non-monotonic but that overall the sup-Wald and, especially, the UDmax tests have higher power, sometimes by a high margin. Hence, the superiority of the proposed tests holds under both a fixed or data dependent rule to select $q\,(P)$. Of course, the power problems are less with a fixed value $q\,(P) = 3$ but they are also much worse with a fixed value $q\,(P) = 9$. This is trivial since if $q\,(P)$ is very small, the estimate becomes similar to using the standard sample variance.

It has by now become standard (good) practice to use a data-dependent method to select the bandwidth. It has the advantage of providing a selection method that is not ad hoc or arbitrary and that, in general, delivers tests with good finite sample size for a wide range of possible DGPs. As stated earlier, using a low fixed value would invariably lead to tests with size distortions for a wide variety of DGPs.

4 The applications revisited

GR applied the tests they proposed to assess the forecasting performance of the UIRP (Uncovered Interest Rate Parity) model relative to a simple random walk model for the UK pound and German Deutsche Mark exchange rate relative to the US$. Large positive values of the fluctuations test provide evidence that the UIRP model is superior to the random walk model. Again, the tests were constructed without a HAC correction assuming a priori uncorrelated forecast losses. They also departed way from the fluctuations test they proposed. Instead of (2), they reported results for the following version of the test

$$F_{t,m}^{OOS} = m^{-1/2} \sum_{j=j-m/2}^{t+m/2-1} \triangle L_j(\hat{\theta}_{j-h,R}, \hat{\gamma}_{j-h,R})/\hat{\sigma}_t$$

where $\hat{\sigma}_t^2 = m^{-1/2} \sum_{j=t-m/2}^{t+m/2-1} \triangle L_j^*(\hat{\theta}_{j-h,R}, \hat{\gamma}_{j-h,R})^2$. In what follows, we consider the original statistic defined with the long-run variance estimated using the full sample. We consider two-sided 5% tests.

Consider first the results for the German Deutsche Mark presented in Figure 3. Here, none of the tests are significant, including the OTR and $UD\max$ not reported. This contrasts with the results of GR who reported a significant rejection using the CW-fluctuations test without a HAC correction.

Consider now the results for the UK pound presented in Figure 4. Here also the OTR is not significant, as well as the $\sup W$ and $UD\max$ based on the GW loss-differences. On the other hand, the fuctuations-based tests offer a contrasting picture. The GW-fluctuations test is barely significant but in favor of the random walk model, contrary to what was reported in GR. On the other hand, the CW-fluctuations test is barely significant in favor of the UIRP, consistent with the result in GR. Based on the CW loss-differences, the $\sup W$ and $UD\max$ are both very highly significant at less than the 1% significance level, which illustrates the higher power of these tests. The estimate of the break date (that which maximizes the sequence of Wald tests for a single change) is 1990:09. To assess the nature of the change in forecasting performance, we estimated the mean of the loss-differences pre and post-1990:09. These are 0.0002 and -0.00004. Hence, this points to better forecasting performance for the UIRP pre-1990:09 and vice-versa post 1990:09. However, a standard CW test applied to the pre 1990:09 sample yields a t-statistic of 0.33. Hence, there is no evidence that the UIRP performed significantly better than the RW in any part of the sample.

5 Conclusions

When constructed properly, it is shown that the tests proposed by GR have undesirable power properties, power that can be low and non-increasing as the alternative gets further from the null hypothesis. In the terminology of Perron (2006), these tests belong to the so-called "partial sums" type tests. These have repeatedly been shown to be inadequate for structural change problems. Tests based on standard Wald statistics are much less prone to such problems. This is again the case here. We have shown that to detect changing relative forecasting accuracy the $\sup W$, and in particular, the $UD\max$, tests applied to test for changes in the mean of the loss-differences have much higher power. Of course, these are not appropriate to test for unequal but constant relative forecast accuracy. In such cases, the original tests of Giacomini and White (2006) and Clark and West (2006) are to be used. The fluctuations versions of these tests, and the OTR test offer no power gains in this case.

References

Andrews, DWK. 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**: 817-858.

Andrews DWK. 1993. Tests for parameter instability and structural change with unknown change point. *Econometrica* **61**: 821-856.

Bai J, Perron P. 1998. Estimating and testing linear models with multiple structural changes. *Econometrica* **66**: 47-78.

Clark, T, West KD. 2006. Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics* **105**: 85-110.

Diebold, FX, Mariano, RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**: 253-263.

Giacomini, R, Rossi B. 2010. Forecast comparisons in unstable environments. *Journal of Applied Econometrics* **25**: 595-620.

Giacomini, R, White H. 2006. Tests of conditional predictive ability. *Econometrica* **74**: 1545-1578.

Kim, D, Perron P. 2009. Assessing the relative power of structural break tests using a framework based on the approximate bahadur slope. *Journal of Econometrics* **149**: 26-51.

Martins, LF, Perron, P. 2015. Online Appendix to "Improved Tests for Forecast Comparisons in the Presence of Instabilities". *Journal of Time Series Analysis.*

Perron, P. 1989. The great crash, the oil price shock and the unit root hypothesis. *Econometrica* **57**: 1361-1401.

Perron, P. 1990. Testing for a unit root in a time series regression with a changing mean. *Journal of Business and Economic Statistics* **8**: 153-162.

Perron P. 2006. Dealing with structural breaks. In *Palgrave Handbook of Econometrics*, Vol. 1: Econometric Theory, K. Patterson and T.C. Mills (eds.), Palgrave Macmillan, 278-352.

West, KD. 1996. Asymptotic inference about predictive ability. *Econometrica* **64**: 1067-1084.

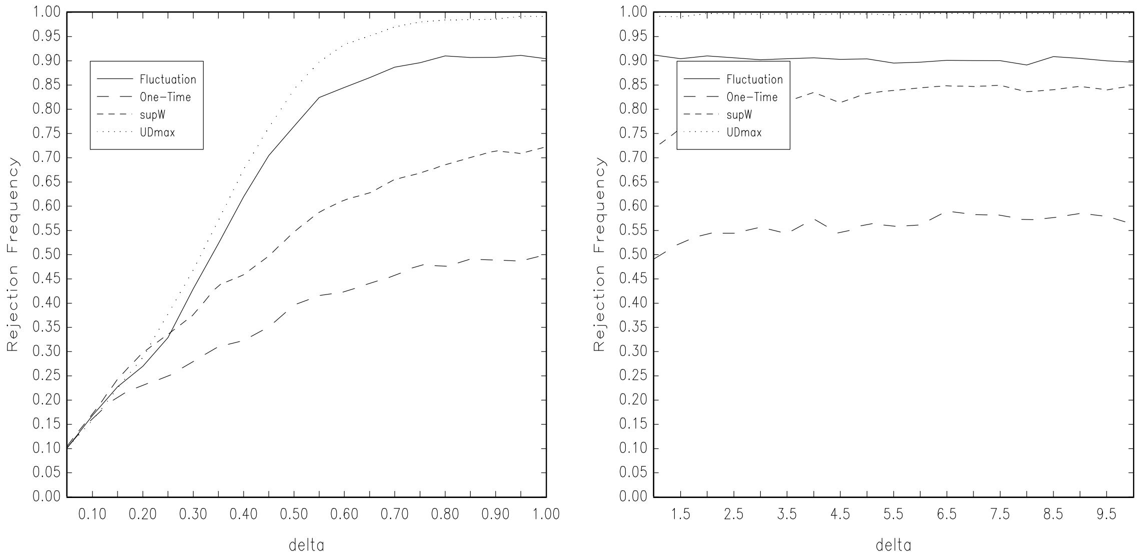

Figure 1.a: Power functions of the GW tests with a break in the relative performance at $\tau = 1/3$, $\mu = 0.3$.

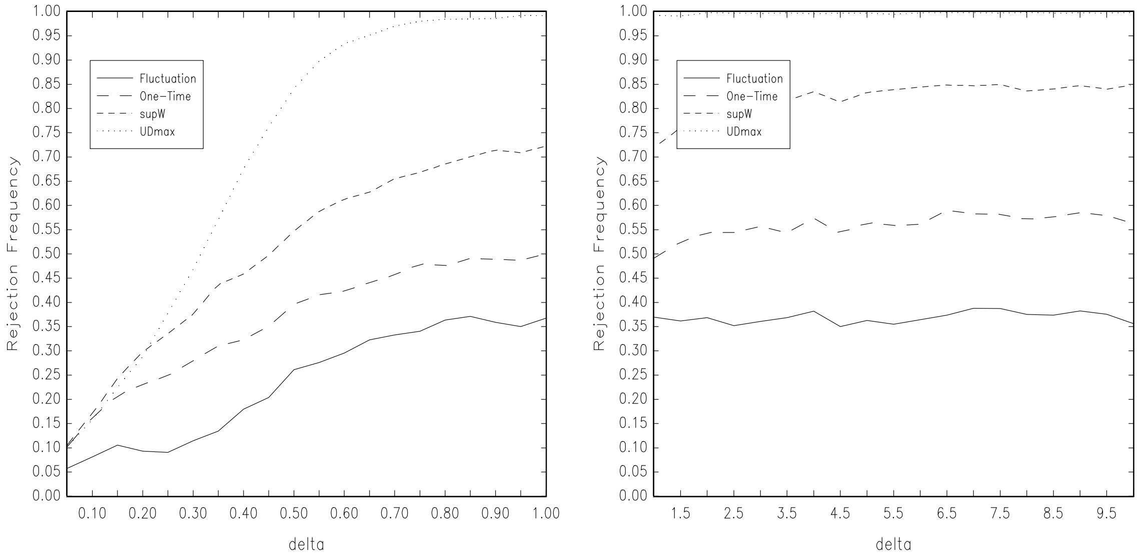

Figure 1.b: Power functions of the GW tests with a break in the relative performance at $\tau = 1/3$, $\mu = 0.7$.

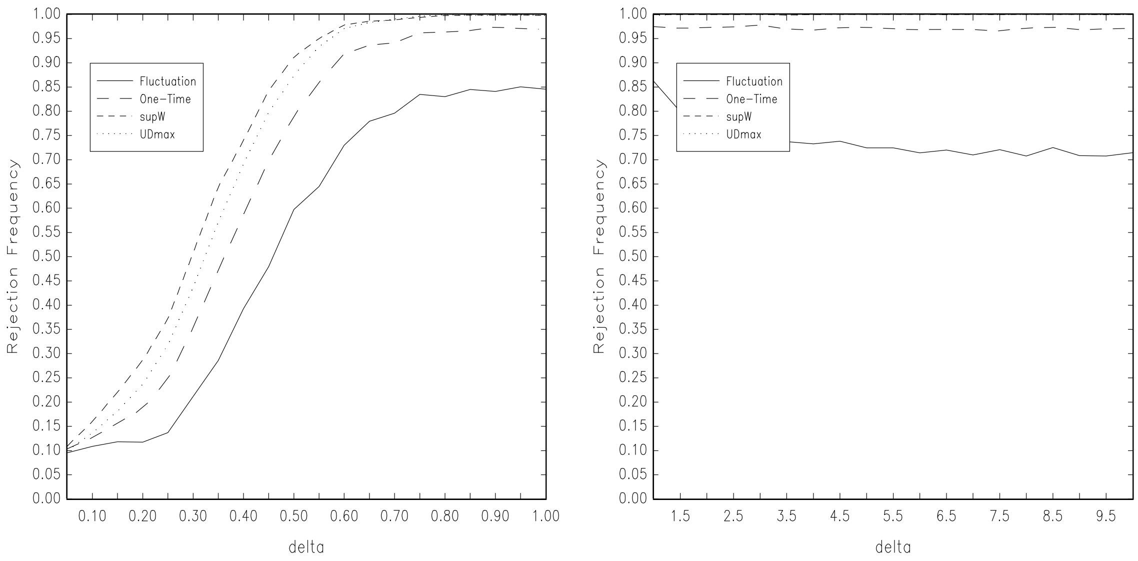

Figure 2.a: Power functions of the GW tests with a break in the relative performance at $\tau = 2/3$, $\mu = 0.3$.

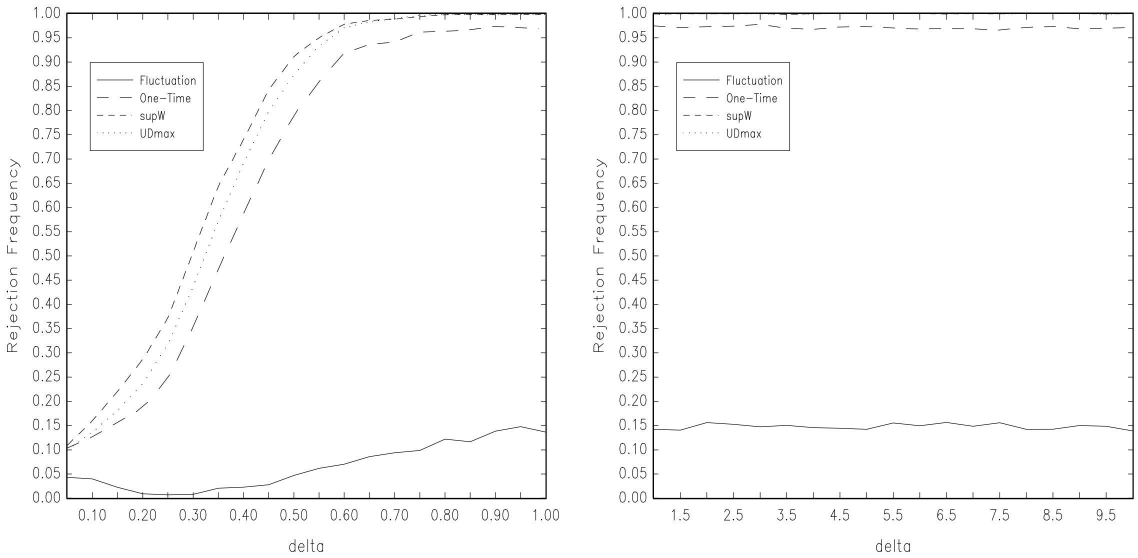

Figure 2.b: Power functions of the GW tests with a break in the relative performance at $\tau = 2/3$, $\mu = 0.7$.

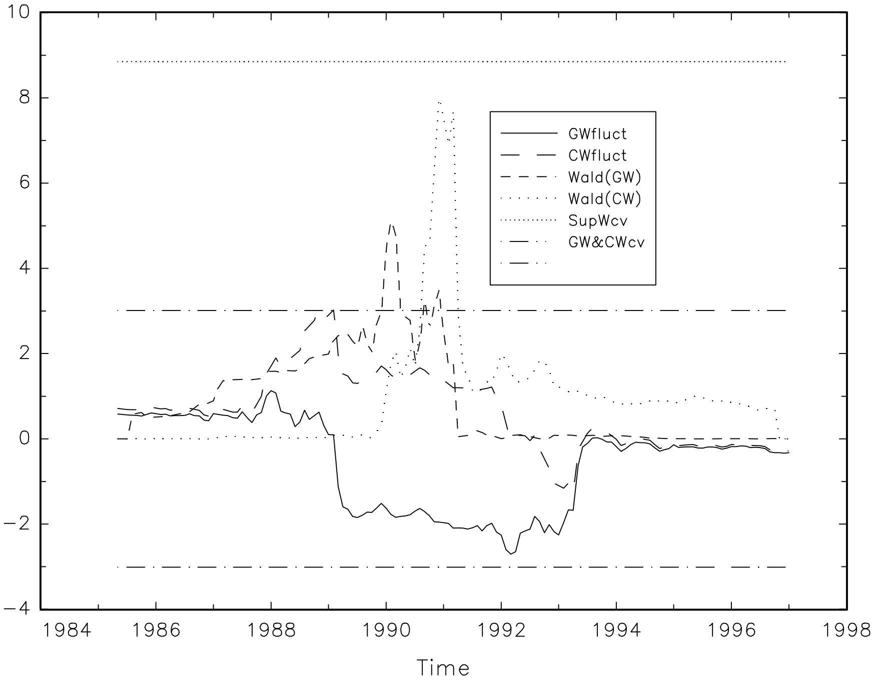

Figure 3: Empirical Results, Deutche Mark.

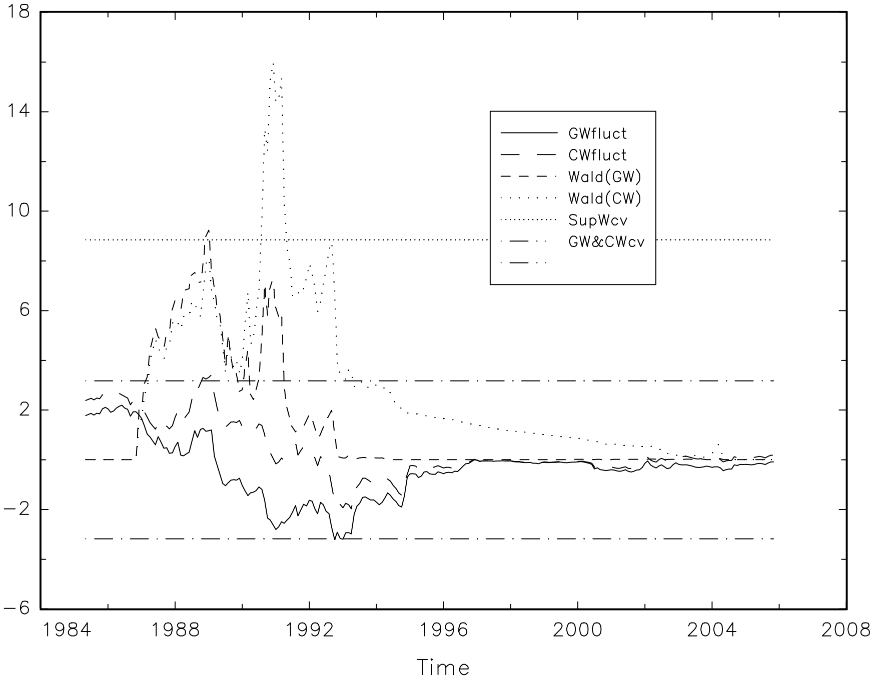

Figure 4: Empirical Results, UK Pound.