Instituto Complutense de Análisis Económico

A statistical test for forecast evaluation under a discrete loss function

Francisco Javier Eransus
Universidad Complutense de Madrid
Departamento de Economía Cuantitativa

Alfonso Novales
Universidad Complutense de Madrid
Departamento de Economía Cuantitativa

Abstract

We propose a new approach to evaluating the usefulness of a set of forecasts, based on the use of a discrete loss function de…ned on the space of data and forecasts. Existing procedures for such an evaluation either do not allow for formal testing, or use tests statistics based just on the frequency distribution of (data , forecasts)-pairs. They can easily lead to misleading conclusions in some reasonable situations, because of the way they formalize the underlying null hypothesis that 'the set of forecasts is not useful.'
Even though the ambiguity of the underlying null hypothesis precludes us from performing a standard analysis of the size and power of the tests, we get results suggesting that the proposed DISC test performs better than its competitors.

Keywords: Forecasting Evaluation, Loss Function.

JL Classification C53, C52, C12

UNIVERSIDAD COMPLUTENSE MADRID

Working Paper nº 1424
July, 2014

ISSN: **2341-2356**
WEB DE LA COLECCIÓN: http://www.ucm.es/fundamentos-analisis-economico2/documentos-de-trabajo-del-icae

A statistical test for forecast evaluation under a discrete loss function

Francisco Javier Eransus, Alfonso Novales

July 2014

Abstract

We propose a new approach to evaluating the usefulness of a set of forecasts, based on the use of a discrete loss function defined on the space of data and forecasts. Existing procedures for such an evaluation either do not allow for formal testing, or use tests statistics based just on the frequency distribution of (data , forecasts)-pairs. They can easily lead to misleading conclusions in some reasonable situations, because of the way they formalize the underlying null hypothesis that '*the set of forecasts is not useful*'. Even though the ambiguity of the underlying null hypothesis precludes us from performing a standard analysis of the size and power of the tests, we get results suggesting that the proposed DISC test performs better than its competitors.

Keywords: Forecasting Evaluation, Loss Function.
JEL Classification:

1 Introduction

Formal tests of the null hypothesis that a set of forecasts lack utility to the user of categorical data can be performed by following the proposals in Merton (19XX), Pesaran and Timmerman (19XX), Merton (1981), Henriksson and Merton (1981), and Greer (2003). These tests are based on different comparisons between the observed frequencies on a 2-way classification of data and forecasts and the frequencies that should have been expected under independence of data and forecasts. They are similar in spirit to tests for continuous data that are based on the degree of linear correlation between forecasts and data. REFERENCES.[1]

Working with continuous data, a test based on comparison of frequencies ignores the relevant information contained in the distance between data and forecasts. That way, the definition of forecast usefulness in all the mentioned tests is independent of the forecasting context, with the user not playing any role in specifying that definition. This situation is not too reasonable, since the cost of missing the sign of the data or the implications of making a given forecast error is bound to differ for each particular forecasting application.

The obvious way out of that is to evaluate the quality of a given forecast through the computation of some distance between forecasts and data, like the Mean Squared Error or the Mean Absolute Error. This has two significant limitations: first, that computing a distance measure is not an option when dealing with categorical data, as in the case of prediction of directional change or when dealing with qualitative data, as in most surveys on expectations.

[1] Practical aplications to macroeconomic or financial forecasts of these tests can be found in Schnader and Stekler (1990), Stekler (1994), Leitch and Tanner (1995), Kolb and Stekler (1996), Ash, Smyth an Heravi (1998), Mills and Pepper (1999), Joutz and Stekler (2000), Oller and Bharat (2000), Pons (2000), Greer (2003) among many others.

Second, that a proper loss function should not have the size of the forecast error as its only argument, but rather be a function of both, the data and its associated forecast. That way, we can allow for the asymmetries and nonlinearities that are natural in so many forecasting applications.

A loss function defined on (data,forecast)-pairs allows us to assign a different loss to a same forecast error, depending on whether or not the sign of the data was predicted correctly, as it would be desirable when forecasting the direction of change of a given variable. Furthermore, by placing an upper bound on the value of the loss a discrete function can be easily isolated from the potentially heavy influence of an occasional unusually large forecast error.

Recently, a quite general approach to forecast evaluation has been advanced by Giacomini and White (2000), that considers the use of such loss functions, defined on the space of (data,forecasts). Unfortunately, the Pesaran-Timerman cannot be embedded into the Giacomini-White approach because the alternative hypothesis cannot be evaluated on the basis of the information up to time-t. But, following the same idea as in Giacomini and White, we propose a test (that we label D-test) that extends the frequentist approach of the Pesaran-Timmerman test to exploit the information provided by a loss function defined on a classification of the data space.

With categorical data, there is no general theory describing which loss function to use, but to reach a sensible decision, it seems necessary that the user be able to measure the relative damage produced by each (data,forecast)-combination. By nature, the loss function would then be with this type of data a discrete function defined on the bivariate finite space of (data,forecast)-pairs. Such loss function can be reinterpreted as assigning weights to each cell in the classification table, thereby incorporating the costs associated to each (data,forecast)-pair in a particular setup and hence, solving some of the limitations of frequentist tests we mentioned above. The need to quantify in advance the cost of each (data,forecast)-pair, with the results of the test being conditional on such characterization, should be seen as a strength of our proposal, rather than as a weakness. The alternative of specifying a continuous loss function like the squared forecast error or its absolute value evades this issue by imposing a very tight structure on the loss function without considering whether such structure is really appropriate for the forecasting application in hand, or without assessing how does it condition the result of the forecasting evaluation exercise.

Hence, at a difference of previous tests, our test analyses the univariate frequency distribution of a loss function defined on the space of (data,forecasts) combinations for categorical data, rather than the observed frequencies in the bi-dimensional distribution of data and forecasts themselves. By doing that, we change the null hypothesis from representing that 'the set of forecasts is independent of the realized data' to expressing that 'the observed loss is not larger than the loss that would have been supported if the set of forecasts happened to be independent of realized data'.

A further interest for our proposal arises from its potential application to continuous data, once it is classified using a given partition of the data space. A stock market investor may be primarily interested on whether the stock market will raise or fall over its investment horizon. An options trader would like to forecast correctly the direction of change of the price of the underlying asset, and whether that price movement, if it is in the right direction, compensates for the option price. There are also instances in which macroeconomic forecasts are used in a qualitative manner: will GDP increase or decrease this quarter? Will it change, positively or negatively, by less than 0.5% or by more than that amount? In all these situations, we have a natural collapse of the data sapace to a few intervals of values.

The discrete setup allows for a simple characterization of the asymptotic distribution of our proposed D-test statistic without correcting for possible autocorrelation of the data, as in the Pesaran and Timermann (19XX) test. Further sections are devoted to comparing the

performance of our D-test with that of the PT test for time independent data, but also for cases when when the data display autocorrelation. Such a comparison is interesting, because the probability distribution of the PT-statistic is hard to characterize under autocorrelated data. Our results suggest that the D-test may perform better than the PT test in many interesting and realistic setups. The D-test is more powerful in situations where the set of forecasts are evidently useful. In cases where there may be some doubt about the usefulness of forecasts, the behavior of the D-test is more reasonable, in the sense that the decision reached will depend in a natural way on the specific loss function chosen by the user.

The paper is organized as follows: in Section 2 we discuss the difficulties associated with the CT and PT tests. In Section 3 we introduce our proposal: we explain the general approach, describe discrete loss functions and derive the D-test. In Section 4 we analyze the performance of the D-test relative to the CT and PT tests through simulation exercises. Finally, Section 5 summarizes the main conclusions.

2 Criticism of standard tests

We denote by y_t the time t realization of the variable being forecasted, and by $\widehat{y}_t$ the associated forecast made at $t-h$. We assume that we have a set of T $(y_t, \widehat{y}_t)$-pairs. The CT and PT test divide the data domain in m regions and therefore, the bidimensional domain of data and forecasts is partitioned into m^2-squares. CT is a nonparametric Pearson test with null hypothesis $p_{ij} = p_{i.}p_{.j}$, where p_{ij} is the joint probability that y_t falls in the i-th region while the forecast $\widehat{y}_t$ falls in the j-th region, and $p_{i.}$, $p_{.j}$ denote the marginal probabilities that y_t and $\widehat{y}_t$ fall in the i-th and j-th regions, respectively. The PT test considers the null hypothesis that $\sum_{i=1}^{m} p_{ii} = \sum_{i=1}^{m} p_{i.}p_{.i}$, and uses the natural statistic $\sum_{i=1}^{m} \widehat{p}_{ii} - \sum_{i=1}^{m} \widehat{p}_{i.}\widehat{p}_{.i}$ that substitutes relative sample frequencies for probabilities, which follows a $N(0,1)$ distribution after normalizing by the corresponding standard deviation [see Pesaran and Timmermann (1992)]. PT is usually implemented as a one-side test, taking just the upper-tail of the $N(0,1)$ distribution as the critical region. CT is a non parametric test of stochastic independence, while the PT test is less restrictive, since it just evaluates the quadrants along the main diagonal of the bidimensional table of frequencies. According to Pesaran and Timmermann (1992), the CT test is, in general, more conservative than the PT test.

As mentioned in the Introduction, the use of these tests to evaluate the predictive ability of a single model is not fully appropriate, since a possible rejection of the null hypothesis is not too informative about forecast quality, as we illustrate in this section. Let us assume that we use a partition of the space of data and forecasts into four regions: L-, S-, S+, L+, where the '+', '-' signs indicate the sign of the data and the forecast, while 'L', 'S' denote whether the data or the forecast are large or small in absolute value, this difference defined by a given threshold. Suppose we have a sample of 100 data points on the period-to-period rate of change of a given time series and the associated forecasts obtained from three alternative forecasting models. The hypothetical information on the predictive results has been summarized in the matrices that follow, whose elements represent the absolute frequencies observed in each cell of the joint partition of data and forecasts:[2]

[2] Actually, the test results we mention were obtained after changing the single 25 value that appears in the first row of each matrix by 24, while the (1,3)-element is 1, rather than 0. This was done to avoid the variance of the PT statistic to be zero.

$$M_1 = \begin{array}{cc|cccc} & & & & \widehat{y}_t & \\ & & \text{L-} & \text{S-} & \text{S+} & \text{L+} \\ \hline y_t & \text{L-} & \mathbf{0} & \mathbf{0} & 0 & \textit{25} \\ & \text{S-} & \mathbf{0} & \mathbf{0} & 25 & \textit{0} \\ & \text{S+} & \textit{0} & 25 & \mathbf{0} & \mathbf{0} \\ & \text{L+} & \textit{25} & 0 & \mathbf{0} & \mathbf{0} \end{array}$$

$$M_2 = \begin{array}{cc|cccc} & & & \widehat{y}_t & & \\ & & \text{L-} & \text{S-} & \text{S+} & \text{L+} \\ \hline y_t & \text{L-} & \mathbf{0} & \mathbf{25} & 0 & \textit{0} \\ & \text{S-} & \mathbf{25} & \mathbf{0} & 0 & \textit{0} \\ & \text{S+} & \textit{0} & 0 & \mathbf{0} & \mathbf{25} \\ & \text{L+} & \textit{0} & 0 & \mathbf{25} & \mathbf{0} \end{array} \qquad M_3 = \begin{array}{cc|cccc} & & & & \widehat{y}_t & \\ & & \text{L-} & \text{S-} & \text{S+} & \text{L+} \\ \hline y_t & \text{L-} & \mathbf{25} & \mathbf{0} & 0 & \textit{0} \\ & \text{S-} & \mathbf{0} & \mathbf{25} & 0 & \textit{0} \\ & \text{S+} & \textit{0} & 0 & \mathbf{25} & \mathbf{0} \\ & \text{L+} & \textit{0} & 0 & \mathbf{0} & \mathbf{25} \end{array}$$

In most applications it will be desirable that the forecasts have the right sign and be as precise as close as possiblein size to actual data. In the previous matrices we have indicated in boldface the squares where the forecasts would be correct in sign. In italics we denote those squares where the forecasts would not only have the wrong sign, *but also they would have the least precision possible.*

Forecasts from Model 1 (M1) have always the wrong sign. They also have the least possible precision 50% of the times, those in cells (1,4) and (4,1). Model 2 (M2) always forecasts the sign correctly, but it never reaches the highest precision, as reflected in an empty main diagonal. Model 3 (M3) is always fully right, in sign as well as in magnitude. Clearly, M1 forecasts have no value whatsoever, those from M3 are optimal given the partition of the data space we consider, while those from M2 might be useful, even though they are not fully precise. The reasonable outcome would be that the tests would not reject the lack of utility of M1 forecasts, rejecting it for M3 forecasts. The desired result for M2 forecasts should depend on the specific forecasting context and the specific definition by the user of what is meant by 'useful predictions'.

Unfortunately, the CT test rejects the null hypothesis in the three sets of forecasts.The PT test rejects the null for M3 and it does not reject it for M1 and M2.[3] Furthermore, the numerical value of the CT test statistic is the same for the three models, 292.31, while that of the PT test statistic is the same for M1 and M2, -353.55. Three forecasting situations as different as those in this example are indistinguishable for the CT test, while M1 and M2 are indistinguishable from the point of view of the PT test.

This example illustrates the potential errors made by existing tests of 'lack of usefulness' of a set of forecasts. This is to be expected, since in particular, CT tests for the null hypothesis of independence of data and forecasts, which is violated in the three situations. For M1, the CT test detects some stochastic dependence between data and forecasts, thereby leading to the rejection of the null hypothesis of independence. This undesired outcome results because the test does not pay attention to the type of dependence, which happens to be negative in this case. For M2, the error comes about from the fact that the CT and PT tests do not take into account any characteristic of the forecast error, like its sign or size, whenever the forecast falls in a region different from that of the data. And this is a consequence of both tests using a too general approach to the concept of 'useful forecasts', that it does not consider the possibility that the concept of 'usefulness' might depend on the specific forecasting situation. For instance, a model that produces forecasts that are always right in

[3] We have used PT as a one-sided test, rejecting the null hypothesis when the test statistic test is positive and large enough.

sign but not in magnitude may be very useful in some applications but not so much in some other setups. Not to mention the convenience of taking into account the size of the forecast error. These are some of the limitations we mentioned in the Introduction.

3 A proposal based on a discrete loss function

We now present an approach alternative to those of the CT and PT tests which, incorporating a specific type of loss functions, may greatly alleviate the limitations of these two tests.

3.1 General overview

Let $f(y_t, \widehat{y}_t)$ be a loss function on two arguments, the data and the associated forecast. We propose testing the hypothesis $H_0 \equiv E(f_t) - E(f_t^{IE}) = 0$ ('forecasts are not useful') against $H_1 \equiv E(f_t) - E(f_t^{IE}) < 0$ ('forecasts are useful'), where f_t^{IE} denotes the loss that would arise with a set of forecasts independent from the data. We consider a set of forecasts not to be useful when the mean loss is at least as large as the one that would obtain from f_t^{IE}. Our approach does not pretend to solve completely the ambiguity in the specification of the null hypothesis when testing for predictive ability, but it is more satisfactory than that of the CT and PT tests. On the one hand, the incorporation of a loss function allows us to define a one-sided alternative hypothesis, avoiding potential mistakes as those made by the CT test under the M1 forecasts. On the other hand, the definition of what is meant by a 'not useful' set of forecasts can be made explicit through the choice of a loss function f . Precisely, under a discrete loss function f , it is easy to adjust that definition to each particular application, as we explain below in some examples. Relative to previous tests, the use of a discrete loss function amounts to assigning a different weight to each $(data, forecast)$-pair. That changes in a non trivial way the frequency distribution of $(data, forecast)$-pairs. Besides, the significance of our proposal can be seen in that our test is no longer based on the frequency distribution of the $(data, forecast)$-pairs but rather, on the implied frequency distribution of the loss function.

After partitioning the domain of y_t and $\widehat{y}_t$ in m regions, so that the joint domain of data and forecasts is naturally partitioned in m^2 cells, we define a discrete loss function f by assigning a nonnegative numerical value to each cell. The discrete loss function can be shown as a matrix, as in the following example:

		$\widehat{y}_t$			
		L-	S-	S+	L+
y_t	L-	0	1	2	3
	S-	1	0	2	3
	S+	3	2	0	1
	L+	3	2	1	0

(1)

with L-, S-, S+, L+ being the $(-\infty, -l)$, $(-l, 0)$, $(0, +l)$, $(+l, +\infty)$ intervals, respectively, for a given constant l , conveniently chosen by the user.[4] Le us denote by a the k-vector $(k \leq m^2)$ of possible losses associated to the different cells, i.e., the k-vector of possible values of f, ordered increasingly. In the example , $a = (0, 1, 2, 3)$. We are associating a high loss to incorrectly forecasting the sign, while the magnitude of the forecast error is of secondary importance. A forecast with the right sign always receives in (1) a penalty lower than any forecast with the wrong sign, with independence of the size of the forecast error. This is a

[4] Here we use the same partition for the data and the forecasts, although the analysis can be extended without any difficulty to the case when the two partitions are different.

loss structure that may be reasonable in many applications, although many other alternative choices for f would also be admissible.

A discrete loss function presents significant advantages for the evaluation of macroeconomic and financial forecasts: *i*) by appropriately choosing the number of elements in the partition and the associated penalties, the user can accommodate the choice of f to each specific forecasting context, *ii*) at a difference from most standard loss functions, which are usually a function of just the forecast error, a discrete loss allows for a rich forecast evaluation that it can pay attention to a variety of characteristics of data and forecasts, *iii*) a discrete loss function can take into account the signs as well as the size of both, data and forecast, which allows for a simple incorporation of different types of asymmetries,[5] *iv*) a discrete loss imposes an upper bound on the loss function, thereby reducing or even eliminating the distorting effect of outliers when evaluating the performance of a set of forecasts, *v*) a discrete loss function is a natural choice for the evaluation of forecasts of qualitative variables.

Besides the notional characteristics we described above in favor of discrete loss functions as an interesting choice for forecast evaluation, they also possess two very significant technical advantages, as we are about to see.

3.2 The D-test

Given a discrete loss function f, we propose to test the lack of usefulness of forecasts by testing a null hypothesis: $H_0 \equiv E(f_t) = E(f_t^{IE})$ against an alternative: $H_1 \equiv E(f_t) < E(f_t^{IE})$. The test should be solved using the asymptotic distribution of the difference between the sample means $\overline{f}$ and $\overline{f}^{IE}$, as unbiased estimates of $E(f_t)$ and $E(f_t^{IE})$. The practical implementation of the test faces one difficulty: while the sample information allows us to compute the $\overline{f}$ estimate, we lack sample observations under independence that could be used to compute $\overline{f}^{IE}$. If we knew the true joint probability distribution of data and forecasts, we could construct an unbiased estimator $\overline{f}^{IE}$ by substituting sample estimates in the analytical expression of $E(f_t^{IE})$ as a function of data and forecasts, but unfortunately, that probability distribution is usually unknown.

Under a discrete loss f , this difficulty can be easily solved: we have $E(f_t) = ap$ and $E(f_t^{IE}) = ap^{IE}$, with $p(r)$ and $p^{IE}(r)$ denoting the probability that the loss function takes the value $a(r)$ under the true distribution and under independence, respectively. Given the discrete nature of f , these probabilities are defined as: $p(r) = \sum_{(i,j)\in C(r)} p_{ij}$ and $p^{IE}(r) = \sum_{(i,j)\in C(r)} p_{ij}^{IE}$, with p_{ij}^{IE} being the probability that data and forecasts fall in the (i,j)-cell if they are stochastically independent, while $C(r)$ represents the set of all quadrants where f takes the value $a(r)$. By definition of stochastic independence, we have $p_{ij}^{IE} = p_{i.}p_{.j}$ and we can easily get the expression for $E(f_t^{IE})$. Its estimator $\overline{f}^{IE}$ is defined substituting in that expression the estimator $\widehat{p}_{ij}^{IE} = \widehat{p}_{i.}\widehat{p}_{.j}$ for p_{ij}^{IE}.

We can now proceed to describing our test proposal. Let $P = (p_{11}, p_{12}, ..., p_{1m}, p_{21}, ..., p_{2m}, ..., p_{m1}, ..., p_{mm})'$ be the m^2-column vector that contains the theoretical probabilities p_{ij} for the quadrants associated to the partition of the space of data and forecasts, and $\widehat{P}$ its maximum likelihood estimator, based on relative frequencies, *assuming the sample observations are independent from each other*. Using *standard results*, we have $\sqrt{T}(\widehat{P} - P) \xrightarrow{L} N(0, V_P)$, with $V_P = \Omega - PP'$ and Ω a diagonal $m^2 \times m^2$ matrix with the elements of P along the diagonal. Let us know consider the differentiable function $\varphi(.)$ of R^{m^2} on R^k defined

[5] Which are so natural in Economics. For instance, it is hard to believe that an investor will regret getting a return much higher than it was predicted when a financial asset was bought.

by: $\varphi(P) = p - p^{IE} = \left(\sum_{C(1)} p_{ij} - p_{i.}p_{.j}, ..., \sum_{C(k)} p_{ij} - p_{i.}p_{.j} \right)'$. Putting together both results we have: $\sqrt{T}\left[(\widehat{p} - \widehat{p}^{IE}) - (p - p^{IE})\right] \xrightarrow{L} N(0, \nabla\varphi(P)V_P\nabla\varphi(P)')$, with $\nabla\varphi(P)$ being the $k \times m^2$ Jacobian matrix for the vector function $\varphi(P)$. Finally, multiplying by a we have the asymptotic distribution of $\overline{f} - \overline{f}^{IE}$ under the null hypothesis $E(f_t) = E(f_t^{IE})$: $\sqrt{T}(\overline{f} - \overline{f}^{IE}) = a(\widehat{p} - \widehat{p}^{IE}) \xrightarrow{L} N(0, G_p)$, with $G_p = a\nabla\varphi(P)V_P\nabla\varphi(P)'a'$.

We use the consistent estimator $\widehat{G}_p$, which allows us to maintain the same limiting distribution. Therefore, the proposed D-test for the null H_0 against H_1 is:

$$D = \sqrt{T}\widehat{G}_p^{-1/2}a(\widehat{p} - \widehat{p}^{IE}) \underset{H_0}{\xrightarrow{L}} N(0,1), \tag{2}$$

with $\widehat{G}_p = a\left[\nabla\varphi(P)\right]_{P=\widehat{P}} \widehat{V}_P \left[\nabla\varphi(P)\right]'_{P=\widehat{P}} a'$ and $\widehat{V}_P = V_P|_{P=\widehat{P}}$. The expression for matrix $\nabla\varphi(P)$ is given in Appendix A. The critical region for the test corresponds to the lower tail of the $N(0,1)$ distribution.

Obviously, the test will be invariant to application of a scale factor λ on the loss function. It is not difficult to show that the one-sided version of the PT test, where the critical region is just the upper-tail of the distribution, is a special case of the D-test when there are only two values in a, one of them being the penalty assigned to every cell along the main diagonal in the loss matrix, and another one for the rest of the cells, the first value being smaller than the second one.

4 Empirical propertis of the D-test

4.1 The D-test under alternative definitions of 'usefulness'

The loss function is central to our analysis, so it is far from surprising that its choice might condition the results of the test. In a given application, it should be expected that the loss function will have been previously chosen, and it should be treated as fixed. But it is interesting to know how decision makers with different loss functions would decide if they faced the same set of data and forecasts. On the other hand, if the decision maker is not sure about his preferences, it is important to have an indication of the sensitivity of the decision to specific changes in the loss function. In particular, it will always be relevant to figure out which aspects of the specification of the loss function condition the decision and which ones are not relevant to reach a conclusion on the usefulness of the set of forecasts.

To illustrate the sensitivity of the D-test to alternative loss functions, we consider again the three hypothetical forecasting situations described in Section 2. We start by implementing the D-test for the three matrices M1, M2 and M3, under two alternative discrete loss functions, the one defined by (1), and an alternative one characterized by:

		$\widehat{y}_t$			
		L-	S-	S+	L+
y_t	L-	0	1.75	2	3
	S-	1.75	0	2	3
	S+	3	2	0	1.75
	L+	3	2	1.75	0

(3)

The difference between (1) and (3) reduces to the loss associated to forecasts that have the right sign but the wrong magnitude, which is equal to 1 in (1) while being 1.75 in (3). The loss function (1) assigns a relatively high value, i.e., a low loss, to predicting the sign

correctly, even if the absolute size is wrong, as it would be reasonable from the point of view of an investor in financial markets. Under this loss function, the main condition for a forecast to be useful is that it has the right sign, although it is of course preferred that it is also correct in size. For the loss function (3) the difference between the losses associated to incorrect predictions with either sign is small. A forecast is not useful if it misses both the sign and the size of actual data, *as it might be the case in macroeconomic forecasting.*

Figure 1. Test results for models M1, M2 and M3

Losses	Function **(1)**			Function **(3)**		
	Observed value for the test statistic					
	CT	PT	D-test	CT	PT	D-test
M1	292.31	-353.55	40.62	292.31	-353.55	33.46
M2	292.31	-353.55	-17.60	292.31	-353.55	2.56
M3	292.31	74.94	-43.77	292.31	74.94	-48.95
	p-value[6]					
	CT	PT	D-test	CT	PT	D-test
M1	0.0	1.0	1.0	0.0	1.0	1.0
M2	0.0	1.0	**0.0**	0.0	1.0	**0.99**
M3	0.0	0.0	0.0	0.0	0.0	0.0

Figure 1 displays the results of applying the D-test as well as the CT and PT tests to the three hypothetical forecasting models. In the three tests, the null hypothesis is that the set of forecasts is not useful. At a difference from the CT and PT tests, the D-test statistic takes different values for the M1 and M2 forecasts. For M3 forecasts, the three tests correctly reject the null hypothesis. For M1, CT also rejects the null, incorrectly, while PT and the D-test do not reject it. Finally, and this is the most relevant case, the CT and PT tests lead to the same decision on the lack of utility of M2 forecasts with independence of the forecasting context in which the observed frequencies were obtained. This is reflected in the fact that the test statistic takes the same value for both tests under the two alternative loss functions. This outcome is unreasonable, as already discussed in Section 2. On the other hand, the D-test rejects the null hypothesis in favor of considering that the predictions are useful if the loss function is (1), while considering the set of forecasts not to be useful if the loss function is given by (3). We consider that to be a reasonable behavior for a forecast evaluation test in a situation like that summarized by M2: to reject the hypothesis of lack of utility of the forecasts if and only if correctly forecasting the sign is relevant enough.

As we see in this application, the D-test solves some of the limitations of the CT and PT tests pointed out in Section 2 and in the Introduction. The D-test does not make unacceptable mistakes like rejecting the null hypothesis of lack of utility of forecasts when they are negatively correlated with the data. Even more important is the fact that the test decision will depend on the definition of 'useful forecasts' made by the user in each specific application through the choice of loss function, *and the D-test can naturally accommodate any level of desired detail in that definition.*

To analyze the sensitivity of the D-test to the definition of 'usefulness' we now characterize how D-test decisions depend on the numerical values chosen for the loss function in the matrix of frequencies M2, the only one in there is such a dependence. To consider a wide array of possibilities, we use the pattern defined by the loss function (4) under the constraint $\lambda_3 > \lambda_2 > \lambda_1 > 0$. This way, we penalize more heavily a forecast that has the wrong sign

[6] The critical region for tests CT and DISC is the upper and lower tail, respectively, of their corresponding test distributions. The critical region for the PT includes both tails.

than one with the right sign, *independently of the size of the error in each case,* although when comparing forecasts with the wrong sign, the size of the forecast error also matters. Furthermore, we guarantee some symmetry in the loss function, with the same numerical loss for quadrants like (L-, S+) and (L+, S-). This is the pattern incorporated in (1) and (3).

$$
\begin{array}{cc|cccc}
 & & \multicolumn{4}{c}{\widehat{y}_t} \\
 & & \text{L-} & \text{S-} & \text{S+} & \text{L+} \\ \hline
 & \text{L-} & 0 & \lambda_1 & \lambda_2 & \lambda_3 \\
y_t & \text{S-} & \lambda_1 & 0 & \lambda_2 & \lambda_3 \\
 & \text{S+} & \lambda_3 & \lambda_2 & 0 & \lambda_1 \\
 & \text{L+} & \lambda_3 & \lambda_2 & \lambda_1 & 0
\end{array}
\tag{4}
$$

Remember that what is distinctive about M2 forecasts is that they always have the right sign, but they are never precise in size. We would expect forecasts to be useful and hence, to reject the null hypothesis, *i*) when the loss associated to having the right sign is small, relative to that of forecasts with the wrong sign and size. Another case in which XXXXX . Finally, *iii*) if the loss values associated to any combination of data and forecast is relatively similar to each other, the test will not reject the null hypothesis, since the performance of the model is almost irrelevant in terms of the loss function chosen by the user. We should also ebar in mind that the CT test would reject the lack of usefulness hypothesis for any combination of values of the λ-parameters, the opposite being true for the PT test.

a) As a first exercise, let us consider values: $\lambda_1 = 1$, $\lambda_2 = 2$, while allowing the value of λ_3 to change over the interval $(2, 3)$.The second exercise is similar, with fixed values: $\lambda_1 = 1$, $\lambda_3 = 3$ while λ_2 takes values in the interval $(1, 3)$. In both exercises, the p-value associated to the D-test is 0.0 for any value of the floating parameter, so that the test would always reject the null hypothesis that the forecasts from M2 are not useful.

b) We might be tempted to conclude from the previous exercise that the D-test will always reject the null hypothesis that the forecasts from M2 are not useful under any loss matrix (4) verifying $\lambda_3 > \lambda_2 > \lambda_1 = 1$, but that is not the case. Under that constraint, if λ_2 and λ_3 are both close enough to λ_1, then the D-test might not reject the null hypothesis, because the penalty associated to forecasts with the right sign but the wrong size (λ_1) is not too different from the loss associated to any forecast that has the wrong sign (λ_2 or λ_3). If we maintain $\lambda_1 = 1$ and let λ_2 and λ_3 vary inside the intervals $(1, 2)$ and $(2, 3)$, respectively, the p-value is above 0.05 in some cases, leading to not rejecting the lack of usefulness of forecasts. This is the case, for instance, when $\lambda_2 = 1.05$ and $\lambda_3 < 2.15$, or when $\lambda_2 = 1.10$ and $\lambda_3 < 2.05$. So, if the loss associated to missing sign but with a relatively small error in size is small, relative to having the right sign, then we need a large penalty associated to forecasts missing sign and size to reject the null hypothesis.

The two previous points show that a clear distinction between the losses associated to forecasts with the wrong sign and some of the forecasts with the right sign is needed for the D-test to conclude in favor of the usefulness of the M2 forecasts, which seems a desirable condition.

c) The λ_1 parameter defines the only loss made by model M2 and hence, it is the most decisive to understand the behavior of the test. So, we now maintain $\lambda_2 = 2$ and $\lambda_3 = 3$, while λ_1 takes values in the interval $(0, 2)$, with the results shown in Figure 2a. The p-value is 0.0 for values of λ_1 up to $\lambda_1 = 1.59$, rapidly increasing to reach 1.0 for $\lambda_1 = 1.76$. This is consistent with the result obtained using (3) as loss function, and emphasizes again that the M2 forecasts will be seen as useful only if there is a substantial difference in value between forecasts with the wrong sign and those with the right sign.

d) Finally, to complete the analysis in *c*), we perform another exercise letting λ_1 and λ_2 vary inside the intervals $(0, 2)$ and $(2, 3)$, respectively, while $\lambda_3 = 3$. The D-test rejects the

null hypothesis whenever $\lambda_1 < 1.6$, in coherence with the results obtained in *c*). If $\lambda_1 > 1.6$, then the decision of the D-test for M2 will depend on the value of λ_2. Once again, the closer are λ_1 and λ_2 to each other, the less relevant will be forecasting the right sign and hence, the more likely will be not to reject the null hypothesis. In Figure 2b we show the p-values for some values of λ_1 and λ_2. Specifically, we draw the curves of p-values as λ_2 changes for some fixed values of λ_1, all of them above 1.6. On the other hand, in Figure 3 we present the combinations (λ_1, λ_2) for which we obtained a p-value equal to 0.01, which allows us to gain some intuition as to the level of the λ_1/λ_2 ratio below which the D-test will reject the null hypothesis of lack of usefulness of the M2 forecasts under a loss matrix (4) with $\lambda_3 = 3$.

Figure 2. p-values of the D-test for example M2, under a loss (4)

Figure 2.a. $\lambda_2 = 2$, $\lambda_3 = 3$

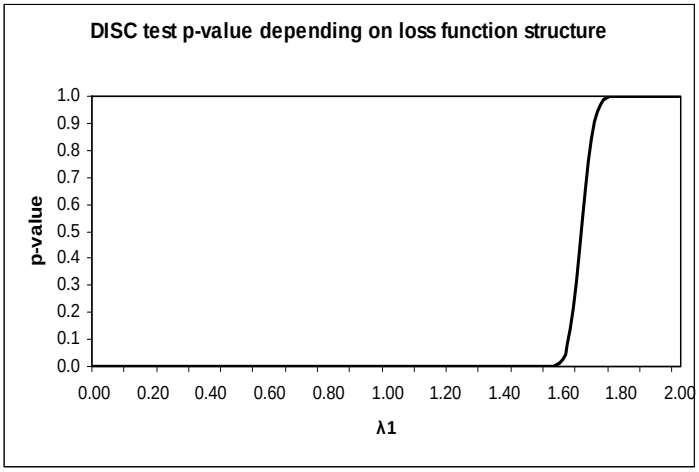

Figure 2.b. $\lambda_3 = 3$

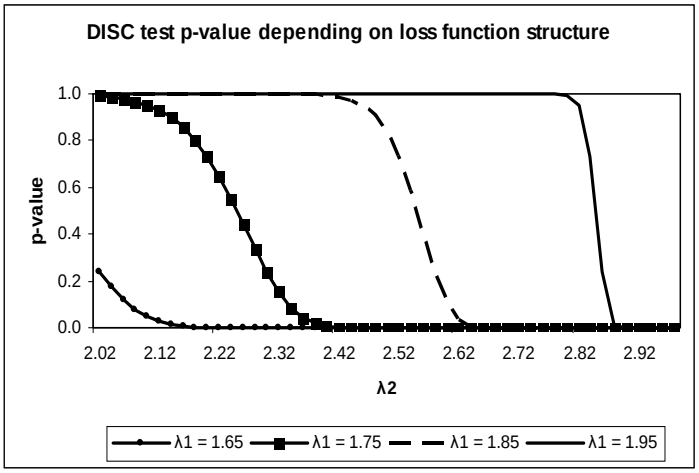

Figure 3. (λ_1, λ_2) combinations such that DISC p-value = 0.01 for M2 forecasts under (4), with $\lambda_3 = 3$.

	λ_1			
	1.65	1.75	1.85	1.95
λ_2	2.14	2.40	2.64	2.88
λ_1/λ_2	0.77	0.73	0.70	0.68

For each value of λ_1, the p-value of the D-test for M2 will be below 0.01 whenever λ_2 is higher than the value associated to λ_1 .

4.2 Simulation results

4.2.1 Experimental design

To obtain further evidence on the different behavior of the D-test and the CT and PT tests, we now perform a simulation experiment. We will sample T $(y_t, \widehat{y}_t)$ pairs from a Bivariate Normal distribution with zero means, correlation coefficient ρ and unit variances, and apply the three tests. The first variable will be considered as the data and the second variable as the forecasts. We will use values: $\rho = 0$, 0.4, 0.75, 0.9 and $T = 10$, 25, 50.[7] The numerical value of the linear correlation coefficient ρ between data and forecasts will allow us to control if the set of forecasts are useful. The test will employ a 4×4 partition of the R^2-space, based on intervals $(-\infty, -l)$, $(-l, 0)$, $(0, +l)$, $(+l, +\infty)$, with $l = 0.8416$, and loss function (1). Under $l = 0.8416$, the marginal probabilities that the data fall in each of the intervals are 0.20, 0.30, 0.30 and 0.20, respectively, and the same applies to the forecasts, which looks reasonable. We repeated the simulation exercises with $l = 0.5244$, which implies interval probabilities of 0.30, 0.20, 0.20 and 0.30, obtaining the same qualitative results.

We can compute the probability associated to each cell in the loss matrix under a joint Normal density with correlation ρ, as well as under independence.That allows us to compute the expected loss in both cases. Given a significance level α,there will be a correlation level ρ^* above which data and forecasts are significantly correlated and hence, the set of forecasts should be considered to be useful. We want to compare the different requirements for usefulness of each testing approach, in terms of the level of correlation between data and forecasts.

Under joint Normality, with $\rho = 0.4$,we have the probability distribution:

$$\begin{pmatrix} & (-\infty, -0.8416) & (-0.8416, 0) & (0, 0.8416) & (0.8416, +\infty) \\ (-\infty, -0.8416) & 0.0762 & 0.0688 & 0.0417 & 0.0133 \\ (-0.8416, 0) & 0.0688 & 0.1017 & 0.0878 & 0.0417 \\ (0, 0.8416) & 0.0417 & 0.0878 & 0.1017 & 0.0688 \\ (0.8416, +\infty) & 0.0133 & 0.0417 & 0.0688 & 0.0762 \end{pmatrix}$$

while for $\rho = 0.9$ we would have:

$$\begin{pmatrix} & (-\infty, -0.8416) & (-0.8416, 0) & (0, 0.8416) & (0.8416, +\infty) \\ (-\infty, -0.8416) & 0.1499 & 0.0481 & 0.0019 & 0.0000 \\ (-0.8416, 0) & 0.0481 & 0.1820 & 0.0679 & 0.0019 \\ (0, 0.8416) & 0.0019 & 0.0679 & 0.1820 & 0.0481 \\ (0.8416, +\infty) & 0.0000 & 0.0019 & 0.0481 & 0.1499 \end{pmatrix}$$

while under independence, we have the probability distribution:

[7] We restrict our attention to samples of length $T \leq 50$, since the case $T > 50$ does not usually arise in practice.

$$\begin{pmatrix} & (-\infty,-0.8416) & (-0.8416,0) & (0,0.8416) & (0.8416,+\infty) \\ (-\infty,-0.8416) & 0.04 & 0.06 & 0.06 & 0.04 \\ (-0.8416,0) & 0.06 & 0.09 & 0.09 & 0.06 \\ (0,0.8416) & 0.06 & 0.09 & 0.09 & 0.06 \\ (0.8416,+\infty) & 0.04 & 0.06 & 0.06 & 0.04 \end{pmatrix}$$

This analysis is opposite to the one we carried out at the end of the previous section. There, we kept the matrix of frequencies M2 fixed, i.e., there was only one set of forecasts set, while we were changing the definition of the discrete loss function. By contrast, we now vary the set of forecasts while maintaining always the same discrete loss function.

Before proceeding, it is crucial to understand that in our framework we cannot perform a standard analysis of size and power. The null hypothesis for the tests is that the set of forecasts lacks usefulness. Therefore, even though each test defines that hypothesis in a specific manner, the null hypothesis is ambiguous in nature, and it will usually not be possible to know before hand whether it is true or false, in spite of the fact that we are running a simulation experiment.

We proceed as follows:

a) if $\rho = 0$, it is clear that the set of forecasts is not useful and the null hypothesis should not be rejected. This is a standard size exercise.

b) if ρ is high enough, the forecasts should be considered useful and the null hypothesis should be rejected, and we can analyze the power of the tests in a standard sense. In our experiment, this will apply to the case $\rho = 0.9$.

c) if ρ takes an intermediate value in our experimental design, like $\rho = 0.4$, as it might be expected in practical applications, we cannot conclusively say whether forecasts are useful. We will then study each sample realization and check whether the decision taken by each test looks 'reasonable' given the partition and loss function that have been defined.

The case when $\rho = 0.75$ can be interpreted as either meaningn that data and forecasts are highly correlated, in which case the D-test should reject the null hypothesis, or as a intermediate case, with the D-test potentially leading to either rejecting or not rejecting the lack of usefulness of forecasts.

4.2.2 Results

Table 1 presents the rejection probabilities for the three tests. We can compare the performance in size of the three tests (when $\rho = 0$) and their performance in terms of power (when $\rho = 0.9$). All tests are reasonably unbiased in size (except for the CT test when $T = 10$), DISC being the test with the highest power for small sample sizes. If we take the view that $\rho = 0.75$ must be interpreted as an exercise in power, i.e., that forecasts that have correlation of $\rho = 0.75$ with the data should be seen as useful, then the results in Table 1 are even more evident in favor of the D-test being more powerful than the CT and PT alternatives.

Table 1. Rejection probabilities (%). $\alpha = 5\%$.

$(y_t, \widehat{y}_t) \sim N(0_{2\times 1}, \Sigma)$, $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$

ρ	T	CT 4×4	PT	DISC
0.00	10	1.7	6.1	8.4
	25	4.1	4.5	6.5
	50	4.4	4.2	5.3
0.40	10	3.1	17.6	29.4
	25	13.5	25.3	45.7
	50	33.3	40.0	69.2
0.75	10	11.5	48.2	68.2
	25	71.2	78.5	95.0
	50	99.0	96.8	99.8
0.90	10	32.8	80.0	89.7
	25	97.9	98.8	99.9
	50	100.0	100.0	100.0

Number of realizations: 5000.

In situations with an intermediate degree of correlation between data and forecasts, the analysis of size and power does not apply. For $\rho = 0.4$, Table 1 shows again that the D-test rejects the null hypothesis of lack of utility of the set of forecasts more often than CT and PT. But, since we do not know a priori whether or not the set of forecasts is then useful, such a result is hard to interpret. Because of that, we have analyzed each of the 5000 samples produced under this design, with the intention of checking how reasonable were the decisions made by each test. We have paid attention to those simulations in which the CT and PT tests take the same decision, while the D-test takes the opposite decision. The percentage of simulations when this circumstance arises for the $\rho = 0.4$ and $\rho = 0.75$ designs is given in Table 2. We can see that it is very unlikely that the D-test will not reject the null hypothesis of lack of utility of the set of forecasts whenever the CT and PT tests reject it. So, the discrepancy between the tests arises in the opposite situation. Under the loss function (1), it is relatively frequent that the D-test rejects the null hypothesis at the same time that the CT and PT tests do not reject it. We will call these 'type-R simulations'. As shown in Table 2, such a probability falls between 10% and 23%, except in the case $\rho = 0.75$, $T = 50$, when the three tests reject the null hypothesis in almost 100% of the samples.

Table 2. Estimated probability (%)
that the D-test will take a decision
contrary to the CT and PT tests

ρ	T	CT and PT: NR DISC: R	CT and PT: R DISC: NR
0.40	10	14.6	0.2
	25	20.3	0.4
	50	22.8	0.4
0.75	10	21.3	0.2
	25	10.1	0.1
	50	0.3	0.0

R and NR denote rejection and not rejection of H_0, respectively.

Table 3 summarizes the results from type-R simulations. For each (ρ, T)-pair we select three specific type-R simulations: those corresponding to the maximum, minimum and median value of $1 - v$, where v refers to the p-value for the D-test. We will denote those simulations by *smax, smin* and *smedian*, respectively. We could interpret this choice as selecting the cases when the discrepancy between DISC and the other two tests was largest (maximum $1 - v$), lowest (minimum $1 - v$) and an intermediate case (median $1 - v$).[8] For each of these three simulations, we present in Table 3 the sample relative frequencies for the four possible loss values according to (1).

The conclusion that the D-test made the right decision is less controversial for small samples. For instance, if $T = 10$, we have simulations like *smax*, where the forecasts have always been correct in sign. Furthermore, when $T = 10$, forecasts have also been correct in size at least 40% of the times for both $\rho = 0.40$ as well as for $\rho = 0.75$. Yet, CT and PT will not reject the lack of utility of the forecasts, while DISC does reject it. Using *smedian* and *smin* simulations the situation is less extreme, but with 80% of forecasts having the right sign, it should be easy to argue that the D-test still leads to the right decision in these simulations. The situation gradually becomes less clear as T increases since CT and PT work better then. But if we revise each one of the simulations in Table 3, it is hard to detect a case in which the D-test makes a decision that we could consider unreasonable. The more arguable case might be the *smin* simulation for $\rho = 0.4$, $T = 50$. There, the forecasts had the right sign 56% of the times, but only 36% of the times were also right in size, while among the 44% of forecasts that missed the sign, only 12% forecasts had the wrong sign and the wrong size. But, even in this case, the decision made by DISC to reject that the forecasts are not useful looks acceptable, according to the loss function (1).

[8] Other criterions lead to similar conclusions. That would be the case if we used the difference between the mean of the p-values obtained for the CT and PT test and v, or if we used the numerical value of $\overline{f}^{IE} - \overline{f}$.

Table 3. Detailed information on representative type-R simulations

ρ	T		$\overline{f}$	$\overline{f}^{IE}$	$\widehat{p}(0)$	$\widehat{p}(1)$	$\widehat{p}(2)$	$\widehat{p}(3)$
0.40	10	*smax*	0.60	1.53	0.40	0.60	0.00	0.00
		smedian	0.80	1.47	0.50	0.30	0.10	0.10
		smin	0.70	1.28	0.50	0.30	0.20	0.00
	25	*smax*	0.84	1.48	0.40	0.40	0.16	0.04
		smedian	0.96	1.36	0.44	0.28	0.16	0.12
		smin	1.04	1.35	0.36	0.32	0.24	0.08
	50	*smax*	0.96	1.41	0.34	0.40	0.22	0.04
		smedian	1.10	1.41	0.36	0.28	0.26	0.10
		smin	1.20	1.46	0.36	0.20	0.32	0.12
0.75	10	*smax*	0.50	1.34	0.50	0.50	0.00	0.00
		smedian	0.70	1.30	0.50	0.30	0.20	0.00
		smin	1.00	1.56	0.30	0.50	0.10	0.10
	25	*smax*	0.84	1.50	0.36	0.44	0.20	0.00
		smedian	0.96	1.43	0.40	0.32	0.20	0.08
		smin	1.00	1.32	0.32	0.36	0.32	0.00
	50	*smax*	0.92	1.37	0.40	0.30	0.28	0.02
		smedian	0.96	1.30	0.42	0.24	0.30	0.04
		smin	1.08	1.35	0.36	0.30	0.24	0.10

$\widehat{p}(i)$: sample relative frequency for the event $f_t = i$.

To summarize the analysis in this section, we can say that the D-test is more powerful than the CT and PT tests in those situations in which the set of forecasts is clearly useful (high values of ρ). In cases when it is unclear a priori whether or not the set of forecasts is useful (intermediate values of ρ), we can at least say that the D-test always behaves reasonably, according to the definition of the chosen loss function. In such situation, we must see the decisions reached by the CT and PT tests as arbitrary, since it would be unclear, by looking just at the relative frequencies of forecasts with the right or wrong sign and size, that the set of forecasts is useful. By contrast, by paying attention at the information provided by each data point and the associated forecast, the D-test makes better decisions.

5 Conclusions

We have analyzed three non parametric tests to evaluate the quality of a set of point forecasts, which can be used even if we ignore the probability distributions of data and forecasts. Two of them are standard in the literature, the Contingency Table test (CT) and the Pesaran and Timmermann test (1992) (PT), while we have introduced the D-test. We have shown how the CT test can easily make unacceptable mistakes even in situations where the forecasts are obviously not useful. Furthermore, given a set of numerical forecasts and data, the conclusion of the CT and PT tests is independent of the particular application in which the data and forecasts have been generated, with a suboptimal performance in many forecasting contexts.

The problem arises because both tests focus just on the independence or lack thereof between data and forecasts, an approach which precludes a finer evaluation of each $(data, forecast)$-

pair and essentially leads to a rigid definition of what we understand by useful forecasts. They are also based on the joint sample frequency distribution of data and forecasts, without fully exploiting the information in their numerical values. On the contrary, the D-test is based on a discrete loss function that characterizes what is meant by useful forecasts in each specific application, solving the mentioned limitations of alternative tests like the CT and PT tests. Discrete loss functions are interesting in many practical situations, as it is the case when a correctly signed forecast is particularly important or when forecasting qualitative data. Discrete loss functions are very flexible, since they do not need to have the forecast error as their only argument. That way, it is very easy to accommodate any type of asymmetry in the valuation of forecast errors, which permits a richer evaluation of forecasts. Besides, the discrete nature of the function allows us to obtain the probability distribution for the D-test statistic, which could not be found under general continuous loss functions.

Our results suggest that the D-test performs better than the two standard tests: it does not make unacceptable mistakes like those occasionally made by CT, and it seems to be more powerful in situations when the set of forecasts is clearly useful. In experimental designs when there is ambiguity about the utility of the set of forecasts, the behavior of the D-test is at least reasonable, according to the utility criterion that the user may have established through the numerical specification of the discrete loss function.

A Appendix: The expression for $\nabla\varphi(P)$

Remember that the function $\varphi(P)$ is

$\varphi(P) = \left(\sum_{C(1)} p_{uv} - p_{u.}p_{.v}, ..., \sum_{C(k)} p_{uv} - p_{u.}p_{.v} \right)'$ and that $C(r)$ is the set of (u,v)-quadrants where f takes the value a_r. We want to obtain the expression for the $k \times m^2$ matrix $\nabla\varphi(P) = \begin{pmatrix} \frac{\partial\varphi_1}{\partial p_{11}} & \frac{\partial\varphi_1}{\partial p_{12}} & ... & \frac{\partial\varphi_1}{\partial p_{mm}} \\ \frac{\partial\varphi_2}{\partial p_{11}} & \frac{\partial\varphi_2}{\partial p_{12}} & ... & \frac{\partial\varphi_2}{\partial p_{mm}} \\ ... & ... & ... & ... \\ \frac{\partial\varphi_k}{\partial p_{11}} & \frac{\partial\varphi_k}{\partial p_{12}} & ... & \frac{\partial\varphi_k}{\partial p_{mm}} \end{pmatrix}$, where φ_r is the r-th element of $\varphi(P)$, i.e., $\varphi_r = \sum_{C(r)} p_{uv} - p_{u.}p_{.v}$.

Before giving the general expression for $\frac{\partial\varphi_r}{\partial p_{ij}}$, let us work with a particular example which may help the reader to understand the ongoing general expression easier:

Consider the 4×4 loss function:

		$\widehat{y}_t$			
		r_1	r_2	r_3	r_4
y_t	r_1	0	1	2	3
	r_2	1	0	2	3
	r_3	3	2	0	1
	r_4	3	2	1	0

and let us see how to calculate the derivative $\frac{\partial\varphi_2}{\partial p_{31}}$. The function is $\varphi_2 = \sum_{C(2)} p_{uv} - p_{u.}p_{.v}$. The set $C(2)$ consists of those quadrants with a loss $a_2 = 1$, i.e., $C(2) = \{(1,2),(2,1),(3,4),(4,3)\}$. Therefore φ_2 takes the expression $\varphi_2 = (p_{12} - p_{1.}p_{.2}) + (p_{21} - p_{2.}p_{.1}) + (p_{34} - p_{3.}p_{.4}) + (p_{43} - p_{4.}p_{.3})$.

We should find those terms that include the parameter p_{31}. As the marginal probabilities $p_{i.}$ and $p_{.j}$ are the sum of the i-th row and j-th column probabilities, respectively, the parameter p_{31} implicitly appears in $p_{3.}$ and $p_{.1}$. As p_{31} appears in $p_{3.}$ and $p_{.1}$, the derivative $\frac{\partial\varphi_2}{\partial p_{31}}$ is $\frac{\partial\varphi_2}{\partial p_{31}} = d_{31}^{(2)} = -p_{2.} - p_{.4}$. Had the $(3,1)$-quadrant also been included in the set $C(2)$,

p_{31} would have been the first element of another term into brackets, and the derivative would have been $1 - d_{31}^{(2)}$.

We are now prepared to understand the general expression for $\frac{\partial \varphi_r}{\partial p_{ij}}$:

$\frac{\partial \varphi_r}{\partial p_{ij}} = d_{ij}^{(r)} = -\left(\sum_{(i,v) \in C(r)} p_{.v} + \sum_{(u,j) \in C(r)} p_{u.} \right)$, if the (i,j)-quadrant is not included in $C(r)$, and $\frac{\partial \varphi_r}{\partial p_{ij}} = 1 - d_{ij}^{(r)}$, otherwise.

References

[1] Ash, J.C.K., Smyth, D.J and Heravi, S.M. (1998). Are OECD Forecasts Rational and Useful?: a Directional Analysis, *International Journal of Forecasting* 14, 381-391.

[2] Greer, M. (2003). Directional Accuracy Tests of Long-Term Interest Rate Forecasts, *International Journal of Forecasting* 19, 291-298.

[3] Henriksson, R.D., Merton and R.C. (1981). On Market Timing and Investment Performance. II: statistical procedures for evaluating forecasting skills, *Journal of Business* 54, 513-533.

[4] Joutz, F. and Stekler, H.O. (2000). An Evaluation of the Predictions of the Federal Reserve, *International Journal of Forecasting* 16, 17-38.

[5] Kolb, R.A. and Stekler, H.O. (1996). How well do Analysts Forecast Interest Rates?, *Journal of Forecasting* 15, 385-394.

[6] Leitch, G. and Tanner, J.E. (1995). Professional Economic Forecasts: Are they Worth their Costs?, *Journal of Forecasting* 14, 143-157.

[7] Merton, R.C. (1981). On Market Timing and Investment Performance. I: an Equilibrium Theory of Value for Market Forecasts, *Journal of Business* 54, 363-406.

[8] Mills, T.C. and Pepper, G.T. (1999). Assessing the Forecasters: an Analysis of the Forecasting Records of the Treasury, the London Business School and the National Institute, *International Journal of Forecasting* 15, 247-257.

[9] Oller, L. and Bharat, B. (2000). The Accuracy of European Growth and Inflation Forecasts, *International Journal of Forecasting* 16, 293-315.

[10] Pesaran, M.H. and Timmermann, A. (1992). A Simple Nonparametric Test of Predictive Performance, *Journal of Business and Economic Statistics* 10, 461-465.

[11] Pons, J. (2000). The Accuracy of IMF and OECD Forecasts for G7 Countries, *Journal of Forecasting* 19, 53-63.

[12] Schnader, M.H. and Stekler, H.O. (1990). Evaluating Predictions of Change, *The Journal of Business* 63, 1, 99-107.

[13] Stekler, H.O. (1994). Are Economic Forecasts Valuable?, *Journal of Forecasting* 13, 495-505.