Span. Econ. Rev. 1, 161–173 (1999)

Spanish Economic Review

Outline of forecast theory using generalized cost functions

Clive W.J. Granger

Department of Economics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0508, USA (e-mail: cgranger@ucsd.edu)

Abstract. The cost functions used to form forecasts in practice may be quite different than the squared costs that is often assumed in forecast theory. The impact on evaluation procedures is determined and simple properties for the derivate of the cost function of the errors are found to provide simple tests of optimality. For a very limited class of situations are forecasts based on conditional means optimal, generally, the econometricians needs to provide the whole conditional predicted distribution. Implications for multi-step forecasts and the combination of forecasts are briefly considered.

JEL classification: C53, C32

Key words: Optimum forecasts, cost functions, predictive distribution

1. Basics

Suppose that, standing at time n, one is interested in forecasting the properties of the vector random variables $\underline{X}_{n+h}$ that consists of m components and whose value will be determined at time $n+h$. Thus, at least in theory, the value of $\underline{X}_{n+h}$ will be known at time $n+h$, where n is now and h is the forecast horizon. The information set that will be used to characterize the properties of $\underline{X}_{n+h}$ will be denoted I_n and will consist only of information available initially at time n and certainly not of information available at a later time. The choice of information is an important one and is discussed further in Sect. 5.

As $\underline{X}_{n+h}$ is a random variable, when viewed at time n as occurring in the future, it is best described by a distribution function and the appropriate one is its conditional cumulative distribution function, $\text{Prob}(X_{n+h} \leq x \mid I_n)$, which will

I would like to thank Mark Machina for his discussion of this paper which was prepared under NSF Grant SBR-0708615. I am also very grateful for the helpful comments by the referees.

sometimes be denoted $P_{n,h}(x)$ when I_n is known. For simplicity this will also be called the "predicted distribution function." Once $P_{n,h}(x)$ is known all the conditional properties of X_{n+h} given I_n are known, such as conditional moments, quantiles, and joint confidence intervals.

For now, just the case $m = 1$ will be considered so that X_{n+h} is univariate. Frequently it is convenient to have a point forecast $f_{n,h}$ of X_{n+h}, which is a single point that in some sense best represents the future occurring random variable. If $f_{n,h}$ is chosen, there will be a forecast error

$$e_{n,h} = X_{n+h} - f_{n,h} \ .$$

Note that $f_{n,h}$ is known at time n, but $e_{n,h}$ is not know until $n+h$. If decisions are based on the forecast, the fact that there are errors will mean that the decisions will be sub-optimal relative to the perfect forecastability situation and so costs will arise. If one assumes that these costs are a function of just the size and sign of the error, denoted $C(e)$, then one would expect that $C(o) = 0$, so that there is no cost if there is no error and the function will be non-decreasing as one moves away from $e = 0$ on each side, i.e., $C(e) > 0$, $e \neq 0$, $C'(e) \geq 0$, $e > 0$, $C'(e) \leq 0$, $e < 0$, where $C'(e)$ is the derivative, which will be assumed to exist for all e. Cost functions are discussed further in Sect. 3.

The optimum point forecast $f_{n,h}$ is chosen to minimize $E[C(X_{n+h} - f_{n,h} \mid I_n]$ i.e.

$$\min_{f_{n,h}} \int C(x - f_{n,h}) dP_{n,h}(x) \ .$$

Taking derivatives with respect to $f_{n,h}$ and assuming that everything is well behaved, gives the first order condition

$$\int C'(x - f_{n,h}) dP_{n,h}(x) = 0 \tag{1.1}$$

which has to be solved for $f_{n,h}$. Some examples in particular cases are discussed in Sect. 4.

2. Properties for the errors from optimum forecasts

At this stage, only the one-step forecast will be considered, so that $h = 1$, and the subscript h will be dropped. Suppose that the solution to (1.1), that is the optimal one-step point forecast using the cost function $C(e)$ based on I_n is $f^0_{n,h}$. The resulting forecast error $e^0_{n,1} = X_{n+1} - f^0_{n,1}$, with corresponding cost $C'(e^0_{n,1})$. Substituting into (1.1) gives

$$\int C'(e^0_{n,1}) d\overline{P}_n(e) = 0 \ , \tag{2.1}$$

where $\overline{P}_n(e)$ is the conditional distribution of $e_{n,1}$ which is immediately derived from $P_n(x)$, the conditional distribution of X_{n+1} given that $e_{n,1} = X_{n+1} - f^0_{n,1}$ and

$f^0_{n,1}$ is just a constant given I_n as it is a function of the contents of I_n. Define the random variable

$$Z_{n+1} = C'(e^0_{n,1}) \tag{2.2}$$

which is the key variable, as it derives from the minimization leading to (1.1). Now (2.1) says that

$$E[Z_{n+1} \mid I_n] = 0$$

and it immediately follows that

$$E[Z_{n+1}k(w_n) \mid I_n] = 0\ , \tag{2.3}$$

where w_n is any (finite) random variable constructed from the contents of I_n and $k(\,)$ is any (finite) function, as conditionally on $I_n, k(w_n)$ will appear to be a constant within the expectation. Consequently, if one formed a regression

$$Z_{n+1} = \alpha + \beta k(w_n) + \text{error} \tag{2.4}$$

one should expect both α and β to be zero if Z_{n+1} is really given by (2.2). As lagged $e^0_{n-j,1}$ will be in I_n, as will be past x_{n-j}, f_{n-j} values, it further follows that

$$\text{corr}(Z_{n+1}, Z_{n+1-j}) = 0 \quad \text{all} \quad j \neq 0 \tag{2.5}$$

so that the Z_n series will be zero mean, white noise. Further

$$\text{corr}(Z_{n+1}, e^0_{n-j,1}) = 0 \quad j > 0$$

and if one ran a regression

$$Z_{n+1} = \alpha + \beta_1 k_1(Z_n) + \beta_2 k_2(f_{n,1}) + \beta_3 k_3(X_n) + e_{n+1} \tag{2.6}$$

for any functions k_1, k_2, k_3 one should expect $\alpha = 0$ and all betas to be zero.

Some important generalizations are easily obtained. The results have been stated for a one step-horizon, $h = 1$, but hold for general horizon h in most cases. The one-step predictive distribution $P(X_{n+1} \mid I_n) = P_n(x) \equiv P_{n,1}(x)$ is used to find the two-step predicted distribution by noting that $X_{n+2} \mid I_n$ is $(X_{n+2} \mid I_{n+1}) \mid I_n$.

$$P(X_{n+2} \mid I_n) = E[P(X_{n+2} \mid I_{n+1}) \mid I_n] = P_{n,2}(x)$$

and similarly for general h.

The optimum forecast is denoted $f^0_{n,h}$ with corresponding error

$$e^0_{n,h} = X_{n+h} - f^0_{n,h}$$

and define $Z^{(h)}_{n+h} = C'(e^0_{n,h})$ from (2.2) extended. This variable will have conditional mean zero, and (2.3) becomes

$$E\left[Z^{(h)}_{n+h}k(w_n) \mid I_n\right] = 0\ . \tag{2.3'}$$

It should be noted that $e^0_{n,k}$ is not in I_n for $k = 1, 2, \ldots, h-1$ and so (2.5) becomes

$$\text{corr}\left(Z_{n+h+1}^{(h)}, Z_{n+h+1-j}^{(h)}\right) = 0 \quad \text{all} \quad j \geq h \tag{2.5'}$$

but is not necessarily zero for $j \leq h-1$. Thus $Z_{n+h}^{(k)}$ has the same autocorrelations as an $MA(h-1)$ process (which, of course, does not mean that it is such a process!).

The section has been considering forecasts of X_{n+h} but one might be interested in forecasting $Y_{n+h} = g(X_{n+h})$ for some given function g ( ), such as the log, exponential, or square functions. If the predicted cumulative distribution function of X_{n+h} is $P_{n,h}^{X}(x)$ then the corresponding predicted CDF of Y_{n+h} is

$$P_{n,h}^{Y} = P_{n,h}^{X}(q(x)) \,, \tag{2.7}$$

where $q(x) = g^{-1}(x)$ assuming $g(\,)$ is a one-to-one transformation over the range of X_t. Thus, if X_t is a positive process the inverse functions of log, exponential, and square are exponential, log and square root respectively. If there is no one-to-one correspondence, the relationship between $P^X(x)$ and $P^Y(y)$ still exists but is more complicated. Thus, at least in theory, the foregoing theory will apply to any well behaved function $Y_t = g(X_t)$ as one is now simply forecasting the mean of Y_t. However, it would be strange behavior to use the same cost function for X_{n+h} and Y_{n+h}.

Suppose that the cost function used to forecast $g(X_{n+1})$ is $C_g(e)$, then one needs to solve

$$\underset{fg}{\text{minimize}} \int C_g[g(X_{n+1}) - fg_{n,1}^{0}]dP_n(x) \,,$$

where $fg_{n,1}^{0}$ is the optimum one-step point forecast of $g(x_{n+1})$. The first-order equation is thus to solve

$$\int C_g{}'\left[g(X_{n+1}) - fg_{n,1}^{0}\right] dP_n(x) = 1$$

for $fg_{n,1}$. The one-step forecast error will be

$$eg_{n,1}^{0} = g(X_{n+1}) - fg_{n,1}^{0} \,.$$

Denoting $Zg_{n+1} = C_g{}'(eg_{n,1}^{0})$ then equations (2.3), (2.4), (2.5) still hold with Zg_{n+1} replacing Z_{n+1}.

3. Cost functions

Cost will arise if forecasting error results in sub-optimal decisions, and so functions of the form $C(e)$ will be considered. In fact, costs could depend on other variables or quantities, such as X_n, the value of the process, or the state of the economy, such as a measure of the business cycle, or time n, but just the simple form will be considered for convenience.

The properties required for a cost function are:

$C(0) = 0$ no error, no cost
$\min_e C(e) = 0$ So $C(e) \geq 0$
$C(e)$ is monotonic non-decreasing as e moves away from zero
i.e. $C(e_1) \geq C(e_2)$ if $e_1 > e_2 > 0$ and if $e_1 < e_2 < 0$.

There are three useful properties that a cost function may have:

Symmetry (property PS)
$C(e)$ is symmetric if $C(-e) = C(e)$.

Homogeneous (property PH)
$C(e)$ is homogeneous if $C(ae) = h(a)C(e)$ for some positive function $h(a)$.

Differentiability to order k (property PD_k)
If D is differentiation with respect to e, then $C(e)$ had PD_k if $D^h C(e)$ is possible for all $h \leq k$ and for all e in some specified range.

If a cost function is PD_k except at a single point, say $e = 0$, there will exist another positive function which is PD_k for all e which is arbitrarily close to it for all $e \neq 0$. In this case the original cost function will be considered to be PD_k. The problem of non-differentiability at a known point is just a technicality.

Some examples of cost functions, using the indicator function $I^+ = 1$ if $x \geq 0$, $= 0$ if $x < 0$ are:

$$(a) \quad (a + bI^+) \mid x \mid^{\theta}, \theta > 0, \quad a \geq 0, b \geq 0 \tag{3.1}$$

This group includes the familiar squared cost function, if $\theta = 2$, $b = 0$ and also the "lin-lin" function (according to Christoffersen and Diebold 1994) if $\theta = 1$. These functions are PD_k, for any k except at $e = 0$, they have the property PH and are PS if $b = 0$.

$$(b) \quad \text{Let } L(x, \alpha) = exp(\alpha x) - \alpha x - 1 \tag{3.2}$$

which is called the Linex function and was introduced by Varian (1974), and define the "double Linex cost function" by

$$C(e) = L(e, \alpha) + L(e, -\beta). \tag{3.3}$$

Note that $L(e, \alpha)$ is not symmetric and, if $\alpha > 0$, is exponential for e positive and linear in e for e negative. The double linex is not symmetric if $\alpha > 0$, $\beta > 0$, $\alpha \neq \beta$, but is exponential for both e positive and negative. If $\alpha = \beta$ the double linex is symmetric. This class is always PD_k, all k and all e but are not PH.

It is very easy to generate further examples of cost functions. If $C_1(e), C_2(e)$ are both cost functions then

(i) $\varphi_1(e) = aC_1(e) + bC_2(e)$, $a \geq 0$, $b \geq 0$ will be a cost function.
(ii) $\varphi_2(e) = [C_1(e)]^a [C_2(e)]^b a > 0$, $b > 0$ will be a cost function.
(iii) $\varphi_3(e) = I^+ C_1(e) + (1 - I^+) C_2(e)$ is a cost function, having $C_1(e)$ for positive e and $C_2(e)$ for negative.

(iv) If $\psi(e)$ is a positive monotonic, non-decreasing function on (o, ∞) with $\psi(o)$ finite, then

$$\varphi_4(e) = \psi(C(e)) - \psi(o)$$

is a cost function if $C(e)$ is a cost function. Note that $\psi(e)$ cannot be $\log e$, as $\psi(0)$ is infinite.

The overwhelming majority of forecast work uses the cost function $C(e) = ae^2$, $a > 0$, largely for mathematical convenience. In practice not a lot is known about cost functions but an assumption of symmetry is probably a poor one, as it is easy to think of examples of non-symmetric functions, the cost of arriving 10 minutes early for an airplane is different from arriving 10 minutes later; the cost of having a computer that is 10% too small for the typical task is different than being 10% too big; the cost of booking a lecture room that is 20 seats too big for your class is different from that of a room that is 20 seats too small, and so forth.

It is also implausible to use the same cost function for point forecasts of X_{n+1} and of $g(X_{n+1})$ where $g(\ )$ is some function, such as the log or the square, if one is interesting in forecasting a form of volatility, say, and yet this commonly occurs in academic reports, and in applied professional reports.

4. Some special cases of optimum forecasts

There are a number of special cases in which particular forms for the optimum point forecast f_n occur. The conditional distribution of X_{n+1} given I_n will be written $P(x, \mu_n, \alpha_n)$ where μ_n is the condition mean and α_n is the (vector of) other conditional parameters. The conditional distribution will be said to be symmetric about the mean if $\frac{dP}{dx}(x - \mu_n, o, \alpha_n) = \frac{dP}{dx}(-(x - \mu_n), o, \alpha_n)$ for all x where $\frac{dP}{dx}$ represents the probability density function corresponding to P. It was shown in Granger (1969) that if $C(e)$ is symmetric and the distribution $P(x)$ is symmetric about the mean μ_n, then $f_n = \mu_n$ provided *either*

(i) $C(e)$ is strictly monotonic increasing as e goes away from o (rather than just non-decreasing;

or

(ii) dP/dx is unimodal.

It is possible that other conditions will also ensure that the optimal forecast equals the conditional mean. The same paper gives an example showing that having both $C(e)$ and dP/dx symmetric is not sufficient to get this result.

It is possible for $f_n = \mu_n$ under weaker conditions, for example if $C(e) = ae^2$ then the optimal forecast is always μ_n regardless of the shape of the distribution or of the properties of α_n.

It is also often difficult to get specific results for particular cost functions. An exception is the "lin lin" function

$$\begin{aligned} C(e) &= (a+b)x \quad x > 0 \quad a \geq 0, b \geq 0 \\ &= a \mid x \mid \quad x < 0 \end{aligned}$$

which can be shown to result in the equation

$$P(f_n, \mu_n, \alpha_n) = (a + b)/(2a + b) \tag{4.1}$$

so that if $a = b$, f_n becomes the conditional median. This will, of course, be equal to the conditional mean if the distribution is symmetric.

For the symmetric double Linex function

$$C(e) = \exp(\theta e) + \exp(-\theta e) - 2$$

the solution is found to be

$$f_n = \mu_n + 1/2\theta \log(M_n(\theta)/M_n(-\theta)) \,, \tag{4.2}$$

where $M_n(\theta) = \int e^{\theta x} dP(x, o, \alpha_n)$ is the conditional central moment generating function and will be a function of α_n.

It is possible to get a few more results by assuming that X_{n+1} is a location/scale process, so that

$$X_{n+1} = \mu_n + \sigma_n \varepsilon_{n+1} \tag{4.3}$$

and the distribution function of ε_{n+1} is independent of I_n, so that the conditional distribution of ε_{n+1} on I_n equals the unconditional distribution. The equation being solved is then

$$\int C'(x - f_n) dP(x, \mu_n, \sigma_n) = 0$$

$$i.e. \qquad \int C'(\overline{x} - \overline{f}_n) \overline{d}P(x - \mu_n, o, \sigma_n) = 0 \,,$$

where $\overline{f}_n = f_n - \mu_n, \overline{x} = x - \mu_n$ and now when one solves this equation, necessarily

$$f_n = \mu_n + \theta(\sigma_n) \tag{4.4}$$

for some function $\theta(\ )$. If the cost function is homogeneous, one gets a more specific results, as the original cost being minimized is

$$\begin{aligned} C(X_{n+1} - f_n) &= C\left[\sigma_n \left(\frac{X_{n+1} - \mu_n}{\sigma_n} - \frac{f_n - \mu_n}{\sigma_n}\right)\right] \\ &= h(\sigma_n) C_1[\varepsilon_{n+1} - \tilde{f}_n] \end{aligned}$$

where $\tilde{f}_n = (f_n - \mu_n)/\sigma_n$. The equation to solve becomes

$$\int C_1'[\varepsilon - \tilde{f}_n] dP(\varepsilon)$$

and the solution is a constant, independent of I_n, depending just on the form of $C_1(e)$, so that

$$f_n = \mu_n + k\sigma_n \tag{4.5}$$

for some constant k.

It is interesting to note that if σ_n= constant, so that the process is homoskedastic, in many of these special cases the optimum forecast takes the form μ_n + constant.

Further progress can be made by assuming a particular distributional form for $P(x, \mu_n, \alpha_n)$ the typical assumption being normality. If X_{n+1} is Gaussian then it will be locational/scale as in (4.3) with $\varepsilon_{n+1} \sim N(0, 1)$ and so (4.4) will hold, as shown by Christoffersen and Diebold (1994). For example, they show that if one has a single Linex cost function (3.2) then

$$f_n = \mu_n + \alpha\sigma_n^2/2 \; . \tag{4.6}$$

Thus, in this case one adds a constant times the conditional variance rather than the conditional standard deviation as in (4.5).

In most cases it is not possible to derive a closed solution for the optimal forecast, but numerical solutions will certainly be available, as discussed by Christoffersen and Diebold (1994).

There are a number of important implications that can be derived from these results. Some can be stated in terms of the modeling strategy selected by the researcher in terms of the forecast provided. Only the one- step case will be considered. Two extremes will be:

(i) Provide an estimate of the complete predicted cumulative distribution function $P_n(x) \equiv \text{prob}(X_{n+1} \leq x \mid \text{i}_n)$ or its corresponding predicted probability function $p_n(x)$

$$\text{where } dP_n(x) = p_n(x)dx.$$

If this is provided, then point forecasts (and corresponding confidence intervals) can be derived for any function of the process and for any cost function, and in theory this can be extended to any horizon.

(ii) Provide an estimate only for the conditional mean μ_n. If one is interested only in forecasting X_{n+1} whilst using a squared cost function ae^2, then the optimum forecast is just μ_n. However, a forecast confidence interval is not available.

Between these two extremes are a variety of partial models, in which some aspects of $P_n(x)$ are modeled; others we merely assumed. An example is to assume X_{n+1} is a location/scale process (4.3), to then provide models or approximations for μ_n and σ_n and to then just assume that ε_{n+1} is iid N(0, 1). If actually ε_{n+1} is not Gaussian, this will be a partial model that may well approximate $P_n(x)$ in some features but not in others. A better strategy may be to check if $\frac{X_{n+1}-\hat{\mu}_n}{\hat{\sigma}_n}$ appears to be iid, independent of I_n and, if so, to find its distribution. The extent that mis-specification affects forecasts will depend on the type of partial model used and on the cost function involved. In all cases where the optimum forecast is μ_n, such as when the cost function is symmetric and monotonic and $P_n(x)$ is symmetric about μ_n, then a model that just produces the conditional mean will produce the optimal forecast. If the cost function is homogeneous then a model based on a location/scale assumption should produce a relevant point forecast. If the correct distribution of ε_{n+1} is given, confidence intervals

will also result. However, if a single or double linex cost function is used, then a complete specification is required.

Discussions in economic methodology often mention the ability of an economic theory to "predict" or forecast. Whereas this concept is not often related to ideas of evaluation, the relevant cost function does need consideration. This is particularly true when discussing how to evaluate rational expectations theory. For example, the result that (2.1) says

$$E[C'(e^o_{n,1}) \mid I_n] = 0$$

implies that the regression

$$C'(x_{n+1} - f_{n,1}) = a + bf_{n,1} + \varepsilon_{n+1}$$

will give $a = b = 0$ and ε_{t+1} a white noise. If $C(e) = e^2$, this translates into the regression

$$X_{n+1} = a + \beta f_{n,1} + \varepsilon_{n+1} \tag{4.7}$$

and one should expect $a = 0$, $\beta = 1$. This regression is often used as a test that "expectations are rational" (or "efficient" in an earlier terminology) but (4.7), with its constraints, will not necessarily hold for other cost functions.

5. Information set

The contents of the information set used to form forecasts is at the choice of the forecaster, and it is often an important choice. If I_n consists just of observed series, its contents can be denoted

$$I_n : X_{n-j}, W_{i,n-j}, i = 1, \ldots, l, j \geq 0 .$$

If I_n contains the past of the series being forecast it will be called "proper," and then I_n will effectively contain past forecast errors, which is important when considering the properties of errors from optimum forecasts.

If $l = 0$, so that I_n consists only of past and present X's, the forecasts are called univariate so that X_{n+h} is forecast just from its own past. Models using this information set are well developed and provide useful comparison forecasts against those arising from larger information sets. If $l > 0$, so that I_n includes other series, one has a multivariate forecast. As l is increased, one would expect to achieve better forecasts, measured in terms of lower expected cost particularly if one keeps adding relevant variables. The difficulty is knowing what variable is relevant and possibly omitting some particularly important variable or group of variables, although economic theory may be helpful in this respect.

If the average cost tends to zero as m becomes large, or perhaps even for a finite value of m, then the process X_t is (effectively) deterministic and thus perfectly forecastable, at least for the one step, $h = 1$, horizon. Whether or not one believes that it is possible to continually improve forecasts by adding further information is largely personal and depends on ones background. Most

statisticians and econometricians would not expect processes to be deterministic whereas physicists and meteorologists would take the opposite viewpoint, at least asymptotically as the sample size becomes very large.

6. Comparing forecasts

Let $P_n(x) \equiv P(x, \mu_n, \alpha_n)$ be the true predicted c.d.f. X_{n+1} given I_n, which leads to the forecast f_n if a cost function $C(e)$ is involved, as discussed in Sects. 1 and 2. Suppose that an alternative theory or model exists which proposes a c.d.f. $M_n(x)$ which is an approximation for $P_n(x)$. It will be convenient to write $M_n(x) \equiv M_n(x, \mu_n^*, \alpha_n^*)$. If one uses this approximation to minimize

$$\int C(X_{n+1} - f_n^*) dM_n$$

resulting in the forecast f_n^*, and thus forecast errors $e_n^* = X_{n+1} - f_n^*$ then these errors will have actual distribution $P(e, \mu_n - f_n^*, \alpha_n) \equiv P_n^*(e)$. The errors from the optimal forecasts f_n^0 will produce errors e_n^0 with distribution $P(e, \mu_n - f_n^0, \alpha_n) \equiv P_n(e)$. If forecasts f^0 and f^* are produced, the errors e^0 and e^* will be observed over time and their conditional cumulative histograms will eventually approximate $P_n(e)$ and $P_n^*(e)$ provided I_n can fall into just a limited number of "states." The obvious question becomes how one can compare these two cumulative distributions. A potentially important result is

Theorem 1.

$$E[C(e_n^0) \mid I_n] \leq E[C(e_n^*) \mid I_n] \tag{6.1}$$

for every cost function.

Proof. The result states

$$\int C\left(X_{n+1} - f_n^0\right) dP_n(e) \leq \int C\left(X_{n+1} - f_n^*\right) dP_n(e)$$

which holds obviously as f_n^0 is chosen to minimize the first integral and any quantity f_n^* will clearly produce costs that are larger, on average. The result immediately holds also for general horizon h and for the point forecast of any well behaved function $g(X_{n+h})$. Thus, the theorem says that, on average, point forecasts of any function of any horizon of the X_n process using any cost function are superior if one uses the true predicted c.d.f. rather than some approximation to it. The result is proved both in Granger and Pesaran (1996) and in Diebold et al. (1996) in the same way. A weakness of the result is that the theorem gives no indication of the amount of the benefit from using $P_n(e)$ rather than its approximation $P_n^*(e)$. It should be noted that the quality of the sub-optimal forecast errors have to be judged in terms of their actual distribution $P_n^*(e, \mu_n - f_n^*, \alpha_n)$ rather than the error distribution suggested by the theory $M_n(e, \mu_n^* - f_n^*, \alpha_n^*)$, i.e. $P_n^*(e)$ rather than $M_n(e)$. However, both of these quantities will,

potentially, be observable, and thus can be compared, whereas in practice $P_n(e)$ is not generally observable, or estimable. [Note that as e_t^* are observable so $P_n^*(e)$ can be estimated and $M_n(e)$ is the distribution of the errors predicted by the model in the post-sample.]

The theorem does not take into account the fact that the errors from a pair of forecasts of the same quantity are likely to be affected by the same outside events and so will be jointly distributed. Some on-going research involving the evaluation of forecasts using stochastic dominance concepts can make use of this information.

7. Some further results and generalization

7.1. Multi-step forecasting with general cost functions

To consider all costs of the form $E[C(x_{t+1} - f_t) \mid I_t]$ one needs the one-step predicted distribution $P(x_{t+1} \mid I_t)$ and similarly for costs occurring further into the future one needs the h-step predicted distribution $P(x_{t+h} \mid I_t)$. If there is general stationarity (not just covariance stationarity) the $P(x_{t+h} \mid I_t)$ will be similar to $P(x_{t+1} \mid I_{t-h+1})$ and this is easily determined from the one-step predicted density by marginalizing out the terms in I_t but not in I_{t-h+1}.

However, these predicted distributions will not be able to cope with cost functions such as $C(x_{t+k} - x_{t+k-1} - f_t)$ or most functions of two or more futures terms in the x series. For such cost functions one needs joint predicted distributions such as $P(x_{t+k}, x_{t+k-1} \mid I_t)$ and obvious generalizations. To cover all cases, one needs $P(x_{t+k}, k = 1, \ldots, M \mid I_t)$ for some appropriate M. This point has previously been made by Clements and Hendry (1993) but in a different form.

7.2. Multi-step forecasting with location-scale processes

Consider the processes generated by

$$X_t = \sum_{j=0}^{\infty} c_j \sigma_{t-1-j} \varepsilon_{t-j}, c_0 = 1$$

where σ_t is a process known at time t, and will be considered to be a function of the information set I_t. Further, suppose that ε_t is iid, independent of I_t. It follows that

$$X_{t+1} = \mu_t + \sigma_t \varepsilon_{t+1} ,$$

where $\mu_t = E[x_{t+1} \mid I_t] = \sum_{j=1}^{\infty} c_j \sigma_{t-j} \varepsilon_{t+1-j}$ and so the results of (4.4) and (4.5) hold for location-scale processes for $h = 1$. However, with general values of h one gets

$$f_{n,h} = E[x_{n+h} \mid I_n] = \sum_{j=h}^{\infty} c_n \sigma_{n+h-j} \varepsilon_{n+h+1-j}$$

so that $e_{n,h} = \sum_{j=0}^{h-1} c_j \sigma_{n+h+j} \varepsilon_{n+h+1-j}$.

Thus $x_{n+h} = f_{n,h} + e_{n,h}$. Now $e_{n,h}$ cannot generally be written in the form $S_n W_{n+h}$ where W_{n+h} is an iid process independent of I_n. If the ε's are normally distributed then x_{n+h} will be a location/scale process but this will not always be the case and so results (4.4) and (4.5) will thus not be true for general h even if they do apply for $h = 1$.

7.3. Combining forecasts with general costs functions

Suppose that two point forecasts are available each point of time t, f_t and g_t. We could ask if there is a simple combined forecast $k_t = \theta_1 f_t + \theta_2 g_t + \theta_3$ which would be superior to both. For a particular cost function C, if the optimum predicted distribution function is used to minimize the conditional expectation of $E[C(x_{t+1} - \theta_1 f_t - \theta_2 g_t - \theta_3) \mid I_t]$, then $\theta_1, \theta_2, \theta_3$ will be chosen as functions of I_t and the combined forecast will, generally, just become the better forecast. In particular, combining asks for constants θ_1, θ_2 so that the unconditional expectation of E[C] (or an estimate of it) is minimized, i.e.

$$\min_{\theta_1,\theta_2} \sum_t C(x_{t+1} - \theta_1 f_n - \theta_2 g_t - \theta_3) .$$

Taking derivatives with respect to the θ's and putting the results equal to zero, give three first-order conditions that have to be solved for the θ's, possibly using a numerical technique. Using these optimum values for the θ's, gives the errors from the combined forecast

$$ek_{t+1} = x_{t+1} - k_t^0$$

and with this notation, the first-order conditions gives

$$\begin{aligned} \sum_t C'(ek_{t+1}) &= 0 \\ \sum_t f_t C'(ek_{t+1}) &= 0 \\ \sum_t g_t C'(ek_{t+1}) &= 0 \end{aligned}$$

Thus, writing $Z_t = C'(ek_t)$ the estimated mean Z_t is zero, and the estimated correlations between Z_t and f_t and between Z_t and g_t are both zero. These are similar, but much weaker, forms of the results (2.3) and (2.5).

The discussion here has been for a specific linear combination, but a similar analysis can handle a particular function that combines forecasts, such as $\lambda(f_t, g_t, \varphi)$ when φ is a set of parameters to be determined by minimizing

$$\sum C(x_{t+1} - \lambda(f_t, g_t, \varphi) .$$

8. Brief literature review

An earlier paper on this topic was Granger (1964), which was greatly expanded in Christoffersen and Diebold (1994) which both preceded and overlaps with some of the results presented here. A more direct approach to evaluation using the whole predictive distribution has been discussed by West (1996) and by Diebold et al. (1996) An early specific contribution discussing asymmetric costs is by Zellner (1986) There is a large literature on combining forecasts; as example is given in the special issued on combining forecasts, *Journal of Forecasting* 8, no. 3, 1989.

References

1. Billingsley, P. (1986) *Probability and Measure*, 2nd edn. John Wiley and Sons, New York
2. Cristoffersen, P.F., Diebold, F.X. (1994) Optimal Prediction Under Asymmetric Loss. NBER Working Paper
3. Clements, M.P., Hendry, D.F. (1993) On The Limitations of Comparing Mean Squared Forecast Errors. *Journal of Forecasting* 12: 617–676 (including discussion)
4. Diebold, F.X., Gunther, T.A., Tay, A.S. (1996) Evaluating Density Forecasts. Working Paper, Department of Economics, University of Pennsylvania
5. Feller, W. (1971) *An Introduction To Probability and its Applications*, Volume II, 2nd edn. John Wiley and Sons, New York
6. Granger, C.W.J. (1969) Prediction With A Generalized Cost of Error Function. *Operations Research Quarterly* 20: 199–207
7. Granger, C.W.J., Pesaran, M.H. (1996) A Decision-Theoretic Approach to Forecast Evaluation. Working Paper 9618, Department of Applied Economics, University of Cambridge, England
8. Ingersoll, J.E. Jr. (1987) *Theory of Financial Decision Making.* Rowman and Littlefield, Totowa, NJ
9. Sin, C.-Y., Granger, C.W.J. (1994) Estimating and Forecasting Quantiles with Asymmetric Least Squares. Working Paper, Department of Economics, UCSD
10. West, K.D. (1996) Asymptotic Inference About Predictive Ability. *Econometrica* 64: 1067–1084
11. Zellner, A. (1986) Bayesian Estimation and Prediction Using Asymmetric Loss Functions. *Journal of Forecasting* 8: 446–451