Robust Tests of Predictive Accuracy

Rosario Dell'Aquila*and Elvezio Ronchetti†

February, 2004

Abstract

We propose robust counterparts to tests of equal forecast accuracy such as those proposed by Diebold and Mariano (1995) and West (1996). We illustrate the robustness problem and evaluate the size and the power properties of the classical and robust tests under various types of deviations from model assumptions. The new robust test has a correct size and larger power across a wide spectrum of distributions including in particular heavy-tailed distributions.

*Quantitative Investment Research, Zürcher Kantonalbank, Switzerland and Università della Svizzera Italiana, Switzerland. e-mail: rosario.dellaquila@lu.unisi.ch

†Dept. of Econometrics, University of Geneva, Blv. Pont d'Arve 40, CH-1211 Geneva, Switzerland and Università della Svizzera Italiana, Switzerland. e-mail: elvezio.ronchetti@metri.unige.ch

1 Introduction

Comparing forecast accuracy is an important task in economics. In a seminal paper Diebold and Mariano (1995) proposed a widely applicable test of the null hypothesis of no difference in forecast accuracy of two competing forecasts. Recently these tests have been extended in several directions. For instance West (1996) and West and McCracken (1998) account for parameter estimation error, McCracken (2000) allows additionally for non differentiable loss functions, Corradi e al. (2001) extend the framework to the case of integrated and cointegrated variables and Sullivan et al. (1999) and White (2000) address the issue of joint comparison of more than two competing models. Other papers approach the issue of predictive accuracy testing by means of encompassing and related tests, see e.g. Chao et al. (2001), Clark and McCracken (2001), Harvey et al. (1997, 1998) and West (2001). Tests for forecast accuracy and forecast encompassing for nested models are proposed by Clark and McCracken (2001) and Corradi and Swanson (2002).

In this paper we analyze the robustness of the size and power properties of tests of equal forecast accuracy such as the tests proposed by Diebold and Mariano (1995) test (DM test henceforth) and propose robust version of this type of test. We mainly focus on the DM test since it is the most basic of the tests mentioned above. Consequences for the other tests, in particular for tests that account for estimation errors as e.g. West (1996) are similar. An additional robustness issue in this case arises because the additional term in the variance covariance matrix contains the estimated parameters which have to be estimated, assuming a reference model.

We formalize and analyze the robustness aspect of these tests and show how to improve them using the framework of *robust statistics*. The general idea of robust statistics is to provide estimators and tests that are stable

when the distributional assumptions are *slightly* different from the model assumptions. In the case of tests, this means, that the test should maintain *approximately* the correct size and not loose too much power when the empirical distribution differs (slightly) from the assumed model distribution. The impact of a slight distributional deviation from the assumptions or the effect of an outlier can be described in a natural way by defining estimators and tests as functionals of the underlying distribution. The main tool for the analysis is the influence function, which corresponds to a derivative in the functional space and helps to describe the influence of deviations in a small neighborhood of the reference distribution on the estimator or the test statistics. A general overview of robust statistics can be found in Hampel et al. (1986).

In our specific context, we have a distribution free test, so it may seem that in principle we should not care too much about deviations from the assumptions. However, there are several reasons for applying robust statistics in this context of tests for predictive accuracy.

First, using the basic tools like the influence function, we are able to show that the general impact of distributional deviations on the DM test statistics is high and we are able to formulate the trade-off between bias and efficiency.

Second, it is plausible to assume that a researcher would like to be aware of possible influential points, that is, points which have a large influence on the size and power of the test. Useful information can be obtained by analyzing the influential points implied by robust estimation or testing procedures as shown recently e.g. in Knez and Ready (1997), Dell'Aquila, Ronchetti and Trojani (2002) and Dell'Aquila and Ronchetti (2002). In a stock and bond return forecasting and risk modelling context, influential points may occur for instance when using e.g. an oil price variable or implied stock mar-

ket volatility as a predictor. In an illustrative example the need for such an analysis is additionally supported (see Section 2.2), by showing that an outlying observation which should help to stress the difference in forecast accuracy leads to a very low power in small and moderate samples. This is important in practice, since we want to make sure that e.g. a judgmental forecaster, who knows that his performance is compared by means of the DM test to other forecasts, does not exploit this fact, e.g. by making a bet which is distant from the mean in order to 'make' the test less powerful.

Third, an empirical distribution may be viewed as a distribution in an 'neighborhood' of the true distribution. We can therefore expect that in small samples a robust test has a stable size and power for a wide variety of distributions, while the performance of the classical test will depend more heavily on the underlying distribution. In fact we show in simulation that for a wide variety of distributions with fatter tails than the normal, the size of the DM test is distorted in small samples and that it has lower power than the proposed robust counterpart. Fatter tails than the normal may arise for several reasons in practice, e.g. a judgmental forecaster may have more or less pronounced view depending on whether one is more or less sure of the forecasts in particular periods or, in a regression setting, fat-tailed forecasts may reflect the fat tailedness of the regressors.

In this paper most of the analyses are carried out for the DM tests, since it is the most basic test for testing forecast accuracy. However, similar results apply to extensions of the DM test as presented in West (1996), McCracken (2000), West and McCracken (1998), where the additional uncertainty due to parameter estimation is taken into account. From a robustness perspective, an additional problem arises when the assumptions underlying the regression are not satisfied e.g. the normality assumption for the errors.

In a short Monte Carlo study, we show that the robustified DM test provides a more accurate size in small samples for fat tailed distributions and is more powerful than the classical test. We conclude that robust versions of tests for forecast accuracy are useful complements to the classical analysis and can be used routinely to support or question classical results and enhance the information set of the analyst.

The paper is structured as follows: In Section 2 we illustrate the robustness problem of the classical DM test by means of several illustrative examples, present the tools for analyzing the sensitivity of a test, and propose a robust version of the DM test. Section 3 presents the Monte Carlo analysis and illustrates the size and power properties of the classical and robust DM test. Section 4 concludes the paper.

2 Testing Equality of Forecast Accuracy

2.1 The Test Proposed by Diebold and Mariano (1995) and West (1996)

We consider a time series $\{y_t\}_{t=1}^T$ to be forecasted and two series of forecasts, $\{\widehat{y}_{1t}\}_{t=1}^T$ and $\{\widehat{y}_{2t}\}_{t=1}^T$. We denote the associated forecast errors by $\{e_{1t}\}_{t=1}^T$ and $\{e_{2t}\}_{t=1}^T$. The quality of the forecast is evaluated by means of a loss function $g(y_t, \widehat{y}_{it})$, and following Diebold and Mariano (1995) we write shortly $g(y_t, \widehat{y}_{it}) = g(e_{it})$. Diebold and Mariano (1995) propose a test for the null hypothesis of equal forecast accuracy, that is $E[d_t] = 0$, where $d_t = g(e_{1t}) - g(e_{2t})$ is the loss differential. Assuming that the loss-differential series $\{d_t\}_{t=1}^T$ is covariance stationary and short memory, they propose to test the null hypothesis that the population mean of the loss-differential μ is 0 by using a

version of the central limit theorem

$$\sqrt{T}(\overline{d} - \mu) \rightarrow N(0, 2\pi f_d(0)), \tag{1}$$

where

$$\overline{d} = \frac{1}{T}\sum_{t=1}^{T} d_t = \frac{1}{T}\sum_{t=1}^{T}[g(e_{1t}) - g(e_{2t})]$$

is the sample mean loss differential,

$$f_d(0) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\gamma_d(\tau)$$

is the spectral density of the loss differential at frequency 0, $\gamma_d(\tau) = E[(d_t - \mu)(d_{t-\tau} - \mu)]$ is the autocovariance of the loss differential at lag τ. Therefore the test statistic is given by

$$DM = \frac{\overline{d}}{\sqrt{\frac{2\pi\widehat{f}_d(0)}{T}}}, \tag{2}$$

where $\widehat{f}_d(0)$ is a consistent estimate of $f_d(0)$ defined by

$$\widehat{f}_d(0) = \frac{1}{2\pi}\sum_{\tau=-(T-1)}^{T-1} l(\frac{\tau}{L(T)})\widehat{\gamma}_d(\tau),$$

where

$$\widehat{\gamma}_d(\tau) = \frac{1}{T}\sum_{t=|\tau|+1}^{T}(d_t - \overline{d})(d_{t-|\tau|} - \overline{d}),$$

$l(\frac{\tau}{L(T)})$ is the lag window, and $L(T)$ is the truncation lag. Diebold and Mariano (1995) chose a rectangular lag window defined by

$$l(\frac{\tau}{L(T)}) = \begin{cases} = 1 & \text{for } \left|\frac{\tau}{L(T)}\right| \leq 1 \\ = 0 & \text{otherwise} \end{cases}$$

where $L(T) = k - 1$, arguing that k-step ahead forecast errors are at most $(k-1)$ dependent. Under the null hypothesis, the test statistics DM is

asymptotically $N(0,1)$ distributed[1]. The DM test can be easily applied to other situations. Harvey, Leybourne and Newbold (1998) propose the DM framework test for forecast encompassing. In particular the test is given by loss differentials of the form $d_t = e_{1t}\left(e_{1t} - e_{2t}\right)$, the null hypothesis is $E(d_t) = 0$, that is, forecast e_{1t} does not encompass forecast e_{2t}.

West (1996) generalizes the approach in Diebold and Mariano (1995) to account also for estimation errors of the unknown parameter β^*. The estimated parameter $\widehat{\beta}_t$ relies on data from perod t and earlier and can be estimated using for instance an expanding window[2], starting with R data points for the first estimation period, $R+1$ for the second, until $R+P-1 := T$. This estimated parameters are then used to produce P predictions τ-step ahead. It turns out that under weak conditions (see West (1996)

$$P^{1/2}(\overline{g} - Eg_t) \sim N(0, \Omega), \tag{3}$$

where

$$\Omega = S_{gg} + V(\beta),$$

and g_t is the loss function, $\overline{g} = \frac{1}{P}\sum_{t=R}^{T} g_{t+\tau}(\widehat{\beta}_t)$, $S_{gg} = \sum_{j=-\infty}^{\infty}(g_t - Eg_t)(g_{t-j} - Eg_t)'$ and the additional term $V(\beta)$ accounts for the parameter estimation error. The additional terms in the variance covariance matrix depend on the estimated parameters, the orthogonality conditions and the type of scheme used for the estimation. Therefore the distribution is still normal and the

[1]Harvey, Leybourne and Newbold (1997) assess the behavior of the DM test in moderate-sized sample and propose the modified test statistic for k-step ahead errors

$$MDM = T^{-1/2}[T + 1 - 2k + T^{-1}k(k-1)]^{1/2}DM.$$

This is based on the use of an estimator of the variance of $\overline{d}$ that is unbiased to order T^{-1}.

[2]Similar reasonings apply for a fixed or a rolling window, see McCracken and West (2000).

parameter estimation error is taken into account by additional terms in the variance covariance matrix. West (1996) discusses the conditions under which the additional terms vanish asymptotically and the test reduces to the DM test. For instance this is the case when there is no estimation error, or when $F := E[\frac{\partial g}{\partial \beta}(\beta^*)] = 0$, e.g. when using mean square prediction errors as a loss function and the predictors are uncorrelated with the prediction errors.

A test for equal forecast accuracy can then be constructed by using (3)[3], e.g. in the case of two forecasts $\alpha = (1, -1)'$

$$P^{1/2}(\alpha'\overline{g}) \sim N(0, \alpha'\widehat{\Omega}\alpha). \tag{4}$$

Thus with regard to testing we have exactly the same structure as in the case of DM.

2.2 Sensitivity Analysis: Some Illustrative Examples

In this section, we present some short examples to illustrate the sensitivity of the classical DM test. Even more pronounced results apply in the case of West (1996), which accounts also for the additional variability due to parameter estimation. In the first two examples we illustrate the effect of an outlying point on level and power of the test. In the first experiment, we draw observations from two independent normal distributions and then add one observation, which lies away from the majority of the data. We would expect that the DM test rejects the null of equal forecast accuracy, since the two series are now different. However, when we draw 10000 times sets of 200 observations from a bivariate normal and each time put one observation at

[3] Alternatively a tests of equal forecast accuracy with a χ^2 distribution (as West (1996) uses in his examples) may be constructed in a similar way e.g. by means of $P(\alpha\overline{f})^2/\alpha'\widehat{\Omega}\alpha \sim \chi^2(1)$. In our one dimensional case they are equivalent.

Table 1: Illustrative Example I

Two independent bivariate normal error series are simulated and an outlying point was added to the first series. In the Table rejection frequencies of the DM test at the nominal level of 5% are reported.

T	32	64	128	256	512	1024
Frequency of rejections	0.0001	0.0001	0.0001	0.001	0.0055	0.0181

10 in the first series, the DM test (with a quadratic loss function) does not reject the null hypothesis of equal forecast accuracy. In Table 1 we report the rejection frequencies for the experiment above for the DM test with nominal size of 5%. We would expect an empirical power above 0.05. Surprisingly the empirical power goes down. The DM test does not detect the difference in the series. This is counter-intuitive: we would expect that adding the observation would strengthen the difference in the error series. However, the test goes in the opposite direction, with an extremely low rejection frequency, because the variance is inflated.

Similarly in the second example we generate two normal error series with zero mean and with variances 1.2 for the first series and 1.0 for the second error series. The DM test should therefore reject the null of no difference in forecast accuracy. In the first row of Table 2, we report the rejection frequencies at the nominal level of 5% when testing with the classical DM test. The rejection rates range from 13% with 32 observations to 89% for 1024 observations. In the third row, the same rejection rates are reported, but a single outlying point (with value 10) has been added to the first error series. We would expect the power to go up, since the outlying point should stress the difference in the forecast accuracy of the two series. Surprisingly

Table 2: Illustrative Example II

Two independent bivariate normal error series are simulated, with variances 1.2 for the first series and 1 for the second error series. In the first and second row we report the rejection frequencies of the classical and the robust DM test with quadratic loss at the nominal level of 5%. In the third and forth row, the same rejection rates are reported, but a single outlying point (with value 10) was added to the first error series.

T	32	64	128	256	512	1024
Cl. DM test	0.130	0.174	0.272	0.428	0.668	0.895
Rob. DM test	0.128	0.172	0.264	0.411	0.650	0.879
Cl. DM test (with outlying obs.)	0.000	0.002	0.046	0.259	0.656	0.928
Rob. DM test (with outying obs.)	0.208	0.243	0.330	0.468	0.688	0.893

the empirical power goes down except when $T = 1024$. If we test instead using the robust version of the DM test defined in Section 2.4, we notice that the power of the robust test is approximately equal in the first case (second row) and rises, as we would expect, when we put an outlying observation in the first series (forth row).

To avoid confusion we stress, that the two illustrative examples above do not simply illustrate the impact of 'outliers' on level and power of the DM test. Since the test is nonparametric, we can interpret the observations of the first series as drawn from a distribution that generates such points with some probability and thus satisfies the conditions to apply the DM test. Since the results in the two illustrative examples above are so striking, we do not include this type of distributions in the Monte Carlo analysis in Section 3.

In the third illustrative example, another aspect of the sensitivity of the DM test is highlighted. Indeed we show, that distributions with longer tails

than the normal may lead to a bias in the size and to a low power (see Section 3) for typical finite samples. Notice again that the DM test does not assume a particular distribution and we thus really focus on small sample aspects of the test. The empirical distribution may be interpreted as a distribution in the 'neighborhood' of the true underlying distribution. We draw 100 observations from two standardized and independent normal, t_5, t_3^{emp}[4] and contaminated normal $CN(0.05, 25)$ distributions and perform the classical and robust DM test (see again Section 2.4 for the robust version of the DM test) with squared error as loss function. This is repeated 10000 times and the empirical sizes are reported in Figure 1. The solid line corresponds to the empirical sizes of the classical DM test, the dashed to the empirical sizes of the robust version of the DM test and the dotted lines are the nominal sizes.

We see from Figure 1 that the empirical size is too small up to sizes of 10% and too large for sizes above 10%. On the other hand the robust empirical sizes are very accurate. Similar results but different curves for the empirical sizes for the DM test arise for different type of error distributions. Notice that the simulations in Diebold and Mariano (1995) are all at the 10% level, where the empirical size seems to be much more accurate, at least for the distributions we look at.

From the illustrative examples above we see, that single observations, or long tailed error distributions can lead the DM test not to differentiate between two different distributions and that the empirical size may not match the nominal size even with a large number of observations.

[4]The t_3^{emp} distribution is the empirical distribution of a sample of 10000 observations drawn from a t_3 distribution. In this way me make sure that the loss differential has finite second moment e.g. when using a quadratic loss.

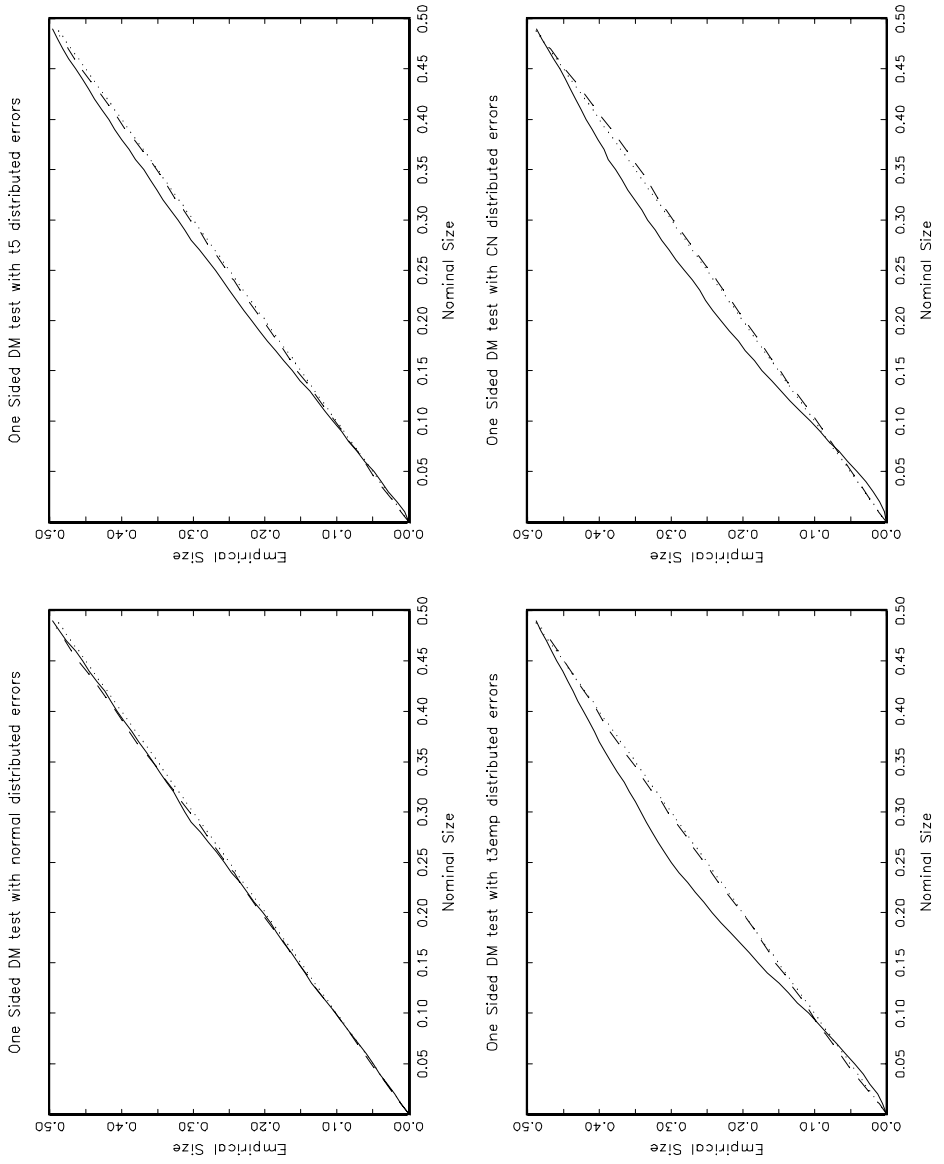

Figure 1: Nominal and empirical sizes of the classical Diebold and Mariano (1995) test and of a robust version of this test. We draw 100 observations from a normal, t_5, t_3^{emp} and CN(0.05,25) distribution and perform the classical and robust DM test (see Section 2.4) with the squared error as loss function. The solid line corresponds to the empirical sizes of the classical DM test, the dashed to the empirical sizes of the robust version of the DM test when repeating this procedure 10000 times. The dotted lines correspond to the 45 degrees line where the empirical size matches the nominal size.

In the following Sections we use the framework of robust statistics to formalize and analyze the sensitivity of the DM test and to construct robust versions of the DM test. Notice that we carry out all the simulations for the DM tests. Similar results (not shown here) apply to extensions of the DM test as presented in West (1996), McCracken (2000), West and McCracken (1998). For these tests an additional robustness issue is due to the fact that the regression parameters have to be estimated and the additional terms of the variance covariance matrix are constructed using the estimated parameter values.

2.3 Approximation of the Asymptotic Level and the Asymptotic Power

In this section, we first analyze how the effect of a general contamination influences the size and the power of a test in a general framework. To be more precise, let $X_1, ..., X_n$ be the sample and assume that X_i is distributed according to F_0. Further let F_n be the empirical distribution given by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \Delta_{X_i}(x)$$

where

$$\Delta_{X_i}(x) = \begin{cases} 0 & X_i > x \\ 1 & X_i \leq x \end{cases}$$

The DM test statistics S_n can now be written as a functional, that is a function of a distribution. In particular we can write the DM test statistics as $S_n = S(F_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$ and the corresponding functional for an arbitrary distribution F is given by $S(F) = E_F(X_1)$.

We now analyze the sensitivity of the size and power of a test based on a

test statistic with an asymptotic normal distribution[5], thus tests of the form

$$\sqrt{n}(S_n - S(F)) \underset{n\to\infty}{\overset{D}{\to}} N(0, V(S,F)) \tag{5}$$

where $V(S,F)$ is the asymptotic variance. This is for example the case of Diebold and Mariano (1995) and West (1996) with S_n being the loss-differentials, see equations (1) and (4). In particular under the null-hypothesis H_0 of no difference in forecast accuracy, $S(F) = 0$.

Then we reject H_0 if $S_n > k_{n,\alpha_0}$, where a_0 is the nominal level of the test. In order to quantify the impact of a slightly different distribution on the size and the power of the test we calculate the asymptotic power, when the underlying distribution is given by[6]

$$\widetilde{F}_{n,\Delta,\varepsilon} = (1 - \frac{\varepsilon}{\sqrt{n}})F_{\frac{\Delta}{\sqrt{n}}} + \frac{\varepsilon}{\sqrt{n}}G, \tag{6}$$

where G is an arbitrary contaminating distribution and $F_{\frac{\Delta}{\sqrt{n}}}$ is a sequence of contiguous alternatives. In Appendix 5.1 we show (c.f. also HRRS (1986)), that the asymptotic power under $\widetilde{F}_{n,\Delta,\varepsilon}$ is given by

$$\lim_{n\to\infty}\beta(\widetilde{F}_{n,\Delta,\varepsilon}) = 1 - \Phi\left(\Phi^{-1}(1-\alpha_0) - \Delta\sqrt{E(S,F_0)} - \varepsilon\frac{\int IF(x;S,F_0)dG(x)}{[V(S,F_0)]^{1/2}}\right), \tag{7}$$

where α_0 is the asymptotic nominal level, Φ and ϕ are the cumulative distribution and the density of the standard normal distribution respectively, $E(S,F_0) = \frac{\left(\frac{\partial}{\partial\Delta}S(F_\Delta)|_{\tilde{\Delta}=0}\right)^2}{V(S,F_0)}$ is the Pitman efficacy of the test, F_0 is the model distribution and $IF(x;S,F_0)/[V(S,F_0)]^{1/2}$ is the self standardized influence function defined in (12) in the Appendix. From there it follows

$$IF(x;S,F_0) = \frac{\partial}{\partial\varepsilon}S((1-\varepsilon)F_0 + \varepsilon\Delta_x)\Big|_{\varepsilon=0}.$$

[5]Similar reasonings can be applied to test with a χ^2 distribution, see Heritier and Ronchetti (1994).

[6]Notice, that we do not consider changes in the serial correlation of the correlation between the series.

Thus, the self-standardized influence function describes the standardized impact of a point mass contamination on the test statistics S. It is the Gâteaux derivative of the functional S and plays the same role as the derivative in real analysis. By means of the influence function, we can derive the first order approximations of the effect of a general contamination on estimators or size and power of a test; see HRRS (1986) for more details.

From (7) we can see, that it is sufficient to bound the self-standardized influence function

$$IF_s(x; S, F_0) = \frac{IF(x; S, F_0)}{[V(S, F_0)]^{1/2}} \tag{8}$$

in order to limit the bias in the power or in the size of a test. In particular, we immediately see that the minimal power of the test in the neighborhood (6) is given by

$$\beta_{\min} = \inf_G(\beta(\widetilde{F}_{n,\Delta,\varepsilon})) = 1 - \Phi\left(\Phi^{-1}(1-\alpha_0) - \Delta\sqrt{E(S, F_0)} - \varepsilon \inf_x IF_s(x; S, F_0)\right) \tag{9}$$

Bounding the influence function is therefore enough to maintain the power in a pre-specified band around β_0, the asymptotic power at the model. The approximation for the asymptotic level can be obtained by putting $\Delta = 0$ in (9), that is

$$\begin{aligned} \lim_{n\to\infty} \alpha(\widetilde{F}_{n,0,\varepsilon}) &= 1 - \Phi\left(\Phi^{-1}(1-\alpha_0) - \varepsilon \int IF_s(x; S, F_0) dG(x)\right) \\ &= \alpha_0 + \varepsilon\phi(\Phi^{-1}(1-\alpha_0)) \int IF_s(x; S, F_0) dG(x) + o(\varepsilon) \end{aligned}$$

where α_0 is the asymptotic nominal level. In this case too, we see, that $\alpha(\widetilde{F}_{n,0,\varepsilon})$ remains between prespecified bounds of α_0 by bounding the influence function $IF_s(x; S, F_0)$. Specifically

$$\alpha_{\max} = \sup_G(\alpha(\widetilde{F}_{n,0,\varepsilon})) = 1 - \Phi\left(\Phi^{-1}(1-\alpha_0) - \varepsilon \sup_x IF_s(x; S, F_0)\right) \tag{10}$$

The choice of the robustness parameter c^* to bound the asymptotic bias of the size can be determined by determining the resulting impact on the size of the test using

$$\alpha_{\max} = 1 - \Phi\left(\Phi^{-1}(1-\alpha_0) - \varepsilon c^*\right), \tag{11}$$

where α_0 is the size at which the test is performed and ε is the assumed contamination. Notice, that the value of $c^* := \sup_x IF_s(x; S, F_0)$ determines the level of robustness and therefore the magnitude of the maximal bias in the size of the test[7]. In Figure 2 we plot the bias of the size as a function of the amount of contamination for various levels of c^*. We notice that when the level of c^* is high, the bias grows very fast around $\varepsilon = 0$. This means that in such cases, small deviations from the model distribution or deviations due to small samples lead to testing results which can be quite different from those obtained under the model or those based on the asymptotic theory.

The constant c^* effectively determines the trade-off between bias and efficiency. A very high value of c^* is equivalent to performing a classical test.

The results of the illustrative examples in Section 2.2 can now be explained by means of our analysis. They are due to the unboundedness of the influence function of the test statistic of the classical DM test.

2.4 A Robust Tests for Forecast Accuracy

Starting from the robust framework above, we can propose different bounded influence versions of the DM test. The chosen test can be based on the particular structure of the data at hand. First, we propose a test for the case where the loss differentials are approximately symmetric and secondly for the case where forecast errors are approximately symmetric.

[7]For a two sided test it follows that $\lim_{n\to\infty} \alpha(\widetilde{F}_{n,0,\varepsilon}) = 2 \cdot \left(1 - \Phi\left(\Phi^{-1}(1-\frac{\alpha_0}{2}) - \varepsilon c^*\right)\right)$

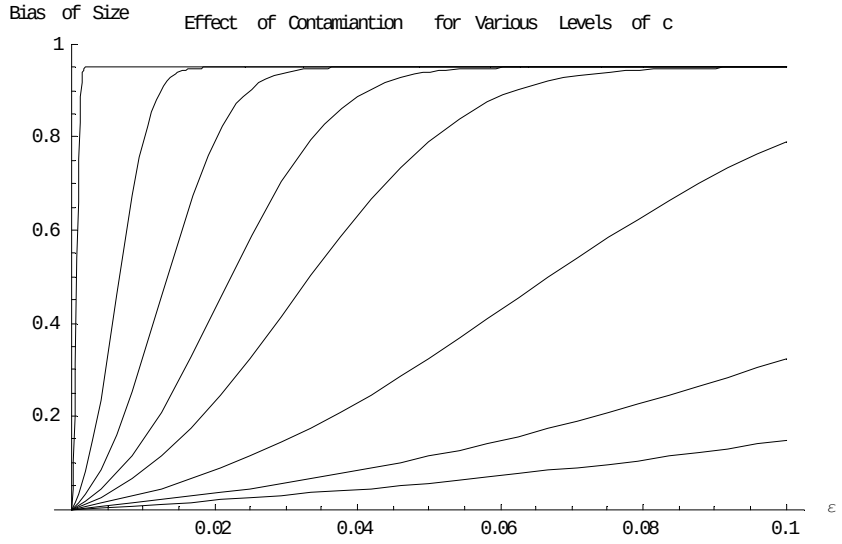

Figure 2: Relation between bias of size and amount of contamination. The different curves correspond to different levels of c^*. ε is the degree of contamination.

Consider first the case where the true underlying distribution of the loss differentials d_t is in a neighborhood of a symmetric (model) distribution. Then we modify the DM test statistic as follows

$$S_{DMR1} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \psi(d_t),$$

where ψ is the Huber function given by[8] $\psi_c(x) = \max(\min(x, c), -c)$. Then, as in Diebold and Mariano (1995) we obtain that under $H_0 : E(\psi(d)) = 0$

$$\sqrt{T} S_{DMR1} \underset{n\to\infty}{\overset{D}{\to}} N(0, 2\pi f_{\psi(d)}(0)),$$

where $f_{\psi(d)}(0) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{+\infty} \gamma_{\psi(d)}(\tau)$ and $\gamma_{\psi(d)}(\tau) = E\left(\psi(d_t)\psi(d_{t-\tau})\right)$. The influence function is given by $IF(d; S_{DMR1}, F_0) = \psi(d)$ and the robustness constant c may be determined making use of (11) and without specifying

[8]Other ψ functions may be chosen. However, for this type of test, ψ cannot be re-descending; see HRRS (1986) for a general discussion.

a reference model. Notice that when $c = \infty$ we obtain the classical DM test statistics. The test may also be simply interpreted as a modification of the DM test, where one is more interested in modelling the majority of the data, in order to check whether the classical testing result is driven by only a few datapoints[9]. Notice also, that the function ψ can be bounded in an asymmetric way by choosing the upper bound to control the maximal level see equation (10) and the lower bound to control the minimum power, see equation (9). This can be useful when the loss differentials are asymmetric.

When asymmetric loss differentials are a concern, but we can still assume that the forecast errors are approximately symmetric, another robust version of the DM test can be constructed

$$S_{DMR2} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \psi(d_t - a),$$

where a is determined such that $E_{F_0}(\psi(d-a)) = E_{F_0}(d)$, that is a corrects for the bias induced by truncating non-symmetric loss-differentials. By standard arguments with the central limit theorem as above, we can then establish the asymptotic normality as in the case above. Notice, that a is 0 for symmetric loss differentials and reduces to the S_{DMR1} test statistic. For non-symmetric loss-differentials it may be determined assuming an explicit reference model F for the errors. Notice, that in robust statistics this is not too restrictive. Indeed when e.g. a normal model is assumed, robustness is also guaranteed

[9]Notice, that a corresponding test when the loss-differentials are symmetric but not around 0 can easily be constructed. Indeed, we can use the test statistics $S^*_{DMR1} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \psi(d_t - L) + L$, where L is a robust measure of location such as the median or an α-trimmed mean. Then using the same CLT as above, we find that the distribution of the S^*_{DMR} under the null of no difference in forecast accuracy is $N(0, 2\pi f_{\psi(d-\mu)+\mu})$ and the influence function of the test statistics given by $IF(d; S_{DMR1}, F_0) = (\psi(d - L(F_0)) - E_{F_0}(\psi(d - L(F_0))) \cdot IF(d; L, F_0)) + IF(d; L, F_0)$ which can again be used to determine the constant c in order to bound the bias of the size below a specific level.

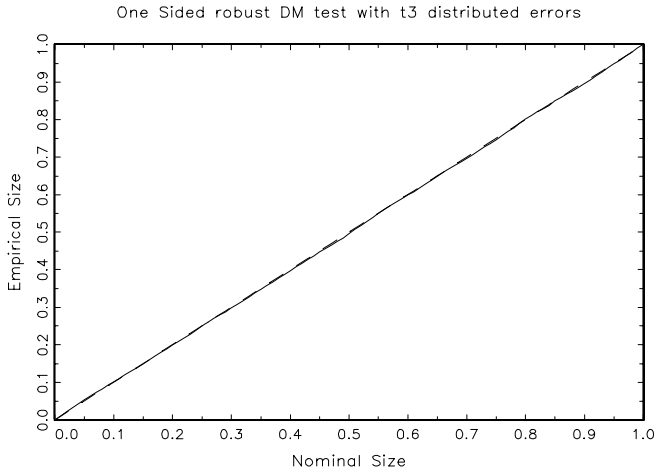

Figure 3: Nominal and empirical sizes of the robust DM test. We draw 100 observations from a t_3 distribution and perform the robust DM test. The solid line corresponds to the empirical sizes of the robust version of the DM test when repeating this procedure 10000 times. The dashed line correspond to the 45 degrees line where the empirical size matches the nominal size.

for error distributions which are in a neighborhood of the normal as e.g. a t_5 distribution or a distribution which is slightly asymmetric. Notice that in this case location, scale and correlation of the forecast error series have to be first estimated.

Notice that the boundedness of ψ ensures that the loss differential has finite second moment. Therefore even when the underlying distribution and the loss function are such that that the DM test cannot be applied (e.g. a t_3 distribution with a quadratic loss function) the boundedness of the loss differentials ensures that equation (5) holds. As an illustration we report in Figure 3 the empirical size of the robust DM test with a quadratic loss function for t_3 distributed errors when drawing 10000 times 100 observations from two independent t_3 distributions and performing the robust DM test.

The robust analysis can be complemented by the analysis of the influential

points resulting from robust testing. Consider the test statistic $S_{DMR1} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} w_c(d_t) \cdot d_t$, where $w_c(d_t) = \frac{\psi(d_t)}{d_t}$. Therefore, S_{DMR1} can be viewed as a weighted version of the classical DM statistic. These weights have values between 0 and 1. A weight of 1 corresponds to no downweighting. The weights arising from the robust analysis can give important hints about possible deviations from the general structure of the data or about particular periods where the forecast/forecaster was performing better or worse.

Notice that the test structure above applies also in the case analyzed by West (1996), who takes into account parameter estimation. Indeed there is simply an additional term in the variance that takes into account the parameter estimation error, see equation (4). An additional and different robustness issue in West (1996) is given by the the fact that the regression parameters have to be estimated e.g. by OLS. Indeed fat tailed error distributions or outliers may induce less efficient or distorted coefficients and an inflated variance covariance matrix estimation. Robust regression estimators may be used in this context (see HRRS (1986) for an overview); in particular an M-estimate $\widehat{\beta}_t$ can be linearized such that $\widehat{\beta}_t - \beta^* = B(t)H(t)$, where $B(t)$ is a $(k \times q)$ matrix, $H(t)$ is a $(q \times 1)$ matrix with $B(t) \xrightarrow{a.s} B$, B a matrix of rank k, $H(t) = \frac{1}{t} \sum_{s=1}^{t} h_s(\beta^*)$ for a $(q \times 1)$ orthogonality condition $h_s(\beta^*)$ and $Eh_s(\beta^*) = 0$, and where equality means "asymptotically equivalent", as required in the framework of West (1996). Notice, that the robust parameter estimation is a different problem from the robust testing of forecast accuracy. Even when estimating robustly the parameters, the loss differential may still be fat tailed and thus a robust test for forecast accuracy is needed.

Finally, notice, that tests for forecast encompassing can be treated in exactly the same way. In this case we put $d_t = e_{1t}(e_{1t} - e_{2t})$.

3 Monte Carlo Analysis

3.1 Experimental Design

In order to evaluate the finite sample size of and power properties of the classical and robust DM test across a large spectrum of underlying distributions, we perform a Monte Carlo analysis using the same experimental design as Diebold and Mariano(1995). We draw realizations of the bivariate forecast-error process, $\{e_{1t}, e_{2t}\}_{t=1}^{T}$, with varying degrees of contemporaneous and serial correlation by first drawing realizations $u_t = \{e_{1t}, e_{2t}\}_{t=1}^{T}$, where e_{1t}, e_{2t} are independent. We then construct forecast-errors with varying degrees of contemporaneous correlation ρ and serial correlation (moving average MA(1) with parameter θ) by premultiplying the independent errors u_t with the Cholesky factor

$$\Gamma = \begin{pmatrix} \sqrt{k} & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix}.$$

The transformed errors $v_t = \Gamma u_t$ have now a contemporaneous correlation ρ. We set $k = 1$ for the analysis of the size and $k > 1$ for the analysis of the power. In a second step we introduce serial correlation by taking

$$\begin{pmatrix} e_{it} \\ e_{jt} \end{pmatrix} = \begin{pmatrix} \frac{1+\theta L}{\sqrt{1+\theta^2}} & 0 \\ 0 & \frac{1+\theta L}{\sqrt{1+\theta^2}} \end{pmatrix} \begin{pmatrix} v_{it} \\ v_{jt} \end{pmatrix}$$

and $v_{0t} = 0$.[10]

For the Monte Carlo analysis, we consider sample sizes of $T = 32, 64, 128, 256, 512, 1024$, contemporaneous correlation parameters $\rho = 0, 0.5, 0.9$, and MA parameters $\theta = 0.5, 0.9$ for the two step ahead forecast errors. We consider the following

[10] Multiplication by $1/\sqrt{1+\theta^2}$ keeps the unconditional variance normalized to 1.

distributions for the forecast-errors e_{1t}, e_{2t} [11]:

- bivariate normal,
- Student t_6 distribution
- Student t_5 distribution
- Student t_3^{emp} distribution (defined in footnote 4)
- contaminated normal

$$F(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(\frac{x}{K}),$$

where Φ is the cumulative distribution function of a standard normal random variable. $\varepsilon = 0.05$, $K = 25; 100$.

We report the empirical sizes for the nominal size $\alpha = 0.05$[12] and use 10000 Monte Carlo replications. The truncation lag is set at 0, 1 respectively. Thus as in the DM simulation setup we consider one-step ahead and two-step ahead forecasting errors.

We analyze the power by constructing two error series in the same manner as shown above, but we set $k \neq 1$. We choose the parameter values $k = 1.4$. Similar results arise for $k = 1.2$.

[11] As in Diebold and Mariano (1995) all error distributions are standardized in order to have variance equal to 1.

[12] The result for other nominal sizes are qualitatively similar, although the bias may go in the opposite direction (too high), as seen in the third illustrative example in Section 2.2.

3.2 Results

In Table 3 we report the simulation results for the two sided $DMR1$ test for the empirical size at 5% in the robust and classical case for one step ahead forecast errors. For all the simulations we use a quadratic loss function.

For all the simulations we choose fixed values of c depending on the value of ρ. In particular, for $\rho = 0,\ 0.5,\ 0.9$, the corresponding values of c are 8, 7 and 3.5 respectively. These are obtained by the two sided version of (11), in order to bound the maximal bias of the size below 0.05% assuming approximately a contamination of $\varepsilon = 1\%$ when the underlying reference model is bivariate normal.

Table 3 reports the simulations for the empirical size when $\theta = 0$. We notice, that in the classical DM case (with truncation lag $L(T) = 0$), the empirical sizes are correct for the normal distribution, but begin to be biased (too low) for t_6 errors. For fatter tailed distributions, the empirical size is downward biased. Notice that for higher nominal size values (e.g. 15%, or 10% one sided test) the empirical size will be upward biased. On the other hand, the robust DM test shows a great stability of the size (around 0.05) across distributions, sample sizes and correlations.

Table 4 presents the same simulation for two step ahead forecast errors with $\theta = 0.5$[13]. The variance was estimated with $L(T) = 1$. In this case, the values of c are slightly lower than in the previous case. Taking into account the serial correlation, we obtain by (11) that for $\rho = 0,\ 0.5,\ 0.9$, the corresponding values of c are approximately 9, 8 and 4 respectively. Again we see a similar structure as in the first case for T above 128. Forecast errors with fat tails lead to distorted sizes in the classical case, while the robust test delivers correct and stable sizes. For small samples the values seem to be too

[13]The results for $\theta = 0.9$ are similar and are omitted for brevity.

high in the robust case, and in the classical case for the bivariate normal. Notice however, that there are two sources of biases in this case. The DM test is oversized in small samples and at the same time undersized for fatter tailed distributions as we can see when looking at large sample sizes for fatter tailed distributions (e.g. see t_3^{emp} series, with $T = 1024$).

We now analyze the power properties of the robust and classical DM tests. We simulate the error distributions as described in the Section 3, and set $k = 1.4$. We report the results for the empirical power for the 5% quantile in the robust and classical case for a one step ahead forecast errors. The tests are one sided.

We can see that the power is broadly similar in the bivariate normal case. Remarkably however, we see, that for distributions with fatter tails, the power of the robust test is always higher than that of the classical test.

All the simulations above have been performed with the $DMR1$ test. Similar results arise when using the $DMR1^*$ and $DMR2$ for the simulation of the power.

4 Conclusions

We propose a robust version of tests of forecast accuracy such as the test proposed by Diebold and Mariano (1995) and by West (1996). The robust version has a stable size and power when the underlying distribution of the errors deviates from the assumed underlying distribution. In particular we have shown the extreme sensitivity of this test to deviations from the model distribution. This has an effect on the small sample properties of the DM test, when the empirical distribution is seen as a contaminated version of the true underlying distribution. In addition, we have shown that a single

observation may produce paradoxical results and that it is therefore useful to complement the classical DM test with a robust version.

We find that the robust versions of the DM test are a good complement to the classical test and can be used routinely to support or question classical results and enhance the information set of the analyst.

Table 3: Empirical Size Under Quadratic Loss

Classical and Robust DM test with squared error as loss function for one step ahead forecast errors. e_1 and e_2 are bivariate standard normal (BN), t_6,t_5, t_3^{emp}, CN(0.05,25) and CN(0.05,100), with correlation given by ρ and $\theta = 0$. Therefore in this simulation truncation lag $L(T) = 0$. All test are two sided. The robustness constant c^* is chosen to bound the maximal bias around 0.5% of the nominal 5% size for an assumed contamination of $\varepsilon = 0.01$. The simulation was performed with 10000 runs. Empirical sizes are reported.

		BN		t_6		t_5		$temp_3$		CN(0.05;25)		CN(0.05;100)	
T	ρ	CL	Rob	CL	Rob	CL	Rob	CL	Rob	CL	Rob	CL	Rob
32	0.0	0.0586	0.0600	0.0487	0.0573	0.0471	0.059	0.0369	0.0516	0.0219	0.0462	0.0235	0.0510
32	0.5	0.0585	0.0593	0.0481	0.0569	0.0507	0.0603	0.0405	0.0562	0.0260	0.0482	0.0249	0.0469
32	0.9	0.0595	0.0599	0.0550	0.0586	0.0519	0.058	0.0412	0.0516	0.0335	0.0438	0.0303	0.0406
64	0.0	0.0533	0.0543	0.0455	0.0530	0.0456	0.0561	0.034	0.0538	0.0213	0.0519	0.0350	0.0541
64	0.5	0.0518	0.0527	0.0486	0.0574	0.0442	0.0548	0.038	0.0533	0.0254	0.0540	0.0303	0.0486
64	0.9	0.0531	0.0544	0.0540	0.0574	0.047	0.0518	0.0388	0.0492	0.0312	0.0472	0.0366	0.0521
128	0.0	0.0506	0.0522	0.0434	0.0498	0.0451	0.0535	0.0342	0.051	0.0320	0.0507	0.0463	0.0581
128	0.5	0.0528	0.0534	0.0476	0.0529	0.0447	0.0509	0.0362	0.0511	0.0337	0.0546	0.0405	0.0499
128	0.9	0.0495	0.0498	0.0503	0.0539	0.0452	0.0515	0.0381	0.0488	0.0362	0.0501	0.0429	0.0515
256	0.0	0.0515	0.0505	0.0482	0.0529	0.0448	0.0513	0.0366	0.0552	0.0385	0.0506	0.0481	0.0496
256	0.5	0.0508	0.0519	0.0507	0.0528	0.048	0.0542	0.0339	0.0507	0.0404	0.0467	0.0481	0.0515
256	0.9	0.0508	0.0510	0.0551	0.0567	0.048	0.0513	0.0387	0.0489	0.0419	0.0499	0.0471	0.0508
512	0.0	0.0559	0.0549	0.0485	0.0531	0.0477	0.0545	0.0344	0.0476	0.0428	0.0509	0.0471	0.0477
512	0.5	0.0531	0.0531	0.0452	0.0471	0.0455	0.0493	0.0362	0.0476	0.0421	0.0501	0.0502	0.0510
512	0.9	0.0512	0.0513	0.0493	0.0498	0.0454	0.0478	0.0385	0.05	0.0486	0.0523	0.0494	0.0493
1024	0.0	0.0465	0.0477	0.0482	0.0522	0.0475	0.0483	0.0355	0.0513	0.0489	0.0494	0.0482	0.0495
1024	0.5	0.0515	0.0521	0.0491	0.0501	0.0464	0.0502	0.0374	0.0499	0.0538	0.0500	0.0522	0.0533
1024	0.9	0.0501	0.0499	0.0488	0.0502	0.0475	0.0512	0.0408	0.0531	0.0483	0.0498	0.0514	0.0503

Table 4: Empirical Size Under Quadratic Loss

Classical and Robust DM test with squared error loss function for two step ahead forecast errors. e_1 and e_2 are bivariate standard normal (BN), t_6, t_3, CN(0.05,25), and CN(0.05,100), with a correlation given by ρ and $\theta = 0.5$. Therefore $L(T) = 1$. All test are two sided. The robustness constant c^* is chosen to bound the maximal bias around 0.5% of the nominal 5% size for an assumed contamination of $\varepsilon = 0.01$. Empirical sizes are reported.

		BN		t_6		t_5		$temp3$		CN(0.05;25)		CN(0.05;100)	
T	ρ	CL	Rob	CL	Rob	CL	Rob	CL	Rob	CL	Rob	CL	Rob
32	0.0	0.0814	0.0823	0.0717	0.0793	0.0683	0.0762	0.0541	0.0703	0.0400	0.0602	0.0465	0.0687
32	0.5	0.0840	0.0847	0.0684	0.0742	0.0684	0.0753	0.0552	0.0690	0.0448	0.0637	0.0441	0.0630
32	0.9	0.0824	0.0829	0.0795	0.0821	0.0751	0.0778	0.0648	0.0703	0.0513	0.0591	0.0528	0.0608
64	0.0	0.0650	0.0658	0.0548	0.0630	0.0513	0.0613	0.041	0.0632	0.0316	0.0584	0.0440	0.0608
64	0.5	0.0611	0.0632	0.0602	0.0664	0.0535	0.0625	0.0484	0.063	0.0347	0.0572	0.0410	0.0588
64	0.9	0.0646	0.0654	0.0589	0.0621	0.0598	0.0636	0.0481	0.059	0.0422	0.0539	0.0458	0.0568
128	0.0	0.0577	0.0578	0.0535	0.0583	0.0534	0.0622	0.0406	0.0581	0.0371	0.0523	0.0493	0.0607
128	0.5	0.0616	0.0617	0.0514	0.0571	0.0471	0.0527	0.0424	0.0584	0.0391	0.0579	0.0476	0.0554
128	0.9	0.0532	0.0535	0.0535	0.0558	0.0548	0.0601	0.0433	0.0555	0.0409	0.0527	0.0468	0.0548
256	0.0	0.0536	0.0536	0.0492	0.0532	0.0472	0.0532	0.0351	0.0536	0.0440	0.0500	0.0480	0.0512
256	0.5	0.0527	0.0529	0.0535	0.0570	0.0484	0.0556	0.0388	0.0505	0.0437	0.0502	0.0510	0.0543
256	0.9	0.0502	0.0508	0.0563	0.0556	0.0511	0.0546	0.0421	0.0527	0.0454	0.0508	0.0499	0.0522
512	0.0	0.0523	0.0532	0.0494	0.0536	0.0458	0.0522	0.0392	0.0515	0.0433	0.0503	0.0490	0.0519
512	0.5	0.0551	0.0555	0.0431	0.0460	0.0474	0.0515	0.0399	0.0538	0.0465	0.0507	0.0537	0.0519
512	0.9	0.0514	0.0505	0.0489	0.0510	0.0483	0.0515	0.041	0.0522	0.0488	0.0540	0.0501	0.0529
1024	0.0	0.0476	0.0465	0.0498	0.0535	0.0516	0.0525	0.0396	0.0517	0.0483	0.0535	0.0486	0.0502
1024	0.5	0.0509	0.0519	0.0525	0.0545	0.047	0.0493	0.0396	0.0516	0.0511	0.0510	0.0531	0.0531
1024	0.9	0.0486	0.0487	0.0506	0.0528	0.0485	0.0485	0.0428	0.0524	0.0496	0.0519	0.0526	0.0571

Table 5: Empirical Power Under Quadratic Loss, Classical and Robust DM Test for one step ahead forecast errors.

Classical and Robust DM test with squared error loss function for one step ahead forecast errors. e_1 and e_2 are bivariate standard normal (BN), t_6, t_3, CN(0.05,25) and CN(0.05,100), with a correlation given by ρ and $\theta = 0.0$. Therefore $L(T) = 0$. The robustness constant c^* is chosen to bound the maximal bias around 0.5% of the nominal 5% size for an assumed contamination of $\varepsilon = 0.01$. Empirical sizes are reported.

		BN		t_6		t_5		$t_{emp}3$		CN(0.05;25)		CN(0.05;100)	
T	ρ	CL	Rob	CL	Rob	CL	Rob	CL	Rob	CL	Rob	CL	Rob
32	0.0	0.1518	0.1572	0.1080	0.1321	0.0976	0.1267	0.0723	0.1092	0.0499	0.1039	0.0360	0.0790
32	0.5	0.1914	0.1976	0.1312	0.1562	0.1148	0.1424	0.0773	0.1138	0.0646	0.1267	0.0410	0.0876
32	0.9	0.5292	0.5414	0.4157	0.4656	0.3855	0.4544	0.2629	0.3628	0.2127	0.3453	0.1310	0.2424
64	0.0	0.2604	0.2614	0.1628	0.1937	0.1474	0.1807	0.0809	0.1363	0.0545	0.1331	0.0528	0.0971
64	0.5	0.3402	0.3428	0.2134	0.2512	0.1474	0.2327	0.1188	0.1824	0.0734	0.1690	0.0643	0.1106
64	0.9	0.8436	0.8454	0.6926	0.7463	0.6422	0.727	0.4307	0.5967	0.2856	0.5102	0.2421	0.3780
128	0.0	0.4732	0.4736	0.2702	0.3193	0.2405	0.3039	0.1234	0.216	0.0798	0.1965	0.0877	0.1287
128	0.5	0.5851	0.5813	0.3643	0.4212	0.3263	0.4003	0.165	0.2949	0.1048	0.2542	0.1191	0.1675
128	0.9	0.9881	0.9878	0.9214	0.9588	0.8819	0.9442	0.6378	0.8747	0.4859	0.7720	0.4867	0.6157
256	0.0	0.7675	0.7600	0.4646	0.5553	0.3982	0.5179	0.1844	0.3736	0.1316	0.3293	0.1447	0.1886
256	0.5	0.8727	0.8669	0.6198	0.7128	0.5386	0.6716	0.2584	0.5098	0.1837	0.4345	0.2076	0.2643
256	0.9	1.0000	1.0000	0.9930	0.9993	0.9823	0.9988	0.8201	0.9917	0.7848	0.9660	0.8206	0.8882
512	0.0	0.9630	0.9589	0.7292	0.8459	0.6356	0.8057	0.2704	0.6291	0.2198	0.5621	0.2630	0.3356
512	0.5	0.9906	0.9892	0.8681	0.9436	0.7868	0.9166	0.3984	0.7942	0.3388	0.7132	0.3903	0.4668
512	0.9	1.0000	1.0000	0.9995	1.0000	0.9964	1.0000	0.8801	1.0000	0.9753	0.9994	0.9889	0.9958
1024	0.0	0.9997	0.9995	0.9335	0.9878	0.8602	0.981	0.4173	0.9067	0.4151	0.8469	0.4785	0.5723
1024	0.5	1.0000	1.0000	0.9810	0.9989	0.9458	0.9977	0.5551	0.9771	0.5994	0.9416	0.6658	0.7448
1024	0.9	1.0000	1.0000	0.9999	1.0000	0.9991	1.0000	0.8627	1.0000	0.9999	1.0000	0.9999	1.0000

5 Appendix

5.1 Asymptotic Power under Distributional Contamination

Let $X_1, ..., X_n$ be a sample of n observations. Assume that X_i is distributed according to some distribution F, and let F_0 be the model distribution. We define a test statistics S_n and assume that

$$\sqrt{n}(S_n - S(F)) \underset{n\to\infty}{\overset{D}{\to}} N(0, V(S, F))$$

We reject $H_0 : S(F) = 0$, if $S_n > k_{n,\alpha_0}$, where a_0 is the nominal level of the test and k_{n,α_0} is given by

$$\alpha_0 = \alpha(F_0) = P_{F_0}(S_n > k_{n,\alpha_0}) = P_{F_0}(\sqrt{n}\frac{S_n - S(F_0)}{\sqrt{V(S, F_0)}} > \sqrt{n}\frac{k_{n,\alpha_0} - S(F_0)}{\sqrt{V(S, F_0)}}).$$

Therefore,

$$\sqrt{n}\frac{k_{n,\alpha_0} - S(F_0)}{\sqrt{V(S, F_0)}} = \Phi^{-1}(1 - \alpha_0) + o(1)$$

and thus

$$k_{n,\alpha_0} = S(F_0) + \frac{1}{\sqrt{n}}\Phi^{-1}(1 - \alpha_0)\sqrt{V(S, F_0)} + o(\frac{1}{\sqrt{n}}),$$

where $\Phi(\cdot)$ si the cumulative distribution of the standard normal distribution.

Now we are interested in the value of the asymptotic power when the underlying distribution deviates slightly from the model distribution. Specifically we define a neighborhood $\widetilde{F}_{n,\Delta,\varepsilon}$ of the model distribution F_0 as in (6); c.f. HRRS (1986) for a discussion of this formalization.

The asymptotic power under $\widetilde{F}_{n,\Delta,\varepsilon}$ can then be obtained by

$$\begin{aligned}
\beta(\widetilde{F}_{n,\Delta,\varepsilon}) &= P_{\widetilde{F}_{n,\Delta,\varepsilon}}(S_n > k_{n,\alpha_0}) \\
&= P_{\widetilde{F}_{n,\Delta,\varepsilon}}\left(\sqrt{n}\frac{S_n - S(\widetilde{F}_{n,\Delta,\varepsilon})}{\sqrt{V(S,\widetilde{F}_{n,\Delta,\varepsilon})}} > \sqrt{n}\frac{k_{n,\alpha_0} - S(\widetilde{F}_{n,\Delta,\varepsilon})}{\sqrt{V(S,\widetilde{F}_{n,\Delta,\varepsilon})}}\right) \\
&= P_{\widetilde{F}_{n,\Delta,\varepsilon}}(\sqrt{n}\frac{S_n - S(\widetilde{F}_{n,\Delta,\varepsilon})}{\sqrt{V(S,\widetilde{F}_{n,\Delta,\varepsilon})}} > \left(\frac{V(F_0)}{V(S,\widetilde{F}_{n,\Delta,\varepsilon})}\right)^{1/2}\Phi^{-1}(1-\alpha_0) \\
&\quad -\sqrt{n}\frac{S(\widetilde{F}_{n,\Delta,\varepsilon}) - S(F_0)}{\sqrt{V(S,\widetilde{F}_{n,\Delta,\varepsilon})}} + o(1)).
\end{aligned}$$

Now we can expand $S(\widetilde{F}_{n,\Delta,\varepsilon})$ around $S(F_0)$ and obtain

$$\sqrt{n}\left(S(\widetilde{F}_{n,\Delta,\varepsilon}) - S(F_0)\right) = \varepsilon\frac{\partial}{\partial\widetilde{\varepsilon}}S((1-\widetilde{\varepsilon})F_0 + \widetilde{\varepsilon}G)\Big|_{\widetilde{\varepsilon}=0} + \Delta\,\frac{\partial}{\partial\widetilde{\Delta}}S(F_{\widetilde{\Delta}})\Big|_{\widetilde{\Delta}=0} + o(\varepsilon) + o(\Delta).$$

By definition

$$\frac{\partial}{\partial\widetilde{\varepsilon}}S((1-\widetilde{\varepsilon})F_0 + \widetilde{\varepsilon}G)\Big|_{\widetilde{\varepsilon}=0} = \int IF(x,S,F_0)dG(x) \tag{12}$$

and with $V(S,\widetilde{F}_{n,\Delta,\varepsilon}) \underset{n\to\infty}{\to} V(S,F_0)$ we obtain

$$\begin{aligned}
\lim_{n\to\infty}\beta(\widetilde{F}_{n,\Delta,\varepsilon}) &= 1 - \Phi\left(\Phi^{-1}(1-\alpha_0) - \Delta\sqrt{E(S,F_0)} - \varepsilon\frac{\int IF(x;S,F_0)dG(x)}{[V(S,F_0)]^{1/2}}\right) \\
&= 1 - \Phi(\Phi^{-1}(1-\alpha_0) - \Delta\sqrt{E(S,F_0)}) \\
&\quad +\varepsilon\phi(\Phi^{-1}(1-\alpha_0) - \Delta\sqrt{E(S,F_0)})\frac{\int IF(x;S,F_0)dG(x)}{[V(S,F_0)]^{1/2}} + o(\varepsilon) \\
&= \beta_0 + \varepsilon\phi(\Phi^{-1}(1-\alpha_0) - \Delta\sqrt{E(S,F_0)})\frac{\int IF(x;S,F_0)dG(x)}{[V(S,F_0)]^{1/2}} + o(\varepsilon),
\end{aligned}$$

where β_0 is the asymptotic power at the model, $\phi(x)$ is the standard normal density, $E(S,F_0) = \frac{\left(\frac{\partial}{\partial\Delta}S(F_{\widetilde{\Delta}})\big|_{\widetilde{\Delta}=0}\right)^2}{V(S,F_0)}$ is the Pitman efficacy of the test.

References

[1] Chao, J.C., V. Corradi, N.R. Swanson, (2001): An out of sample test for granger causality, *Macroeconomic Dynamics* 5,598-620.

[2] Clark, T.E., M.W. McCracken (2001): Tests of equal forecast accuracy and encompassing for nested models, *Journal of Econometrics,* 105, 85-110.

[3] Corradi, V., N.R. Swanson (2002): A consistent test for nonlinear out of sample predictive accuracy, *Journal of Econometrics,* 110, 353-381.

[4] Corradi,V., N.R. Swanson and C. Olivetti (2001): Predictive ability with cointegrated variables, *Journal of Econometrics,* 104,315-358.

[5] Dell'Aquila R., E. Ronchetti and F. Trojani (2003): Robust GMM Estimation of Models of the Short Rate, *Journal of Empirical Finance*, 10, 373-397.

[6] Dell'Aquila R. and E. Ronchetti (2002): Resistant nonparametric estimation of models of the short rate, Working Paper.

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

[8] Hampel F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986): *Robust Statistics: The Approach Based on Influence Functions.* New York: Wiley.

[9] Harvey, D.I., Leybourne, S. J. and P. Newbold (1997): Testing the equality of prediction mean squared errors, *International Journal of Forecasting*, 13, 281-291.

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

[11] Heritier S. and E. Ronchetti (1994): Robust bounded influence tests in general parametric models, *Journal of the American Statistical Association*, Vol. 89, No. 427,897-904.

[12] Kitamura Y. (2001): Predictive inference and the bootstrap, Working Paper.

[13] Knez, P.J., and M.J. Ready (1997): On the Robustness of Size and Book–to–Market in Cross–Sectional Regressions, *Journal of Finance* 52, 1355-1382.

[14] McCracken,M.W., (2000): Robust out of sample inference, *Journal of Econometrics,* 99,195-223.

[15] McCracken, M.W., and K.D. West (2000): Inference about predictive ability, Working Paper.

[16] Randles, R. (1982): On the asymptotic normality of statistics with estimated parameters, *Annals of Statistics*, 10, 462-474.

[17] Sullivan R., A. Timmerman, and H. White (1999): Data-snooping, technical trading rules performance and the bootstrap, *Journal of Finance*, 54, 1647-1692.

[18] West K.D. (1996): Asymptotic inference about predictive ability, *Econometrica*, Vol. 64., Issue 5, 1067-1084.

[19] West K.D. (2001): Ecompassing tests when no model is encompassing, *Journal of Econometrics*, 105, 287-308.

[20] West,K. and M.W. McCracken (1998): Regression based tests of predictive ability, *International Economic Review* 39, 817 840.

[21] White, H. (2000): A Reality Check For Data Snooping, *Econometrica*, 69, 1097-1127.