Banco de México

Documentos de Investigación

Banco de México

Working Papers

N° 2007-14

Optimality Tests for Multi-Horizon Forecasts

Carlos Capistrán
Banco de México

December 2007

Documento de Investigación
2007-14

Working Paper
2007-14

Optimality Tests for Multi-Horizon Forecasts*

Carlos Capistrán†
Banco de México

Abstract

This paper develops and analyzes a series of tests to evaluate the optimality of forecasts when forecasts for more than one horizon are available. The tests are based on the property that the unconditional expected loss of optimal forecasts should not decrease with the forecast horizon (e.g., under quadratic loss the variance of optimal forecast errors should not decrease with the horizon). The tests complement existing methods of forecast evaluation, such as Mincer-Zarnowitz-type tests, by using an implication of optimality that directly concerns forecasts made at different horizons. The finite sample performance of the tests is analyzed and an illustration using the Survey of Professional Forecasters is provided.
Keywords: Forecast Evaluation, Composite Hypothesis.
JEL Classification: C12, C53, E27.

Resumen

Este documento desarrolla y analiza una serie de pruebas para evaluar la eficiencia de pronósticos cuando éstos se encuentran disponibles para más de un horizonte. Las pruebas se basan en la propiedad de que la pérdida esperada no condicional de los pronósticos óptimos no debe decrecer con el horizonte de pronóstico (e.g., para el caso de una pérdida cuadrática, la varianza de los errores de un pronóstico óptimo no debe decrecer con el horizonte). Estas pruebas complementan métodos existentes para la evaluación de pronósticos, como lo son las pruebas del tipo Mincer-Zarnowitz, al utilizar una implicación de eficiencia que concierne directamente a pronósticos hechos para distintos horizontes. Se analiza el desempeño de estas pruebas para muestras finitas y se provee un ejemplo empleando la Encuesta de Pronosticadores Profesionales (SPF, por sus siglas en inglés).
Palabras Clave: Evaluación de Pronósticos, Hipótesis Compuestas.

*I am grateful to Graham Elliott and Allan Timmermann for posing the problem and for their comments. Conversations with Clive W.J. Granger were very helpful. I thank Vince Crawford, Marjorie Flavin, Antonio E. Noriega, Yixiao Sun, Elliot Williams and Carla Ysusi for useful comments.
† Dirección General de Investigación Económica. Email: ccapistran@banxico.org.mx.

1 Introduction

The usual way to proceed in the forecasting literature when analyzing a set of forecasts is to define what an optimal forecast is, and then to use the properties derived from optimality to evaluate it. Two properties, both derived under mean squared error loss (MSE), are routinely used for this purpose. The first, called unbiasedness property, is that forecast errors should have zero mean. The second, called efficiency property, is that one should not be able to predict those errors using information available to the forecaster when the forecasts were made (Clements and Hendry (1998); Granger and Newbold (1996)). Current tests of forecast optimality are designed to assess these two properties.[1]

But these properties refer to forecasts for a given horizon, which is in sharp contrast with the way forecasts are produced. Forecasters do not predict only one step ahead, but usually do so for two, three, or more steps at the same time, producing multi-horizon forecasts. Failure to incorporate this information when assessing the quality of forecasts could lead to mistakenly conclude that they are optimal. What is needed is a property that directly relates optimal forecasts at different horizons.

A good candidate is a property that, again under MSE, says that the further ahead one forecasts into the future, the less precise the optimal forecast is, where the precision is measured by the variance of the forecasts errors: a "decreasing precision" property (Granger and Newbold (1986)). When considering fixed-event forecasts, that is, forecasts of the same event made at different horizons, this property implies that the closer the event of interest, the more precise the forecast: a "convergence" property (Swidler and Ketcher (1990)). This property clearly establishes a relation between optimal forecasts at different horizons.

Although the possible use of decreasing precision as a tool for assessing forecast optimality has been considered before, so far no test using this property has been developed. For the convergence case, tests have been used (Bakhshi et al. (2005); Batchelor and Dua (1991); Swidler and Ketcher (1990)) but the properties of those tests and the validity of their application have not been considered.

Although the properties of optimal forecasts have been typically considered under MSE, recent literature on forecasting has explored the role of more general, possible asymmetric, loss functions (Diebold and Lopez (1996); Granger (1999)). Under general loss functions, the properties of optimal forecasts are somewhat different. For example, it is not true anymore that forecast errors should have zero mean, but an optimal bias may exist (Zellner (1986), Elliott et al. (forthcoming)). For the case of the property relating different horizons, the generalization shows that is not the variance of the errors the object that should be non-

[1] See the survey by Stekler (2002).

decreasing with the horizon, but their (unconditional) expected loss, which in some cases (such as MSE) coincides with the variance (Patton and Timmermann (forthcoming)).

It is the aim of this paper to further investigate the property of decreasing precision, or convergence, under general loss functions, and to develop tests for it.

2 Optimal Forecasts

The traditional way to assess the quality of forecasts made by the same forecaster for different horizons is to compare some statistical measure of the size of the errors, such as mean absolute error (MAE) or mean squared error (MSE). Table 1 is an example of what is usually reported.[2] How can Table 1 be interpreted? If the forecasters were forecasting the same quarters (which they are not) one may want to compare the sizes of the RMSE across forecasters for a fixed horizon. For that, one could use some test of predictive accuracy (e.g., Diebold and Mariano (1995)) But how about the information across different horizons? How can that be interpreted? The first thing that can be observed is that for most forecasters the root MSE (RMSE) increases monotonically with h, the forecast horizon, whereas for forecasters 1, 3, 96 and 452 this is not the case.[3] What are the implications of this behavior? Can the quality of the forecasters be assessed based on this difference?

A review of the literature on evaluation of optimal forecasts sheds light on how to answer these questions. But the literature has gone two separate ways when testing optimality. Both ways are related because they are based on the same properties of optimal forecasts. But they are different because of the structure of the data to which the properties have been applied. The first structure is known as rolling-event forecasts, where the forecaster is forecasting different events by fixing the forecast horizon and varying the initial date (Mincer and Zarnowitz (1969)). The second structure is known as fixed-event forecasts because, as described before, fixes the event but varies the horizon, so a forecaster is naturally approaching the event (Nordhaus (1987)).[4] In view of the division of the literature, the review in this section is divided in three parts, with the first subsection reviewing the common ground, the properties of optimal forecasts, and the next two reviewing the specifics of each structure.

[2]The table contains information used in the application at the end of the paper, where details about its construction can be found. The forecasters are a random sample taken from the Survey of Professional Forecasters. RMSE_h is the root MSE at horizon h.

[3]The MSE is the variance of the errors if the forecasts are unbiased.

[4]With data on multi-horizon forecasts one can always see the data as having either structure.

2.1 Optimal Forecasts in the Traditional Linear-Quadratic Case

The traditional framework for analyzing optimal forecasts is the linear-quadratic framework under known parameters. The quadratic part refers to the loss function employed, MSE, and the linear part to the form of the model assumed for the forecast, which is linear in the parameters.

MSE loss implies that the optimal forecast is the conditional mean of the variable of interest, Y_t, the conditioning made upon the information set $\Omega_t : Y_{t-j}, j = 0, 1, ...$ which is called a proper information set.[5] The conditional mean is approximated by a linear model, which may be a good approximation if Y_t is a covariance stationary process.[6]

Denote $f_{t+h,t}$ the forecast of the random variable Y_{t+h} made at time t (the h-step ahead forecast). The optimal forecast is : $f^*_{t+h,t} = \sum_{j=0}^{\infty} \psi_{h+j}\varepsilon_{t-j}$ with corresponding forecast error: $e^*_{t+h,t} = y_{t+h} - f^*_{t+h,t} = \sum_{j=0}^{h-1} \psi_j \varepsilon_{t+h-j}$ (Granger and Newbold (1986)). The usual properties of optimal forecasts in this context are (Granger and Newbold (1986)):

1. The optimal forecast error has conditional and unconditional mean of zero. That is, the optimal forecast is unbiased.

2. The optimal forecast error is orthogonal to any function of any element contained in the information set Ω_t. An implication of this property with a proper information set is that the h-step ahead optimal errors have the same autocorrelation as an $MA(h-1)$.

These properties of optimal forecasts have been used extensively in economics. Most of the time linked to the rational expectations literature, but also in other contexts.[7]

The properties refer to a single forecast horizon. However, Nordhaus (1987) presents a way to use the second property when data on multi-horizon forecasts are available. He notices that if the forecasts are of the fixed-event type, property 2 implies that forecasts' revisions should be unforecastable with information available before the revision. By using information contained in various horizons, the test is helpful in analyzing the way new information is incorporated into the forecasts.

But optimal forecasts have other properties that link forecasts at different horizons more directly. In particular, from the variance of the optimal forecast errors, $var(e^*_{t+h,t}) = \sigma^2_\varepsilon \sum_{j=0}^{h-1} \psi^2_j$, a third property can be derived (Granger and Newbold (1986)):

[5] In the traditional linear-quadratic case, the information set typically includes the past and present of the variable of interest. But the set can include other variables as well, as long as they are known when the forecast is made.

[6] The Wold representation theorem asserts that any covariance stationary process can be represented as an infinite moving average plus a linearly deterministic term $y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + v_t,$, where $\varepsilon_t \sim WN(0, \sigma^2_\varepsilon)$, as long as the coefficients of the MA are absolutely summable (Hamilton (1994)).

[7] For a review of various uses, as well as the tests that have been developed, see Diebold and Lopez (1996).

3. The variance of an optimal forecast is a non-decreasing function of h, the forecast horizon.

In this setting, the conditional and the unconditional variance of the forecast error are the same, so the property applies to both of them. When the forecasts are of the rolling-event type, property 3 applies in a forward way and is interpreted as a loss of precision as one forecast further into the future. When the forecasts are of the fixed-event type, property 3 applies in a backward way and is interpreted as a gain of precision as one approaches the event of interest.

2.2 Rolling-Event Forecasts

In contrast to the first two properties, no test for the property of decreasing precision can be found in the literature, despite several researchers finding the property useful to asses the optimality of forecasts.[8] For example, Granger (1981) notes: "If $f_{t+h,t}$ is an optimal forecast [under MSE] based on a proper information set, then the variance of the h-step forecast error will be a non-decreasing sequence as h increases... if forecasts do not make optimal use of an information set the forecast error variances need not increase as one forecasts further ahead" or more recently, in his undergraduate textbook on forecasting, Diebold (2001) writes: "We have learned a number of lessons about optimal forecasts while ignoring parameter estimation uncertainty, such as: Forecast error variance grows as the forecasts horizon lengthens,... Such lessons provide valuable insight and intuition regarding the workings of forecasting models and provide a useful benchmark for assessing actual forecasts...".

For the example at the beginning of this section (Table 1), the literature so far advises that the optimality of forecasters 1, 3, 96 and 452 (the non-monotonic ones) must be called into question, although no formal method to do it has been proposed.

Perhaps the principal reason for which tests of the third property have not been developed is because when the assumption of parameter certainty is relaxed the property vanishes. The result goes back in the economic forecasting literature at least to Schmidt (1974), and is known as the "non-monotonicity of the MSE".

The problem is that under parameter uncertainty, the conditional variance has a second term that depends on the difference between the estimate and the population parameter, and this term can be increasing in h. There are no general results on this second term, but some special cases have been worked out. For example, for the $AR(1) : y_t = \phi y_{t-1} + \varepsilon_t$,

[8] A few tests exists for the "converge" form of the property, that is, when applied to fixed-event forecast. These tests are discussed later.

$|\phi| < 1$, $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$, Clements and Hendry (1998) present the following result:

$$var(\widehat{e}^*_{T+h,T} \mid \Omega_T) = \sigma_\varepsilon^2 \frac{\left(1-\phi^{2h}\right)}{1-\phi^2} + E\left[\left(\phi^h - \widehat{\phi}^h\right)^2\right] y_T^2,$$

which they approximate by:

$$var(\widehat{e}^*_{T+h,T} \mid \Omega_T) = \sigma_\varepsilon^2 \frac{\left(1-\phi^{2h}\right)}{1-\phi^2} + h^2\phi^{2(h-1)}T^{-1}\left(1-\phi^2\right) y_T^2,$$

where T is the total number of observations used in the OLS estimation of the $AR(1)$ model (see also Box et al., 1994, p. 304). The first term, the asymptotic variance of the error, is the same as in the case of parameter certainty, so it is non-decreasing in h. The second term is where the problem lies. Chong and Hendry (1986) noticed that $h^2\phi^{2(h-1)}$ is non-monotonic in h as it has a maximum. So the second term can make the conditional variance non-monotonic in h. The result is true only in finite samples, because the term vanishes asymptotically ($T \to \infty$). But this means that with finite samples it is more difficult to test the third property.

This result led to conclusions like the following, again from Diebold's textbook (2001): "[On the lessons learned under known parameters] they sometimes need modification when parameter estimation uncertainty is acknowledged: Forecast error variance need not grow monotonically with horizon. Typically we *expect* forecast error variance to increase monotonically with horizon, but it doesn't *have* to." (Italics in the original). Going back to Table 1, under parameter uncertainty the advice on how to interpret the non-monotonicity of forecasters 1, 3, 96 and 452 is not clear. Something could be wrong, but because the variance *does not have to increase* (according to Diebold), one cannot be conclusive.

2.3 Fixed-Event Forecasts

Swidler and Ketcher (1990), looking at fixed-event forecasts, notice that because the information set is non-decreasing in the forecast horizon, the optimal forecast of event τ made h steps before should be a more precise forecast than the optimal forecasts for the same event made $h+j$ steps before ($j > 0$). This is the same as property 3 above, but applied to fixed-event forecasts. To test this property they looked at the R^2s of efficiency regressions ($Y_\tau = \alpha_h + \beta_h f_{\tau,\tau-h} + u_{\tau,\tau-h}$) made at different horizons. Their argument is that if the property is satisfied by the forecasts, the R^2s should be non-increasing when h increases.[9]

[9]This makes sense in a linear-quadratic set-up because the $R\hat{}2$ is a monotone transformation of the variance of the forecast error.

Analyzing Blue Chip surveys of real GNP growth and inflation forecasts they find that for both variables the forecasters seem to be able to make better forecasts the closer the event of interest. The problem with their approach is that the sample distribution of the $R^2 s$ is not taken into account.

Batchelor and Dua (1991) give the property the name of convergence. They write it as:

$$E\left(e^{*2}_{\tau,\tau-h} \mid \Omega_{\tau-h}\right) - E\left(e^{*2}_{\tau,\tau-h-j} \mid \Omega_{\tau-h-j}\right) \leq 0, j > 0,$$

and propose two tests. One is to run a regression of the squared errors on a constant and h, and test for a non-negative coefficient for h. The problem with this test is that the relation does not have to be a linear one, as they assume. Also, the autocorrelation in the residuals has to be properly addressed, as from property 2 above, even optimal forecast errors present serial correlation. The second test is a non-parametric binomial test, where the test statistic is the number of increases of the squared errors from month to month. But they apply the test using only the forecasts for one event at a time, losing the information contained when pooling the forecasts across events. Looking at real GNP, the price deflator of GNP, Unemployment, and T-bill rates, they find that the consensus forecast in each case seems rational, but that analysis of individual forecasters gives evidence of irrationality (which they pair with lack of optimality, as many in the literature) for some forecasters.

From this literature, the non-monotonic forecasters from Table 1 would be labelled non-optimal (further, irrational), as they apparently fail to satisfy the convergence property.

3 The Property of Non-Decreasing Expected Loss

In recent years, several generalizations to the traditional linear-quadratic framework of optimal forecasting have been made (e.g., Granger (1999)). Non-linear models have been considered, as in Granger and Terasvirta (1993), Clements and Hendry (1998), and Patton and Timmermann (2007) and other loss functions, including asymmetric ones, have also been employed, for example Granger (1969), Christoffersen and Diebold (1996) and Patton and Timmermann (2007).

In what follows, the theory behind the convergence, or decreasing precision, property of optimal forecasts is presented in a general setting, building on the work of Granger (1999) and Patton and Timmermann (forthcoming, 2007). Asymmetric loss functions as well as different, possible nonlinear, models for different horizons are allowed. The property is then used to set the null hypothesis of interest.

3.1 The Property of Interest

If the variable of interest is Y_t and the forecast is for h steps ahead, then one is concerned with forecasting Y_{t+h}. The conditional distribution of this random variable is $F_{Y_{t+h}|\Omega_t}(y) = \Pr(Y_{t+h} \leq y \mid \Omega_t)$, the conditioning been made with respect to the information known at time t, denoted by Ω_t. The information set is such that it contains the past and present of the variable of interest and any other variable available to the forecaster at the time the forecast is produced, $\Omega_t : Y_{t-j}, X_{i,t-j} \quad i = 1, ..., l \quad j = 0, 1...$ It is assumed that any deterministic part has been taken away from Y_t. The loss function $L : \mathbb{R} \longrightarrow \mathbb{R}^+$ takes a forecast error and returns its loss. This is what Christoffersen and Diebold (1997) called a prediction-error loss function, $L(y_{t+h} - f_{t+h,t}) = L(e_{t+h,t})$. The only requirements for the loss function is that it be weakly convex, with the normalization $L(0) = 0$, and that it be at least once differentiable. In this setting, an optimal forecast is defined as follows:

Definition 1 (Optimal forecast) *An* optimal forecast $f^*_{t+h,t}$ *is the forecast that minimizes expected loss. In a parametric framework, this can be put as:*

$$f^*_{t+h,t} \equiv g_h(x_t, \beta^*_h)$$

$$\beta^*_h \equiv \arg\min_{g_h} E\left[L\left(y_{t+h} - g_h(x_t, \beta_h)\right) \mid \Omega_t\right],$$

where

$$E\left[L\left(y_{t+h} - g_h(x_t, \beta_h)\right) \mid \Omega_t\right] = \int_e L\left(y_{t+h} - g_h(x_t, \beta_h)\right) dF_{e|\Omega_t}(e),$$

$$x_t \in \Omega_t, \quad \beta_h \text{ is a parameter.}$$

If a forecaster wants to jointly forecast more than one step-ahead at a time, the problem may be better modelled using a multi-horizon loss function. However, this is not common in the literature. For this paper, one can think of the forecaster as having a loss function over multiple horizons but separable among them. In this case the results are equivalent to the ones presented here.[10]

Because the optimal forecast is defined as the solution to a minimization problem, the usual two mathematical objects to look at to examine its properties are the first order condition (FOC) and the value function.

[10]For example, using MSE error loss, if one wants to generalize the loss, by defining $e_H = (e_{t,t+1}, ..., e_{t,t+H})'$ one can restate the problem as $\min e'_H W e_H$. The problem is equivalent to minimizing the loss one horizon at a time (i.e. the problem is separable), as long as W is a diagonal positive definite matrix (Capistrán (2006)).

Definition 2 (FOC) *The* first order condition *to solve for the optimal forecast is:*[11]

$$E\left[\left.\frac{\partial L\left(y_{t+h}-g_h(x_t,\beta_h)\right)}{\partial\beta_h}\right|_{\beta_h^*} \mid \Omega_t\right]=0.$$

The literature on forecasts evaluation has concentrated on testing the optimality of a set of forecasts by using this condition, usually referred to as the martingale difference sequence (mds) property (when $h=1$) (e.g., Granger (1999); Patton and Timmermann (forthcoming)). Any test based on this property has to specify a loss function and the information set used by the producer of the forecast. For example, when the information set includes the past of the series, then the mds property implies that the h-step ahead marginal loss, evaluated at the optimum, has zero conditional and unconditional mean and a moving average structure which is at most $MA(h-1)$. If the loss is MSE, then the marginal loss is the forecast error and these properties apply to it, showing that this is indeed a generalization of the linear-quadratic framework. As mentioned before, these properties only concern one forecast horizon at a time. All the efficiency tests in the literature can be seen as variants of this property. The unbiasedness property is obtained when a constant is consider (as it is always a part of the information set), and the efficiency property is obtained with any other variable that belongs to the information set.

Definition 3 (Value function) *The* value function *is a function of the time the forecast is made for, the time the forecast is made at, and the information set:*[12]

$$V\left(t+h,t,\Omega_t\right)=E\left[L\left(y_{t+h}-g_h(x_t,\beta_h^*)\right)\mid\Omega_t\right].$$

So far, the set up corresponds to rolling-event forecasts, so the property of decreasing precision is the relevant one. Patton and Timmermann (forthcoming) prove it using definition 3. It relates optimal forecasts for different horizons by comparing their value function, holding everything else constant. Their result is summarized in the following proposition, taken from Patton and Timmermann (forthcoming), but repeated here for completeness and to state it in the notation of this paper:

Proposition 4 (Decreasing Precision) *(From Patton and Timmermann (forthcoming)). If the optimal forecast error is strictly stationary and the loss function is of the prediction*

[11] This is true as long as F fulfills some regularity conditions that allow the interchange of differentiation and integration. The FOC is necessary for an optimum. If L is strictly convex, then the FOC is also sufficient.

[12] In statistical terminology, the value function is the risk evaluated at the optimum.

error type and time invariant, then the unconditional expected loss of an optimal forecast is a non-decreasing function of the forecast horizon, h, for a given information set Ω_t. i.e.

$$V(t+h,t) \leq V(t+h+j,t) \qquad \forall j > 0$$

Proposition 4 can be interpreted as saying that in general it is harder to forecast the longer the horizon (i.e. the precision of the forecasts is decreasing in h, hence the name).

The property is general enough to allow for asymmetric loss functions. It is stated here assuming a prediction-error loss function, but the property can be proved for more general loss functions provided one makes appropriate assumptions on F, for example, strict stationarity of Y_{t+h}. Conversely, when one makes stronger assumptions on the loss function, requirements on F can be relaxed, for example, under MSE loss only covariance stationarity is needed.

The requirement of a strictly stationary optimal forecast error can be obtained in several ways. One is by directly assuming that Y_{t+h} is a strictly stationary random variable. But, as stated, proposition 4 allows for a non-strictly-stationary Y_{t+h}. This is theoretically possible as the optimal forecast cancels any dominant property of the variable of interest. In this way, what is relevant is that the forecast error satisfies the strict stationarity assumption.

When the multi-horizon forecasts are viewed as fixed-event forecasts, the relevant property is convergence. The property can also be proved using definition 3 (changing notation in the obvious manner).[13]

Proposition 5 (Convergence) *If the loss function is time invariant, and the information set is a filtration, then the unconditional expected loss of an optimal forecast is a non-decreasing function of the forecast horizon, h, for a given information set $\Omega_{\tau-h}$. i.e.*

$$V(\tau,\tau-h) \leq V(\tau,\tau-h-j) \qquad \forall j > 0$$

Proposition 5 can be interpreted as saying that in general it is easier to forecast the closer one is to the target (i.e. the forecasts converge to the outcome of the event of interest, hence the name).

The property is general enough to allow for asymmetric loss functions. The advantage of using convergence is that the assumption of strict stationarity is not needed. Also, the interpretation of the forecasters converging to the actual value of the variable seems more intuitive. Likewise, failure of forecasts to fulfill convergence are easily interpreted as failure of the forecasters to incorporate new information. For the rest of the paper, the converge

[13] Proofs of the propositions are presented in the Appendix.

form of the property is used, although the discussion and tests to come apply to both forms.

3.2 Comments on the Property of Interest

The first thing to notice is that the property concerns the unconditional expected loss, but that the optimal forecast employed in its calculation is a function of the information set. In addition, note that the *conditional* expected loss need not be a non-decreasing function of h.

When the loss function is MSE, the unconditional value function equals the unconditional variance of the optimal forecast error, $E\left[L\left(y_\tau - f^*_{\tau,\tau-h}\right)\right] = var(e^*_{\tau,\tau-h})$. In this case, property 3 states then that the unconditional variance is non-decreasing in h, which shows again that the traditional linear-quadratic framework is a particular case. In some situations the result is also applicable to the conditional variance (e.g., if the forecasts are generated by an $AR(1)$).

It is possible in some cases to use a test applied to the variance as a test of the property of interest, even when the loss function is not quadratic. The result is summarized in the following proposition:

Proposition 6 *If the loss function is of the prediction error type, time invariant,* $E\left[e^*_{\tau,\tau-h}\right] = E\left[e^*_{\tau,\tau-h-j}\right]\ \forall j > 0$, *and provided one of the following two conditions hold,*

1. *The loss function is at most twice differentiable, or*

2. *At least the first two moments of the conditional distribution function exists,*

then,

$$E\left[L(e^*_{\tau,\tau-h})\right] - E\left[L(e^*_{\tau,\tau-h-j})\right] \approx var(e^*_{\tau,\tau-h}) - var(e^*_{\tau,\tau-h-j}), \quad \forall j > 0.$$

When are the conditions required for proposition 6 met? Whenever the loss function is time invariant and $f^*_{\tau,\tau-h} = E\left[Y_\tau \mid \Omega_{\tau-h}\right]$, the result will follow, as is clear from Granger's (1969) requirements under which the optimal forecast is the conditional mean. The conditions are also met if the conditional distribution is Gaussian, even if the loss is asymmetric, because in that case $f^*_{\tau,\tau-h} = E\left[Y_\tau \mid \Omega_{\tau-h}\right] + \gamma$, but γ is independent of the information set (Granger (1999)). The conditions are not going to be met in general with asymmetric loss functions if the conditional distribution function depends on the information set, because in that case the optimal bias is time varying (Granger (1999); Christoffersen and Diebold (1997)).

3.3 The Hypothesis of Interest

With data on multi-horizon forecasts, one may want to test the optimality of the forecasts by testing proposition 5, the convergence property. This suggests the composite null hypothesis:

$$H_0 : E\left[L\left(e^*_{\tau,\tau-1}\right)\right] \leq E\left[L\left(e^*_{\tau,\tau-2}\right)\right] \leq ... \leq E\left[L\left(e^*_{\tau,\tau-H}\right)\right], \tag{1}$$

where H is the largest forecasting horizon. Hypothesis (1) is composite, involves multiple inequalities, and is specified in terms of the optimal population parameter.[14]

In the rest of the paper, hypothesis (1) will be referred to as the hypothesis of interest.

4 Testing for Convergence

4.1 Simplification of the Hypothesis of Interest

The hypothesis of interest is a complex object. It is a composite hypothesis that involves multiple inequalities. There are two approaches to deal with it. One is to perform a simultaneous test. This approach has the advantage that it can bound the overall size of the test, but in general is complicated (Gourieroux and Monfort (1995)).[15] Further, when a rejection of the null occurs, one cannot tell which horizon is driving the problem.

A second approach is known as multiple comparison. This approach is less complicated and also bounds the size of the overall test (although the rejection regions are different, see Goldberger (1992)). The advantage is that it allows the detection of the horizon(s) that cause trouble. The second approach is followed in this paper to test the hypothesis of interest.

The Union-Intersection Principle (UIP) is applied to implement a multiple comparison test (Roy (1953); Sen and Silvapulle (2002)).[16] The hypothesis of interest can be written as an intersection of $H-1$ hypotheses:

$$H_0 : E\left[L\left(e^*_{\tau,\tau-1}\right)\right] \leq E\left[L\left(e^*_{\tau,\tau-2}\right)\right] \cap ... \cap \left[EL(e^*_{\tau,\tau-(H-1)})\right] \leq E\left[L\left(e^*_{\tau,\tau-H}\right)\right], \tag{2}$$

[14]The property exists for the population parameter but not for the estimate (i.e., under parameter uncertainty), so one can only test if the property holds in population.

[15]Wolak (1987) proposes simultaneous tests that can be used in this setting. The tests are Wald tests with complicated distributions that are a weighted average of χ^2 distributions, and that have to be calculated in a case by case basis.

[16]When the null hypothesis can be expressed as an intersection, $H_0 : \theta \in_{i\in I} \Theta_\gamma$, where I is an arbitrary index, and tests are available for each of the component hypotheses: $H_{0,i} : \theta \in \Theta_i \; vs \; H_{a,i} : \theta \in \Theta_i^c$, the UIP can be used to find an appropriate testing procedure for H_0. According to the UIP, if the rejection region for the ith test is $\{x : T_i(x) \in R_i\}$, then the rejection region for H_0 is $_{i\in I}\{x : T_i(x) \in R_i\}$. H_0 is rejected if at least one of the component nulls $H_{0,i}$ is rejected.

so the UIP can be applied to separate it in $H-1$ component hypotheses, where the ith hypothesis is:

$$H_{0,i} : E\left[L\left(e^*_{\tau,\tau-i}\right)\right] \leq E\left[L\left(e^*_{\tau,\tau-(i+1)}\right)\right]$$
$$H_{a,i} : E\left[L\left(e^*_{\tau,\tau-i}\right)\right] > E\left[L\left(e^*_{\tau,\tau-(i+1)}\right)\right].$$

4.2 Bounds for the Size of the Overall Test

If tests are available for each of the component hypothesis, the issue is how to control the overall size of the testing procedure for the hypothesis of interest. Assume for now that those tests are available. The tests will be built later on in the paper.

According to the UIP, each of the component hypothesis has to be tested, and the hypothesis of interest is rejected if at least one of the component hypotheses is rejected. To control the overall size, each of the component tests has to be performed with a significance level smaller than the level desired for the overall hypothesis.

If the tests for the component hypotheses were independent, then to get a level α test the appropriate level for each test would be $\alpha_i = \left[1-(1-\alpha)^{\frac{1}{(H-1)}}\right]$. But the tests are likely to be highly correlated due to the $MA(h-1)$ property of the (generalized) forecast errors. The Bonferroni bounds provide a way to deal with this dependence. The individual significant levels under these bounds are $\alpha_i = \frac{\alpha}{(H-1)}$, for a level α test.

But the Bonferroni bounds are known to be very conservative, especially when the tests are highly dependent. To deal with this issue, Simes (1986) proposed a procedure based on a modification of these bounds. Simes's modification is simple: Let $pv_{(1)}, ..., pv_{(H-1)}$ be the ordered p-values for testing the component hypotheses $H_{0,(1)}, ..., H_{0,(H-1)}$, with $pv_{(1)}$ the largest value. Then the hypothesis of interest H_0 is rejected if $pv_{(i)} \leq \frac{i\alpha}{(H-1)}$ for any $i = 1, ..., (H-1)$. Simes (1986) shows that this procedure has size α for independent tests and a significance level much closer to the nominal level than the classical Bonferroni procedure in the dependent case.[17] Simes also shows that the modified procedure also has higher power than the Bonferroni Bounds when the tests are highly correlated.

The use of the UIP and Simes's modification to the Bonferroni bounds suggest the following testing procedure for the composite hypothesis:

1. Calculate the $(H-1)$ p-values of the component hypotheses.

2. Take the largest p-value and compare it to $\alpha_1 = \frac{\alpha}{(H-1)}$. If the p-value is smaller than α_1 stop. Reject the null hypothesis of interest. Proceed otherwise,

[17] For an analysis under dependence see Sarkar and Chang (1997).

3. Take the second-to-largest p-value and compare it to $\alpha_2 = \frac{2\alpha}{(H-1)}$. If the p-value is smaller than α_2 stop. Reject the null hypothesis of interest. Proceed otherwise,

4. The procedure continues until the smaller p-value is compared to $\alpha_{(h-1)} = \frac{(H-1)\alpha}{(H-1)} = \alpha$. If the p-value is smaller the null hypothesis of interest is rejected. If not, the conclusion is that there is not enough evidence in the sample to reject the null hypothesis of interest.

The procedure is a little involved, but there are some possible shortcuts:

1. If the (sample) expected loss is increasing with the forecast horizon for a given forecaster, then the immediate conclusion for him is that there is not enough evidence to reject the null of interest.[18] So, Simes's procedure can be applied only to the non-monotonic forecasters. This should give the same results but with a possible increase in power.

2. If H is small, yielding a small number of component hypotheses, then it is well known that Bonferroni bounds are close to the individual significance levels one would use under independence (Miller (1980)). So, for small H (according to Miller $H < 6$ is small enough) one is in solid ground by using the classical Bonferroni Bounds.[19]

Given a test for each of the component hypotheses, p-values can be calculated and the hypothesis of interest can be tested. Not only that, but by using Simes's procedure the overall size can be bounded. The next section develops tests that can be applied to each of the component hypotheses.

5 Tests for each of the Component Hypotheses

In order to construct tests for each of the component hypotheses, a probability structure is needed. Based on that structure, two tests are proposed based on asymptotic results. A nonparametric test is also discussed. Finally, the finite sample properties of the tests are examined via a Monte Carlo experiment.

[18] In this case all the p-values will be $>> \alpha$.

[19] For example, under independence, for $H = 4$ and $\alpha = 0.10$, $\alpha_i = 0.0345$, whereas by using the Bonferroni bounds, $\alpha_i = 0.0333$ gives $\alpha \leq 0.10$.

5.1 Description of the Environment

A $P \times H$ matrix of fixed-event forecasts is available, where P is the number of forecasts and H the number of forecast horizons available. If $f_{t,t-h}$ is the forecast of y_t done in $t-h$, the matrix contains data $\{f_{t,t-1}, ..., f_{t,t-H}\}_{t=1}^{P}$. The forecasts can be based on regression estimators $\widehat{\beta}_{h,R}$, as in West (1996) and White (2000), in which case the following assumptions are made: recursive estimation has been used, a total of $T+1$ observations are available, the first sample employed is of size $R-h+1$ for the first h-step-ahead forecast so that $T = R+P-1$ is the last date for which a one-step-ahead forecast is made, and the observed data are generated by a stationary strong ($\alpha-$) mixing sequence $\{X_i\}$.[20]

Following West (1996) and White (2000), the object of interest is the $H \times 1$ vector of moments $E(v_t^*)$ where $v_t^* \equiv L\left(y_t - f_{t+h,t}^*\right) \equiv g(X_t, \beta^*)$ is an $H \times 1$ vector, X_t is a random vector with marginal distributions equal to those of $\{X_i\}$, and under the null hypothesis of interest, $p\lim \widehat{\beta}_R \equiv \beta^*$.

Given a loss function, the actual values, and the forecasts, one can form the $H \times 1$ vector $v_t\left(\widehat{\beta}_n\right) = \left[\begin{array}{ccc} L\left(\widehat{e}_{t,t-1}\right), & ... & L\left(\widehat{e}_{t,t-H}\right) \end{array}\right]'$, where the dependence on the estimators is made explicit, and calculate the $H \times 1$ statistic:

$$\overline{v} \equiv P^{-1} \sum_{t=1}^{P} v_t\left(\widehat{\beta}_i\right).$$

Theorem 4.1 in West (1996) gives regularity conditions on the moments, dependence and heterogeneity of the vector sequence $\left\{v_t\left(\widehat{\beta}_n\right)\right\}_{t=1}^{P}$, such that a CLT exists:

$$P^{\frac{1}{2}}\left(\overline{v} - E\left(v^*\right)\right) \xrightarrow{d} N\left(0, S\right). \tag{3}$$

The $H \times H$ matrix

$$S \equiv \lim_{P\to\infty} var\left[P^{-\frac{1}{2}} \sum_{t=1}^{P} v_t^*\right] \tag{4}$$

is the spectral density of $(\overline{v} - E(v^*))$ at frequency zero scaled by 2π, provided that either $K \equiv E\left[\frac{\partial v_t^*}{\partial \beta}\right] = 0$ or $P/R \longrightarrow 0$ as $T \longrightarrow \infty$. If neither of these conditions hold, the expression for the variance is more complex (West (1996)). K will be zero if the loss function used by the producer of the forecasts is the same as the loss used for the evaluation, and if the forecasts are optimal (notice that the last part is true under the null). The second condition requires the in-sample number of observations to increase faster than the out-of-

[20] The timing for the forecasts is a little different than the usual case for one-step-ahead. This is because the usual case employs rolling-event forecasts.

sample number of observations.[21]

Because of the autocorrelation that is likely to be present, especially when H is large, and because in general there is no further information about it, care must be exercised when estimating S.[22]

West's CLT only applies to loss functions that are twice differentiable, to models that are linear in the parameters, and if the models used to produce the forecasts at different horizons are not nested.[23] So, although the setting in section 3 allows for more complex models, the CLT available restricts the application of the results.

5.2 Tests for the Component Hypotheses

To develop a test for each of the component hypotheses is sufficient to construct a test for the first component, the one involving forecasts at one and two steps ahead:

$$H_{0,1} : E\left[L\left(e^*_{t,t-1}\right)\right] \leq E\left[L\left(e^*_{t,t-2}\right)\right]$$
$$H_{a,1} : E\left[L\left(e^*_{t,t-1}\right)\right] > E\left[L\left(e^*_{t,t-2}\right)\right].$$

The tests for the other components follow as the obvious generalization.

Using the CLT in (3), two testing procedures can be constructed, one akin to Diebold and Mariano's test (1995) and another inspired in a variance ratio test.

5.2.1 A test based on the difference of the expected losses

The hypotheses can be stated as:

$$H_{0,1} : E\left[L\left(e^*_{t,t-1}\right) - L\left(e^*_{t,t-2}\right)\right] \leq 0$$
$$H_{a,1} : E\left[L\left(e^*_{t,t-1}\right) - L\left(e^*_{t,t-2}\right)\right] > 0,$$

and a test statistic can be based on the average:

$$\overline{d}_{P,1} \equiv \frac{1}{P}\sum_{t=1}^{P} d_{t,1} = \frac{1}{P}\sum_{t=1}^{P}\left[L\left(\widehat{e}_{t+1,t}\right) - L\left(\widehat{e}_{t+2,t}\right)\right].$$

[21] For an alternative asymptotic theory, using different assumptions see Giacomini and White (2006).

[22] Diebold and Mariano (1995) propose to use the information that an optimal h-step ahead forecast has the same autocorrelation than (at most) an $MA(h-1)$. But this is misleading, because this autocorrelation refers to the mds property of the *derivative* of the cost function, not to the cost function itself, which is what is needed.

[23] McCraken (2000) has extended West's results for cases when the cost function is not differentiable, but $E(v^*_t)$ is. This allows the use of MAE cost and, in general, Lin-lin costs.

$H_{0,1}$ is rejected for large values of $\overline{d}_{P,1}$. The following proposition states that the asymptotics of the proposed test statistic is standard.

Proposition 7 *If $\widehat{S}$ is a consistent estimator of S (2×2 matrix in this case) and the CLT in (3) holds then, using the element in the null that is least favorable to the alternative,*

$$T_1 = \frac{P^{\frac{1}{2}} \overline{d}_{p,1}}{\widehat{se_d}} \overset{a}{\sim} N(0,1),$$

where $\widehat{se_d} = \left(G' \widehat{S} G\right)^{\frac{1}{2}}$ and G is a 2×1 vector with 1 in the $(1,1)$ position, -1 in the $(2,1)$ position and zeros elsewhere.

The p-value is $\Pr(Z \geq T_1)$, where Z is distributed standard normal.

5.2.2 A test based on the ratio of the expected losses

Alternatively, the hypotheses can be stated as:

$$H_{0,1} : \frac{E\left[L\left(e^*_{t,t-1}\right)\right]}{E\left[L\left(e^*_{t,t-2}\right)\right]} \leq 1$$
$$H_{a,1} : \frac{E\left[L\left(e^*_{t,t-1}\right)\right]}{E\left[L\left(e^*_{t,t-2}\right)\right]} > 1,$$

as long as $E\left[L\left(e^*_{t,t-2}\right)\right] \neq 0$, and a test statistic can be based on the ratio of the sample means

$$\overline{R}_{P,1} = \frac{\frac{1}{P}\sum_{t=1}^{P} L\left(\widehat{e}_{t,t-1}\right)}{\frac{1}{P}\sum_{t=1}^{P} L\left(\widehat{e}_{t,t-2}\right)} = \frac{\overline{v}_1}{\overline{v}_2}.$$

Since $\overline{R}_{P,1}$ tends to be large when $E\left[L\left(e^*_{t,t-1}\right)\right] > E\left[L\left(e^*_{t,t-1}\right)\right]$, $H_{0,1}$ is rejected for large $\overline{R}_{T,1}$. As for its distribution:

Proposition 8 *If $\widehat{S}$ is a consistent estimator of S (2×2 matrix in this case) and the CLT in (3) holds then, using the element in the null that is least favorable to the alternative,*

$$TR_1 = \frac{P^{\frac{1}{2}}\left(\overline{R}_{P,1} - 1\right)}{\widehat{se_R}} \overset{a}{\sim} N(0,1),$$

where $\widehat{se_R} = \left(G' \widehat{S} G\right)^{\frac{1}{2}}$ and G is a 2×1 matrix with $\frac{1}{\overline{v}_2}$ in the $(1,1)$ position, $-\frac{\overline{v}_1}{(\overline{v}_2)^2}$ in the $(2,1)$ position and zeros elsewhere.

The p-value is $\Pr(Z \geq TR_1)$, where Z is distributed standard normal.

5.2.3 A nonparametric test based on the difference of the expected losses

Nonparametric tests have been used in the forecasting literature because of its robustness against departures from assumptions such as normality and because they are usually more reliable in the small samples typically available in forecasting exercises (particularly in macroeconomics), although it is well known that they may are very sensible to outliers. For related tests used for the evaluation of forecasts see Campbell and Ghysels (1995) and Diebold and Lopez (1996).

When one is interested on testing whether one random variable in a pair (X, Y) tends to be larger than the other variable, the sign test is typically used. Following Conover (1980), three assumptions are needed for the sign test to be a reliable test: (i) the random variables (X_i, Y_i), $i = 1, ...n$ are mutually independent, (ii) the measurement scale is at least ordinal, and (iii) the pairs (X_i, Y_i) are internally consistent, in that if the $\Pr(X_i < Y_i) > \Pr(Y_i < X_i)$ for one i, then it is also true $\forall\ i$. Under these requirements, the sign test may be used to test:

$$H_0 : E[X_i] \leq E[Y_i]$$
$$H_a : E[X_i] > E[Y_i],$$

by using as a test statistic the total number of times that $Y_i < X_i$. By substituting $L\left(e^*_{t,t-1}\right)$ for X_i, and $L\left(e^*_{t,t-2}\right)$ for Y_i, the sign test can be seen as testing the same hypothesis as the difference test. In this case, a test statistic can be based on:

$$sn_{P,1} \equiv \sum_{t=1}^{P} \mathbf{1}_{\left(L\left(e^*_{t,t-1}\right) > L\left(e^*_{t,t-2}\right)\right)},\ \text{for } t = 1, ...P,\ \text{where } \mathbf{1} \text{ is the indicator function.}$$

H_0^1 is rejected for large values of $sn_{P,1}$. Following Conover (1980), under the three assumptions stated before, $sn_{P,1}$ is distributed as a binomial with parameters $\rho = \frac{1}{2}$ and $n = P$, and for large values of P, $(P > 20)$,

$$TS_1 = \frac{sn_{P,1} - \frac{P}{2}}{\frac{\sqrt{P}}{2}} \overset{a}{\sim} N(0, 1).$$

The p-value can be calculated as $p - value_1 = \Pr(Z \geq TS_1)$, where Z is distributed standard normal.

From the three assumptions required for the non-parametric test, the first is clearly not going to be satisfied by the expected losses, as high autocorrelation is expected. But the sign test has proven to be robust to departures from this assumption (Diebold and Mariano

(1995)), and may actually be useful in these conditions. The sign test is considered here as an alternative test because of its simplicity.

5.2.4 Comments on the tests

In all the tests the null is enforced by using the element in it that is least favorable to the alternative. The problem with this approach is that when the population moment is far from equality, the proposed tests will be undersized. This effect is usual when an inequality is present under the null hypothesis. The further away the population moment is from equality, the larger the size problem.

To implement the first two tests, a consistent estimator of S is needed. There are several consistent estimators in the literature that can be applied. For example, the nonparametric ones discussed in Andrews (1991) or the parametric of Den Haan and Levin (1997). The problem with all of them is that although consistent, they can perform poorly in finite samples. This has as a consequence that the tests will be oversized in finite samples.

There are three reasons why the first test seems, a priori, a better choice. It is based on a linear hypothesis, while the second is based on a non-linear one, it takes into account the autocorrelation structure, which the third does not, and it is related to the Diebold-Mariano (1995) test, that is well-known in the forecasting literature.

5.3 Finite Sample Performance

5.3.1 Experiment design

Several Monte Carlo experiments were performed to evaluate the finite sample size of the test statistics T_h, TR_h, and TS_h. For all the experiments the DGP under the null was taken to be a univariate AR(1) and the loss function, both to estimate the model and to evaluate the forecasts, was taken to be MSE.[24] The estimator used to estimate the long run covariance matrix when required is the VARHAC of Den Haan and Levin (1997).[25] All the tests were performed at $\alpha = 0.10$. Each experiment was repeated 5,000 times.

The first division among the experiments is between those in which the parameter of the AR(1) was taken to be known and those in which it has to be estimated. For the later case, recursive OLS estimation was employed.

For each of these two parts, two different assumptions about the stochastic part of the DGP were made. One assumption is that the error is drawn from a standard normal dis-

[24]The initial condition for the AR(1) is zero. The first 500 observations were dropped in each repetition to avoid the dependence of the results on the initial condition.

[25]For the VARHAC, the BIC model selection criteria was used to choose the lags of the VAR. The maximum lag was set to the integer part of $P^{\frac{1}{3}}$, as recommended by Den Haan and Levin.

tribution, the Gaussian errors case. The other assumption (following Diebold and Mariano (1995)) is that the error is drawn from a fat-tailed distribution. In this case the distribution used to generate the errors is a t distribution with six degrees of freedom. The errors were standardized by dividing them by $\sqrt{\frac{3}{2}}$.

The four experiments are shown in Tables 2 to 5. One table for each experiment.

For Tables 2 and 3, when the parameter of the AR(1) is known, three different number of forecasts were considered: $P = 50$, 100, 500, and three values for the AR(1) parameter: $\phi = 0.2$, 0.5, 0.8. The results are presented for the component hypotheses one, three and seven for the three tests.

For Tables 4 and 5, with parameter uncertainty, three different number of forecasts were considered: $P = 50$, 100, 500, and three cases for R, the size of the first sample used in the estimation: $R = 50$, 100, 500. Only the relevant combinations are reported (e.g., the combination where an initial sample of 50 is used to construct 500 forecasts was dropped). The DGP in these cases was taken to be the AR(1) with $\phi = 0.2$. The results are presented for the component hypotheses one, three and seven.

5.3.2 Results from the experiment

The first result to be derived from the tables is that the level of the tests is correct, as the actual size is smaller or close to 10%.

On specific results, for the cases where the DGP is known (Tables 2 and 3) the following effects can be noted for the first two tests:

1. As expected, the fact that the long-run covariance matrix has to be estimated causes the tests to be oversized. This effect can be seen by looking at a fixed column in Table 2. Everything else constant, as the number of forecasts decreases, the size increases.

2. Also as expected, the tests are undersized when the population moment is far from equality (in the null). In fact, the further the population moment is from equality the more undersized the tests. This effect can be seen in two ways in the tables: First, the more persistent the DGP, the smaller the size; and second, the smaller the horizon, the smaller the size. In both cases this is true because for an AR(1) the difference in the variance of the forecast errors between the h and the $h+1$ horizons is: $var(h) - var(h+1) = \frac{\phi^{2(h+1)} - \phi^{2h}}{1-\phi^2}\sigma_\varepsilon^2$, and its magnitude increases with ϕ and decreases with h (a similar result holds for the ratio).

The results show that the performance of the first two tests is very similar. The actual size is slightly better when the DGP is Gaussian, but the tests show a good performance when

applied to the non-Gaussian DGP. The non-parametric test performs remarkably, showing that it can be used as a first approximation

For the cases where the parameter of the DGP has to be estimated (Tables 4 and 5) the performance is very similar to the case under known DGP. In particular, as the sample size used in the estimation, R, increases, the size approaches the size of the tests under known parameter. This conclusion shows that even under parameter uncertainty the property of decreasing precision can be tested using one of the tests developed here.

6 An Illustrative Example

The property of decreasing precision can be tested to assess if a set of multi-horizon forecasts are optimal. Further, a testing procedure and tests that can be used for each of the component hypotheses are now available. The next step is to illustrate the application of the testing procedure. This is done by applying the tests to answer the questions raised by Table 1 concerning the performance of a sample of forecasters from the Survey of Professional Forecasters (SPF).

The SPF is a quarterly survey conducted by the Federal Reserve Bank of Philadelphia. The survey involves private economists that produce regular forecasts of economic variables as part of their jobs. The forecasters are from Wall Street, private banks, consulting firms and research centers, among others. The variables forecasted include output and its components, inflation and interest rates. Among the forecasts provided, quarterly forecasts for one to four steps ahead are made for most variables, with the first predictions going back to 1968 in some cases. The identity of the forecasters is maintained private, but a number is associated with each forecaster.[26]

For the application a sample of twenty forecasters was taken from the survey. The forecasts are for the level of nominal output (GNP before 1992, GDP from then on) for 1 to 5 steps-ahead. The sample was taken randomly from those forecasters with at least 10 observations per quarter, for every forecasting horizon, taking care of using the same number of observations for each horizon for the same forecaster. Not all the forecasters are predicting the same dates, and the number of observations per forecaster is different across forecasters, so a comparison across forecasters is impossible. But this is not a problem here, as the test developed in this paper concerns the evaluation of one forecaster at a time, not a comparison across forecasters. The forecasts were transformed from levels to growth rates for the analysis (therefore, only 1 to 4 step-ahead forecasts are available).

[26] A complete description of the survey can be found in Croushore (1993), or in the Federal Reserve Bank's web page: www.phil.frb.org/econ/spf.

The actual values are taken from the Bureau of Economic Analysis. The series is quarterly GDP (GNP before 1992), seasonally adjusted, in billions of dollars. Growth rates were calculated in the same way as for the forecasts. An assumption is made here that the forecasters were actually trying to forecasts the final values of nominal output, not the first released value.

Table 1 presents the RMSE at each horizon for each forecaster, where MSE loss is assumed. The first thing to notice is that for most of the forecasters the RMSE (the standard deviation if the forecasts are unbiased) is increasing with the horizon. This is common to most forecasts in the survey, as mentioned in a comprehensive study by Zarnowitz and Brown (1992): "Forecast errors generally increase as the number of periods in the forecast horizon increases". But one can also notice that for some forecasters this is not the case. As mentioned before, for forecasters 1, 3, 96 and 452 the RMSE decreases with the horizon at some point. In some cases it goes up again. One can test if this implies that the forecasters are not optimal by applying one of the tests proposed in this paper to the non-monotonic forecasters.

The first step is to calculate the p-values for each of the component hypotheses using one of the tests proposed before. The parametric test based on the difference of the losses is used in this example.

An easy way to perform the test is by regressing the sample differences $d_{t,h} = \widehat{e}^2_{t,t-h} - \widehat{e}^2_{t,t-(h+1)}$ on a constant and then to perform a one-sided t-test to see if the constant is negative or equal to zero. The autocorrelation of the errors in the regression has to be modelled in each case. To do this, after the first regression is estimated, an ARMA structure is selected for the residuals by looking at the autocorrelograms. Then the regression is estimated again with the selected ARMA model included. The p-value of interest is the one associated with the constant. For each forecaster three p-values are obtained in this way.[27]

The next step is to order the p-values for each forecaster. After that, to perform a test with a significance level of 0.10, each of the p-values is compared to $\alpha_j = \frac{j(0.10)}{3}$, $j = 1, 2, 3$ where $j = 1$ corresponds to the largest p-value. The null hypothesis of interest is rejected for each forecaster if at least one of his p-values is smaller than the corresponding α_j.

The results are shown in Table 6. The test only rejects the null hypothesis of decreasing precision for forecaster 452. For the other forecasters there is evidence that the decrease in the RMSE is not statistically significant.

According to the results, forecaster 1, 3 and 96 cannot be label non-optimal. But a problem is detected for forecaster 452. Further, it is clear that d_2 is the difference causing

[27]This is an alternative way to proceed with the test. It has the advantage that all the tools developed for regression can be applied, for instance, here we are explicitly modelling the autocorrelation.

trouble, which means that the forecasts for two or three steps-ahead, or both, are misspecified and can be improved. For example, using the forecasts from forecaster 452, a better forecast for two step-ahead is simply to use the three step-ahead forecast, but starting at $t-1$ instead of at t.[28]

All that is needed for the application, apart from the data on the forecast errors, is an assumption on the loss function of the producer of the forecasts. This is in contrast to other tests of forecast optimality, as other tests need also an assumption about the information used by the forecaster. For this example MSE loss is used. But even if the true loss is not MSE, as long as the forecasts' optimal biases are not time-varying, the use of the squared errors may shed light on the optimality of the forecast, as their movement may be proportional to the object of interest (proposition (6)).

7 Conclusions

The property of optimal forecasts that the expected loss of forecast errors has to be non-decreasing in the forecast horizon can be used to assess the optimality of multi-horizon forecasts.

A testing procedure can be applied provided the null hypothesis is set so that it relates population moments, not sample moments. The hypothesis of interest set in this way ends up being a composite hypothesis with multiple inequalities. The Union-Intersection Principle is used to deal with it. The overall size of the test is bounded by a procedure suggested by Simes that modifies the traditional Bonferroni bounds so as to make them less conservative for the case of dependent component tests. Three different tests for the component hypotheses are provided. All tests have correct level, but the test based on the differences of the expected losses, akin to Diebold and Mariano's test, appears to perform slightly better. This testing procedure can be applied using general loss functions and when forecasts are taken to be the product of estimated models, provided the models are linear and non-nested.

By applying this test one can study if a set of multi-horizon forecasts is using information efficiently, and therefore if decisions can be based on them with certain confidence, although this test should be complemented with other efficiency test, for example those surveyed by Diebold and Lopez (1996). In case of a rejection, the procedure indicates where the problem lies, and therefore indicates the horizons that have to be improved upon.

Advantages of the testing procedure developed here are that it directly uses the relation between forecasts for different horizons, and that it only needs an assumption about the forecaster's loss function, but not one about his information set. Under certain conditions,

[28] Using simple checks for outliers, none could be found.

even the assumptions on the loss function can be relaxed, as the test could be applied to compare the variance of forecast errors across horizons, provided the required conditions are met.

An interesting direction for future research is to investigate the degree of complementarity between the tests presented here and other optimality tests (i.e., the power of different tests). Further experience with the application of the tests is also needed.

References

[1] Andrews, D., 1991, "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817-858.

[2] Bakhshi, H., G. Kapetanios, and A. Yates, 2005, "Rational Expectations and Fixed-events Forecasts: An Application to UK Inflation," *Empirical Economics,* 30, 539-553.

[3] Batchelor, R. and P. Dua, 1991, "Blue Chip Rationality Tests," *Journal of Money, Credit and Banking*, 33, 692-705.

[4] Box, G.E.P., G.M. Jenkins, and G.C. Reinsel, 1994, *Time Series Analysis. Forecasting and Control*, 3rd edition. New Jersey: Prentice-Hall.

[5] Campbell, B. and Ghysels, E., 1995, "Federal Budget Projections: A Nonparametric Assessment of Bias and Efficiency," *Review of Economics and Statistics*, 17-31.

[6] Capistrán, C., 2006, "On Comparing Multi-Horizon Forecasts," *Economics Letters*, 93, 176-181.

[7] Chong, Y.Y. and D.F. Hendry, 1986, "Econometric Evaluation of Linear Macro-Economic Models," *Review of Economic Studies*, 53, 671-690.

[8] Christoffersen, P. and F.X. Diebold, 1997, "Optimal Prediction under Asymmetric Loss," *Econometric Theory,* 13, 806-817.

[9] Clements, M.P. and D.F. Hendry, 1998, *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

[10] Conover, W.J., 1980, *Practical Nonparametric Statistics*, 2nd edition. New York: Wiley & Sons.

[11] Croushore, D., 1993, "Introducing: the Survey of Professional Forecasters," *Business Review*, November/December, Federal Reserve Bank of Philadelphia, 3-13.

[12] Den Haan, W.J. and A. Levin, 1997, "A Practitioner's Guide to Robust Covariance Matrix Estimator," in G.S. Maddala and C.S. Rao (eds.), *Handbook of Statistics* 15, pp. 299-342. Amsterdam: North Holland.

[13] Diebold, F.X., 2001, *Elements of Forecasting.* 2nd edition. Ohio: South-Western.

[14] Diebold, F.X. and J.A. Lopez, 1996, "Forecast Evaluation and Combination," in G.S. Maddala and C.R. Rao (eds.), *Handbook of Statistics 14*, pp. 241-268. Amsterdam: North Holland.

[15] Diebold, F.X. and R. Mariano, 1995, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics,* 13, 253-263.

[16] Elliott, G. and A. Timmermann, 2004, "Optimal Forecast Combinations under General Loss Functions and Forecast Error Distributions," *Journal of Econometrics*, 122, 47-79.

[17] Elliott, G., I. Komunjer, and A. Timmermann, forhtcoming, "Biases in Macroeconomic Forecasts: Irrationality or Asymmetric Loss," *Journal of the European Economic Association.*

[18] Giacomini, R., and H. White, 2006, "Tests of Conditional Predictive Ability," *Econometrica,* 74, 1545-1578.

[19] Goldberger, A.S., 1992, "One-sided and Inequality Tests for a Pair of Means," in R. Bewley and T.V. Hoa (eds.), *Contributions to Consumer Demand and Econometrics*, pp. 140-162, New York: St. Martin's Press.

[20] Gourieroux, C. and A. Monfort, 1995, *Statistics and Econometric Models*, Volumes 2, (translated by Quang Voung from the 1989 French edition) Cambridge: Cambridge University Press.

[21] Granger, C.W.J., 1981, "The Comparison of Time Series and Econometric Forecasting Strategies," in J. Kmenta, and J.B. Ramsey (eds.), *Large-Scale Macro-Econometric Models*, pp. 123-128. North-Holland Publishing Company.

[22] Granger, C.W.J., 1969, "Prediction with a Generalized Cost Function," *Operational Research* 20, 199-207, reprinted in E. Ghysels, N.R. Swanson and M.W. Watson (eds.), *Essays in Econometrics: Collected Papers of Clive W.J. Granger,* Volume I, 2001. Cambridge: Cambridge University Press.

[23] Granger, C.W.J., 1999, "Outline of Forecast Theory using Generalized Cost Functions," *Spanish Economic Review,* 1, 161-173.

[24] Granger, C.W.J. and P. Newbold, 1986, *Forecasting Economic Time Series*, 2nd edition. Orlando: Academic Press.

[25] Granger, C.W.J. and T. Terasvirta, 1993, *Modelling Nonlinear Economic Relationships.* Oxford: Oxford University Press.

[26] Hamilton, J.D., 1994, *Time Series Analysis.* Princeton, N.J.: Princeton University Press.

[27] Hendry, D.F. and M.P. Clements, 2001, "Economic Forecasting: Some Lessons from Recent Research," *European Central Bank, Working Paper* No. 82.

[28] McCracken, M.W., 2000, "Robust Out-of-Sample Inference," *Journal of Econometrics*, 99, 195-223.

[29] Miller, R.G. Jr., 1980, *Simultaneous Statistical Inference.* 2nd edition. New York: Springer-Verlag.

[30] Mincer, J. and V. Zarnowitz (1969), "The Evaluation of Economic Forecasts" in J. Mincer (ed.) *Economic Forecasts and Expectations.* New York: National Bureau of Economic Research.

[31] Nordhaus, W.D., 1987, "Forecasting Efficiency: Concepts and Applications," *The Review of Economic and Statistics*, 69(4), 667-674.

[32] Patton, A.J. and A. Timmermann, forthcoming, "Testable Implications of Forecasts Optimality," *Journal of the American Statistical Association.*

[33] Patton, A.J. and A. Timmermann, 2007, "Properties of Optimal Forecasts under Asymmetric Loss and Nonlinearity" *Journal of Econometrics*, 140, 884-918.

[34] Roy, S.N., 1953, "On a Heuristic Method of Test Construction and its Use in Multivariate Analysis," *Annals of Mathematical Statistics*, 24, 220-238.

[35] Sarkar, S.K., and C-K. Chang, 1997, "The Simes Method for Multiple Hypothesis Testing with Positively Dependent Test Statistics." *Journal of the American Statistical Association*, 92, 1601-1608.

[36] Schmidt, P., 1974, "The Asymptotic Distribution of Forecasts in the Dynamic Simulation of an Econometric Model," *Econometrica*, 42, 303-309.

[37] Sen, P.K. and M.J. Silvapulle, 2002, "An Appraisal of Some Aspects of Statistical Inference under Inequality Constraints," *Journal of Statistical Planning and Inference*, 107, 3-43.

[38] Simes, R.J., 1986, "An Improved Bonferroni Procedure for Multiple Tests of Significance." *Biometrika*, 73, 751-754.

[39] Stekler, H.O., 2002, "The Rationality and Efficiency of Individuals' Forecasts," chapter 10 in M.P. Clements and D.F. Hendry. (eds), *A Companion to Economic Forecasting.* Oxford: Blackwell Publishers.

[40] Swidler, S. and D. Ketcher, 1990, "Economic Forecasts, Rationality, and the Processing of New Information over Time," *Journal of Money, Credit and Banking*, 22, 65-76.

[41] West, K.D., 1996, "Asymptotic Inference about Predictive Ability," *Econometrica,* 64, 1067-1084.

[42] White, H., 2000, "A Reality Check for Data Snooping," *Econometrica,* 68, 1097-1126.

[43] White, H., 2001, *Asymptotic Theory for Econometricians*, revised edition. San Diego: Academic Press.

[44] Wolak, F.A., 1987, "An Exact Test for Multiple Inequality Restrictions and Equality Constraints in the Linear Regression Model," *Journal of the American Statistical Association*, 82, 782-793.

[45] Zarnowitz, V., and P. Braun, 1992, "Twenty-Two Years of the NBER-ASA Quarterly Economic Outlook Surveys: Aspects and Comparison of Forecasts Performance," *NBER working paper* No. 3965.

[46] Zellner, A., 1986, "Biased Predictors, Rationality and the Evaluation of Forecasts," *Economics Letters*, 21, 45-48.

Appendix. Proofs

Proof of proposition 5. By optimality, $f^*_{\tau,\tau-h} \in \Omega_{\tau-h}$, and $f^*_{\tau,\tau-h-j} \in \Omega_{\tau-h-j}$ $\forall j > 0$. Since $\Omega_{\tau-h}$ is a filtration, $\Omega_{\tau-h-j} \subset \Omega_{\tau-h}$, so $f^*_{\tau,\tau-h-j} \in \Omega_{\tau-h}$. Then, if the loss function does not depend on time,

$$V(\tau, \tau-h, \Omega_{\tau-h}) \leq V(\tau, \tau-h-j, \Omega_{t-h}) \quad \forall j > 0,$$

and by the Law of Iterated Expectations,

$$E[V(\tau, \tau-h, \Omega_{\tau-h})] \leq E[V(\tau, \tau-h-j, \Omega_{t-h})] \quad \forall j > 0,$$

$$V(\tau, \tau-h) \leq V(\tau, \tau-h-j) \quad \forall j > 0.$$

■

Proof of proposition 6. Following Elliott and Timmermann (2004), a Taylor expansion around the mean forecasts, $\mu_e^{\tau-h} = E\left[e^*_{\tau,\tau-h}\right]$, gives:

$$\begin{aligned} L(e^*_{\tau,\tau-h}) = {} & L(\mu_e^{\tau-h}) + L'(\mu_e^{\tau-h})(e^*_{\tau,\tau-h} - \mu_e^{\tau-h}) \\ & + \frac{1}{2}L''(\mu_e^{\tau-h})(e^*_{\tau,\tau-h} - \mu_e^{\tau-h})^2 + \sum_{k=3}^{\infty} \frac{1}{k!} L^K(\mu_e^{\tau-h})(e^*_{\tau,\tau-h} - \mu_e^{\tau-h})^k, \end{aligned}$$

taking the mean,

$$\begin{aligned} E\left[L(e^*_{\tau,\tau-h})\right] = {} & L(\mu_e^{\tau-h}) + \frac{1}{2}L''(\mu_e^{\tau-h})E\left[(e^*_{\tau,\tau-h} - \mu_e^{\tau-h})^2\right] \\ & + \sum_{k=3}^{\infty} \frac{1}{k!} L^K E\left[(\mu_e^{\tau-h})(e^*_{\tau,\tau-h} - \mu_e^{\tau-h})^k\right]. \end{aligned}$$

When one of the two conditions stated in the proposition is met, the result collapses to:

$$E\left[L(e^*_{\tau,\tau-h})\right] = L(\mu_e^{\tau-h}) + \frac{1}{2}L''(\mu_e^{\tau-h})E\left[(e^*_{\tau,\tau-h} - \mu_e^{\tau-h})^2\right],$$

in this case, provided that the loss function is stable over time,

$$\begin{aligned} E\left[L(e^*_{\tau,\tau-h})\right] - E\left[L(e^*_{\tau,\tau-h-j})\right] = {} & L(\mu_e^{\tau-h}) - L(\mu_e^{\tau-h-j}) + \\ & \frac{1}{2}L''(\mu_e^{\tau-h})E\left[\left(e^*_{\tau,\tau-h} - \mu_e^{\tau-h}\right)^2\right] - \\ & \frac{1}{2}L''(\mu_e^{\tau-h-j})E\left[(e^*_{\tau,\tau-h-j} - \mu_e^{\tau-h-j})^2\right], \end{aligned}$$

further, if $\mu_e^{\tau-h} = \mu_e^{\tau-h-j}$,

$$E\left[L\left(e^*_{\tau,\tau-h}\right)\right] - E\left[L\left(e^*_{\tau,\tau-h-j}\right)\right] = \frac{1}{2}L''(\mu_e^{\tau-h})\left[E\left(e^*_{\tau,\tau-h} - \mu_e^{\tau-h}\right)^2 - E(e^*_{\tau,\tau-h-j} - \mu_e^{\tau-h-j})^2\right],$$

and the result from the proposition follows. ■

Proof of proposition 7. The result follows from the application of the Delta rule and the Slutsky theorem. ■

Proof of proposition 8. The result follows from the application of the Delta rule and the Slutsky theorem. ■

Table 1: Root Mean Squared Errors for different forecasters and horizons.

Forecaster	T	$RMSFE_1$	$RMSFE_2$	$RMSFE_3$	$RMSFE_4$
1	16	1.04	1.41	1.39	1.62
3	17	1.24	1.88	2.14	2.02
7	38	1.31	1.60	1.88	2.19
15	49	0.76	1.16	1.33	1.56
20	58	0.78	1.17	1.80	2.35
54	35	1.18	1.84	2.18	2.25
60	59	0.74	1.18	1.59	1.96
62	45	0.98	1.69	2.33	2.96
86	56	1.05	1.71	2.12	3.06
96	21	1.14	1.55	5.27	3.37
102	25	1.54	2.13	2.55	2.68
125	36	1.87	3.21	4.45	5.47
144	32	1.11	1.53	1.77	1.88
147	15	1.69	2.88	3.77	4.20
170	10	1.12	1.55	2.22	4.09
404	30	0.44	0.66	0.88	1.14
411	33	0.71	0.85	1.14	1.29
414	29	0.45	0.65	0.89	0.96
452	14	0.83	1.28	1.00	1.00
483	11	0.60	1.13	1.64	1.89

Notes: Sample of 20 forecasters from the Survey of Professional Forecasters. Original forecasts are for nominal output between 1968 and 2001. Shown are the RMSEs of 1 to 4 steps ahead forecasts for the annualized growth rate. T is the number of forecasts for each horizon.

Table 2: % of rejection of the true null when the $AR(1)$ is generated using Gaussian errors and the critical values are compared to a $N(0,1)$.

P	ϕ	d_1	d_3	d_7	r_1	r_3	r_7	s_1	s_3	s_7
50	0.20	2.36	9.78	10.44	2.52	9.52	10.86	4.20	10.00	9.94
50	0.50	0.06	4.76	8.00	0.08	4.60	7.84	0.70	6.84	9.76
50	0.80	0.00	0.26	1.00	0.00	0.26	0.94	0.08	1.62	5.54
100	0.20	1.18	9.40	10.38	1.14	9.10	10.36	2.50	9.44	9.86
100	0.50	0.00	2.56	6.70	0.00	2.44	6.60	0.12	5.00	9.18
100	0.80	0.00	0.00	0.18	0.00	0.02	0.14	0.00	0.46	4.02
500	0.20	0.00	9.16	9.76	0.00	9.16	9.74	0.18	9.28	10.12
500	0.50	0.00	0.14	3.66	0.00	0.12	3.66	0.00	1.84	9.04
500	0.80	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	1.06

Notes: The tests are at the 10% level. 5,000 Monte Carlo replications are performed. P is the number of forecasts available for each horizon. ϕ is the parameter in the $AR(1)$. d_h corresponds to the difference test for the difference between horizon h and $h+1$. r_h corresponds to the ratio test for the ratio between h and $h+1$. s_h corresponds to the sign test for the difference between h and $h+1$.

Table 3: % of rejection of the true null when the $AR(1)$ is generated using fat-tailed errors and the critical values are compared to a $N(0,1)$.

P	ϕ	d_1	d_3	d_7	r_1	r_3	r_7	s_1	s_3	s_7
50	0.20	2.30	10.78	9.86	2.50	10.24	9.68	3.76	10.24	9.38
50	0.50	0.08	4.12	7.60	0.08	4.24	7.20	0.46	6.20	9.92
50	0.80	0.00	0.18	0.80	0.00	0.14	0.68	0.04	1.12	5.82
100	0.20	1.28	9.02	10.34	1.46	8.78	10.40	2.30	9.44	9.46
100	0.50	0.00	2.64	6.28	0.00	2.70	6.40	0.06	5.10	8.26
100	0.80	0.00	0.00	0.12	0.00	0.00	0.08	0.00	0.48	4.62
500	0.20	0.04	8.12	10.22	0.04	8.14	10.12	0.20	8.88	9.88
500	0.50	0.00	0.18	3.68	0.00	0.18	3.50	0.00	1.62	9.24
500	0.80	0.00	0.00	0.08	0.00	0.00	0.08	0.00	0.00	1.24

Notes: As in Table 2.

Table 4: % of rejection of the true null when the $AR(1)$ is generated using Gaussian errors, the parameter (0.2) is estimated and the critical values are compared to a $N(0,1)$.

P	R	d_1	d_3	d_7	r_1	r_3	r_7	s_1	s_3	s_7
50	50	5.08	11.86	12.22	5.16	11.30	10.80	4.38	9.70	9.96
50	100	3.42	10.88	11.72	3.36	10.48	11.02	4.12	9.58	10.48
50	500	2.42	10.26	11.12	2.40	10.22	10.78	3.98	9.54	9.94
100	100	1.48	10.04	11.46	1.50	9.92	10.74	2.42	9.64	10.30
100	500	1.26	9.46	10.12	1.16	9.52	10.26	2.16	9.10	9.56
500	500	0.00	8.86	9.34	0.00	8.82	9.42	0.20	9.50	9.88

Notes: As in Table 2. R is the number of observations of the first sample used to estimate the $AR(1)$.

Table 5: % of rejection of the true null when the $AR(1)$ is generated using fat-tailed errors, the parameter (0.2) is estimated and the critical values are compared to a $N(0,1)$.

P	R	d_1	d_3	d_7	r_1	r_3	r_7	s_1	s_3	s_7
50	50	4.62	12.22	13.12	4.90	10.98	11.02	4.20	10.38	10.04
50	100	3.42	10.44	11.84	3.90	9.60	9.82	3.82	9.70	9.98
50	500	2.58	9.80	10.34	2.66	9.52	9.36	3.20	8.82	10.36
100	100	1.36	10.30	10.64	1.54	9.86	9.64	2.42	8.94	9.18
100	500	1.10	9.64	9.94	1.26	9.00	9.60	2.18	9.16	9.20
500	500	0.02	7.82	9.76	0.02	8.00	9.88	0.22	8.18	9.52

Notes: As in Table 4

Table 6: P-values, size and conclusions from the application of the t-stat and Simes's procedure to the apparently non-monotonic forecasters.

Component	d_j	p–value$_j$	α_j	Conclusion
		Forecaster 1		
$j = 1$	$RMSE_1$–$RMSE_2$	0.84	0.03	cannot reject
$j = 2$	$RMSE_2$–$RMSE_3$	0.47	0.10	cannot reject
$j = 3$	$RMSE_3$–$RMSE_4$	0.77	0.07	cannot reject
		Forecaster 3		
$j = 1$	$RMSE_1$–$RMSE_2$	0.98	0.07	cannot reject
$j = 2$	$RMSE_2$–$RMSE_3$	0.99	0.03	cannot reject
$j = 3$	$RMSE_3$–$RMSE_4$	0.41	0.10	cannot reject
		Forecaster 96		
$j = 1$	$RMSE_1$–$RMSE_2$	0.99	0.03	cannot reject
$j = 2$	$RMSE_2$–$RMSE_3$	0.86	0.07	cannot reject
$j = 3$	$RMSE_3$–$RMSE_4$	0.20	0.10	cannot reject
		Forecaster 452		
$j = 1$	$RMSE_1$–$RMSE_2$	0.98	0.07	cannot reject
$j = 2$	$RMSE_2$–$RMSE_3$	0.00	0.10	reject
$j = 3$	$RMSE_3$–$RMSE_4$	0.99	0.03	cannot reject