Loss Functions in Time Series Forecasting

Tae-Hwy Lee
Department of Economics
University of California, Riverside
Riverside, CA 92521, USA
Phone (951) 827-1509
Fax (951) 827-5685
taelee@ucr.edu

March 2007

1 Introduction

The loss function (or cost function) is a crucial ingredient in all optimizing problems, such as statistical decision theory, policy making, estimation, forecasting, learning, classification, financial investment, etc. However, for space, the discussion here will be limited to the use of loss functions in econometrics, particularly in time series forecasting.

When a forecast $f_{t,h}$ of a variable Y_{t+h} is made at time t for h periods ahead, the loss (or cost) will arise if a forecast turns to be different from the actual value. The loss function of the forecast error $e_{t+h} = Y_{t+h} - f_{t,h}$, is denoted as $c(Y_{t+h}, f_{t,h})$. The loss function can depend on the time of prediction and so it can be $c_{t+h}(Y_{t+h}, f_{t,h})$. If the loss function is not changing with time and does not depend on the value of the variable Y_{t+h}, the loss can be written simply as a function of the error only, $c_{t+h}(Y_{t+h}, f_{t,h}) = c(e_{t+h})$.

Granger (1999) discusses the following required properties for a loss function: (i) $c(0) = 0$ (no error and no loss), (ii) $\min_e c(e) = 0$, so that $c(e) \geq 0$, (iii) $c(e)$ is monotonically non-decreasing as e moves away from zero so that $c(e_1) \geq c(e_2)$ if $e_1 > e_2 > 0$ and if $e_1 < e_2 < 0$.

When $c_1(e), c_2(e)$ are both loss functions, Granger (1999) shows that further examples of loss functions can be generated: (i) $c(e) = ac_1(e) + bc_2(e), a \geq 0, b \geq 0$ will be a loss function. (ii) $c(e) = c_1(e)^a c_2(e)^b$, $a > 0, b > 0$ will be a loss function. (iii) $c(e) = 1(e > 0)c_1(e) + 1(e < 0)c_2(e)$ will be a loss function. (iv) If $h(\cdot)$ is a positive monotonic non-decreasing function with $h(0)$ finite, then $c(e) = h(c_1(e)) - h(0)$ is a loss function.

2 Loss functions and risk

Granger (2002) notes that an expected loss (a risk measure) of financial return Y_{t+1} that has a conditional predictive distribution $F_t(y) \equiv \Pr(Y_{t+1} \leq y|I_t)$ with $\mathbf{X}_t \in I_t$ may be written as

$$\mathbb{E}c(e) = A_1 \int_0^\infty |y - f|^\theta dF_t(y) + A_2 \int_{-\infty}^0 |y - f|^\theta dF_t(y),$$

with A_1, A_2 both > 0 and some $\theta > 0$. Considering the symmetric case $A_1 = A_2$, one has a class of volatility measures $V_\theta = \mathbb{E}\left[|y - f|^\theta\right]$, which includes the variance with $\theta = 2$, and mean absolute deviation with $\theta = 1$.

Ding, Granger, and Engle (1993) study the time series and distributional properties of these measures empirically and show that the absolute deviations are found to have some particular properties such as the longest memory. Granger remarks that given that the financial returns are known to come from a long tail distribution, $\theta = 1$ may be more preferable.

Another problem raised by Granger is how to choose optimal L_p-norm in empirical works, to minimize $\mathbb{E}[|\varepsilon_t|^p]$ for some p to estimate the regression model $Y_t = X_t\beta + \varepsilon_t$. As the asymptotic covariance matrix of $\hat{\beta}$ depends on p, the most appropriate value of p can be chosen to minimize the covariance matrix. In particular, Granger (2002) refers to a trio of papers (Nyguist 1983, Money et al. 1982, Harter 1977) who find that the optimal $p = 1$ from Laplace and Cauchy distribution, $p = 2$ for Gaussian and $p = \infty$ (min/max estimator) for a rectangular distribution. Granger (2002) also notes that in terms of the kurtosis κ, Harter (1977) suggests to use $p = 1$ for $\kappa > 3.8$; $p = 2$ for $2.2 \leq \kappa \leq 3.8$; and $p = 3$ for $\kappa < 2.2$. In finance, the kurtosis of returns can be thought of as being well over 4 and so $p = 1$ is preferred.

We consider some variant loss functions with $\theta = 1, 2$ below.

3 Loss functions and regression functions

Optimal forecast of a time series model extensively depends on the specification of the loss function. Symmetric quadratic loss function is the most prevalent in applications due to its simplicity. The optimal forecast under quadratic loss is simply the conditional mean, but an asymmetric loss function implies a more complicated forecast that depends on the distribution of the forecast error as well as the loss function itself (Granger 1999), as the expected loss function if formulated with the expectation taken with respect to the conditional distribution. Specification of the loss function defines the model under consideration.

Consider a stochastic process $Z_t \equiv (Y_t, X_t')'$ where Y_t is the variable of interest and X_t is a vector of other variables. Suppose there are $T + 1$ ($\equiv R + P$) observations. We use the observations available at time t, $R \leq t < T + 1$, to generate P forecasts using each model. For each time t in the prediction period, we use either a rolling sample $\{Z_{t-R+1}, ..., Z_t\}$ of size R or the whole past sample $\{Z_1, ..., Z_t\}$ to estimate model

parameters $\hat{\beta}_t$. We can then generate a sequence of one-step-ahead forecasts $\{f(Z_t, \hat{\beta}_t)\}_{t=R}^{T}$.

Suppose that there is a decision maker who takes an one-step point forecast $f_{t,1} \equiv f(Z_t, \hat{\beta}_t)$ of Y_{t+1} and uses it in some relevant decision. The one-step forecast error $e_{t+1} \equiv Y_{t+1} - f_{t,1}$ will result in a cost of $c(e_{t+1})$, where the function $c(e)$ will increase as e increases in size, but not necessarily symmetrically or continuously. The optimal forecast $f_{t,1}^*$ will be chosen to produce the forecast errors that minimize the expected loss

$$\min_{f_{t,1}} \int_{-\infty}^{\infty} c(y - f_{t,1}) dF_t(y),$$

where $F_t(y) \equiv \Pr(Y_{t+1} \leq y | I_t)$ is the conditional distribution function, with I_t being some proper information set at time t that includes Z_{t-j}, $j \geq 0$. The corresponding optimal forecast error will be

$$e_{t+1}^* = Y_{t+1} - f_{t,1.}^*$$

Then the optimal forecast would satisfy

$$\frac{\partial}{\partial f_{t,1}} \int_{-\infty}^{\infty} c(y - f_{t,1}^*) dF_t(y) = 0.$$

When we may interchange the operations of differentiation and integration,

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial f_{t,1}} c(y - f_{t,1}^*) dF_t(y) \equiv \mathbb{E}\left(\frac{\partial}{\partial f_{t,1}} c(Y_{t+1} - f_{t,1}^*) | I_t \right)$$

the "generalized forecast error", $g_{t+1} \equiv \frac{\partial}{\partial f_{t,1}} c(Y_{t+1} - f_{t,1}^*)$, forms the condition of forecast optimality:

$$H_0 : \mathbb{E}\left(g_{t+1} | I_t\right) = 0 \qquad a.s.,$$

that is a martingale difference (MD) property of the generalized forecast error. This forms the optimality condition of the forecasts and gives an appropriate regression function corresponding to the specified loss function $c(\cdot)$.

To see this we consider the following two examples. First, when the loss function is the squared error loss

$$c(Y_{t+1} - f_{t,1}) = (Y_{t+1} - f_{t,1})^2,$$

the generalized forecast error will be $g_{t+1} \equiv \frac{\partial}{\partial f_t} c(Y_{t+1} - f_{t,1}^*) = -2e_{t+1}^*$ and thus $\mathbb{E}\left(e_{t+1}^* | I_t\right) = 0$ $a.s.$, which implies that the optimal forecast

$$f_{t,1}^* = \mathbb{E}\left(Y_{t+1} | I_t\right)$$

is the conditional mean. Next, when the loss is the check function, $c(e) = [\alpha - \mathbf{1}(e < 0)] \cdot e \equiv \rho_\alpha(e_{t+1})$, the optimal forecast $f_{t,1}$, for given $\alpha \in (0, 1)$, minimizing

$$\min_{f_{t,1}} \mathbb{E}\left[c(Y_{t+1} - f_{t,1}) | I_t\right]$$

can be shown to satisfy

$$\mathbb{E}\left[\alpha - \mathbf{1}(Y_{t+1} < f^*_{t,1})|I_t\right] = 0 \qquad a.s.$$

Hence, $g_{t+1} \equiv \alpha - \mathbf{1}(Y_{t+1} < f^*_{t,1})$ is the generalized forecast error. Therefore,

$$\alpha = \mathbb{E}\left[\mathbf{1}(Y_{t+1} < f^*_{t,1})|I_t\right] = \Pr(Y_{t+1} \leq f^*_{t,1}|I_t),$$

and the optimal forecast is $f^*_{t,1} = q_\alpha\left(Y_{t+1}|I_t\right)$ is the conditional α-quantile.

4 Loss functions for transformations

Granger (1999) note that it is implausible to use the same loss function for forecasting Y_{t+h} and for forecasting $h_{t+1} = h(Y_{t+h})$ where $h(\cdot)$ is some function, such as the log or the square, if one is interested in forecasting volatility. Suppose the loss functions $c_1(\cdot), c_2(\cdot)$ are used for forecasting Y_{t+h} and for forecasting $h(Y_{t+h})$, respectively. Let $e_{t+1} \equiv Y_{t+1} - f_{t,1}$ will result in a cost of $c_1(e_{t+1})$, for which the optimal forecast $f^*_{t,1}$ will be chosen from $\min_{f_{t,1}} \int_{-\infty}^{\infty} c_1(y - f_{t,1})dF_t(y)$, where $F_t(y) \equiv \Pr(Y_{t+1} \leq y|I_t)$. Let $\varepsilon_{t+1} \equiv h_{t+1} - h_{t,1}$ will result in a cost of $c_2(\varepsilon_{t+1})$, for which the optimal forecast $h^*_{t,1}$ will be chosen from $\min_{h_{t,1}} \int_{-\infty}^{\infty} c_2(h - h_{t,1})dH_t(h)$, where $H_t(h) \equiv \Pr(h_{t+1} \leq h|I_t)$. Then the optimal forecasts for Y and h would respectively satisfy

$$\begin{aligned}
\int_{-\infty}^{\infty} \frac{\partial}{\partial f_{t,1}} c_1(y - f^*_{t,1})dF_t(y) &= 0, \\
\int_{-\infty}^{\infty} \frac{\partial}{\partial h_{t,1}} c_2(h - h^*_{t,1})dH_t(h) &= 0.
\end{aligned}$$

It is easy to see that the optimality condition for $f^*_{t,1}$ does not imply the optimality condition for $h^*_{t,1}$ in general. Under some strong conditions on the functional forms of the transformation $h(\cdot)$ and of the two loss functions $c_1(\cdot), c_2(\cdot)$, the above two conditions may coincide. Granger (1999) remarks that it would be strange behavior to use the same loss function for Y and $h(Y)$. We leave this for further analysis in a future research.

5 Loss functions for asymmetry

The most prevalent loss function for the evaluation of a forecast is the symmetric quadratic function. Negative and positive forecast errors of the same magnitude have the same loss. This functional form is assumed because mathematically is very tractable but from an economic point of view, it is not very realistic. For a given information set and under a quadratic loss, the optimal forecast is the conditional mean of the variable under study. The choice of the loss function is fundamental to the construction of an optimal forecast. For asymmetric loss functions, the optimal forecast can be more complicated as it will depend not only on the

choice of the loss function but also on the characteristics of the probability density function of the forecast error (Granger, 1999).

As Granger (1999) notes the overwhelming majority of forecast work uses the cost function $c(e) = ae^2$, $a > 0$, largely for mathematical convenience. Asymmetric loss function is often relevant. A few examples from Granger (1999) are as follows: The cost of arriving 10 min early in the airport is quite different from arriving 10 min late. The cost of having a computer that is 10% too small for a task is different than being 10% too big. The loss of booking a lecture room that is 10 seats too big for your class is different from that of a room that is 10 seats too small. In dam construction an underestimate of the peak water is usually much more serious than an overestimate (Zellner 1986).

There are some commonly used asymmetric loss functions. The check loss function $c(y, f) \equiv [\alpha - \mathbf{1}(y < f)] \cdot (y - f)$, or $c(e) \equiv [\alpha - \mathbf{1}(e < 0)] \cdot e$, which makes the optimal predictor f the conditional quantile. The check loss function is also known as tick function or lil-lin loss. The asymmetric quadratic loss $c(e) \equiv [\alpha - \mathbf{1}(e < 0)] \cdot e^2$ can also be considered. A value of $\alpha = 0.5$ gives the symmetric squared error loss.

A particularly interesting asymmetric loss is the linex function of Varian (1975) that takes the following form

$$c_1(e, \alpha) = \exp(\alpha e_{t+1}) - \alpha e_{t+1} - 1,$$

where α is a scalar that controls the aversion towards either positive ($\alpha > 0$) or negative forecast errors ($\alpha < 0$). The linex function is differentiable. If $\alpha > 0$, the linex is exponential for $e > 0$ and linear for $e < 0$. If $\alpha < 0$, the linex is exponential for $e < 0$ and linear for $e > 0$. To make the linex more flexible, it can be modified to the "double linex loss function" by

$$\begin{aligned} c(e) &= c_1(e, \alpha) + c_1(e, -\beta), \qquad \alpha > 0, \beta > 0, \\ &= \exp(\alpha e) + \exp(-\beta e) - (\alpha - \beta)e - 2 \end{aligned}$$

which is exponential for all values of e (Granger 1999). When $\alpha = \beta$, it becomes the symmetric double linex loss function.

6 Loss functions for forecasting financial returns

Some simple examples of the loss function to evaluate the point forecasts of financial returns are the out-of-sample mean of the following loss functions as studied in Hong and Lee (2003): the squared error loss $c(y, f) = (y - f)^2$, absolute error loss $c(y, f) = |y - f|$, trading return $c(y, f) = -\text{sign}(f) \cdot y$ (when y is a financial asset return), and the correct direction $c(y, \hat{y}) = -\text{sign}(f) \cdot \text{sign}(y)$, where $\text{sign}(x) = \mathbf{1}(x > 0) - \mathbf{1}(x < 0)$ and $\mathbf{1}(\cdot)$ takes the value of 1 if the statement in the parenthesis is true and 0 otherwise. The negative

signs in the latter two is to make them the loss to minimize (rather than to maximize). The out-of-sample mean of these loss functions are the mean squared forecast errors (MSFE), mean absolute forecast errors (MAFE), mean forecast trading returns (MFTR), and mean correct forecast directions (MCFD):

$$\begin{aligned} MSFE &= P^{-1}\sum_{t=R}^{T}(Y_{t+1}-f_{t,1})^2, \\ MAFE &= P^{-1}\sum_{t=R}^{T}|Y_{t+1}-f_{t,1}|, \\ MFTR &= -P^{-1}\sum_{t=R}^{T}\text{sign}(f_{t,1})\cdot Y_{t+1}, \\ MCFD &= -P^{-1}\sum_{t=R}^{T}\mathbf{1}(\text{sign}(f_{t,1})\cdot\text{sign}(Y_{t+1})>0). \end{aligned}$$

These loss functions may further incorporate issues such as interest differentials, transaction costs and market depth. Because the investors are ultimately trying to maximize profits rather than minimize forecast errors, MSFE and MAFE may not be the most appropriate evaluation criteria. Granger (1999) emphasizes the importance of model evaluation using economic measures such as MFTR rather than statistical criteria such as MSFE and MAFE. Note that MFTR for the Buy-and-Hold trading strategy with $\text{sign}(f_{t,1})=1$ is the unconditional mean return of an asset because $MFTR^{\text{Buy\&Hold}}=-P^{-1}\sum_{t=R}^{T}Y_{t+1}\to-\mu$ in probability as $P\to\infty$, where $\mu\equiv\mathbb{E}(Y_t)$. MCFD is closely associated with an economic measure as it relates to market timing. Mutual fund managers, for example, can adjust investment portfolios in a timely manner if they can predict the directions of changes, thus earning a return higher than the market average.

7 Loss functions for estimation and evaluation

When the forecast is based on an econometric model, to the construction of the forecast, a model needs to be estimated. We often observe inconsistent choices of loss functions in estimation and forecasting. We may choose a symmetric quadratic objective function to estimate the parameters of the model but the evaluation of the model-based forecast may be based on an asymmetric loss function. This logical inconsistency is not inconsequential for tests assessing the predictive ability of the forecasts. The error introduced by parameter estimation affects the uncertainty of the forecast and, consequently, any test based on it. However, in applications, it is often the case that the loss function used for estimation of a model is different from the one(s) used in the evaluation of the model. This logical inconsistency can have significant consequences with regards to comparison of predictive ability of competing models. The uncertainty associated with parameter estimation may result in invalid inference of predictive ability (West 1996). When the objective function in estimation is the same as the loss function in forecasting the effect of parameter estimation vanishes. If

we believe that a particular criteria should be used to evaluate forecasts then it may also be used at the estimation stage of the modelling process. Gonzalez-Rivera, Lee, and Yoldas (2007) show this in the context of the VaR model of RiskMetrics, which provides a set of tools to measure market risk and eventually forecast the Value-at-Risk (VaR) of a portfolio of financial assets. A VaR is a quantile return. RiskMetrics offers a prime example in which the loss function of the forecaster is very well defined. They point out that a VaR is a quantile and thus the check loss function can be the objective function to estimate the parameters of the RiskMetrics model.

8 Loss function for binary forecast and maximum score

Given a series $\{Y_t\}$, consider the binary variable $G_{t+1} \equiv \mathbf{1}(Y_{t+1} > 0)$. We consider the asymmetric risk function to discuss a binary prediction. To define the asymmetric risk with $A_1 \neq A_2$ and $p = 1$, we consider binary decision problem of Granger and Pesaran (2000), Granger and Machina (2007), and Lee and Yang (2006) with the following 2×2 payoff or utility matrix:

Utility	$G_{t+1} = 1$	$G_{t+1} = 0$
$G_{t,1}(\mathbf{X}_t) = 1$	u_{11}	u_{01}
$G_{t,1}(\mathbf{X}_t) = 0$	u_{10}	u_{00}

where u_{ij} is the utility when $G_{t,1}(\mathbf{X}_t) = j$ is predicted and $G_{t+1} = i$ is realized $(i, j = 1, 2)$. Assume $u_{11} > u_{10}$ and $u_{00} > u_{01}$, and u_{ij} are constant over time. $(u_{11} - u_{10}) > 0$ is the utility gain from taking correct forecast when $G_{t,1}(\mathbf{X}_t) = 1$, and $(u_{00} - u_{01}) > 0$ is the utility gain from taking correct forecast when $G_{t,1}(\mathbf{X}_t) = 0$. Denote

$$\pi(\mathbf{X}_t) \equiv \mathbb{E}_{Y_{t+1}}(G_{t+1}|\mathbf{X}_t) = \Pr(G_{t+1} = 1|\mathbf{X}_t).$$

The expected utility of $G_{t,1}(\mathbf{X}_t) = 1$ is $u_{11}\pi(\mathbf{X}_t) + u_{01}(1 - \pi(\mathbf{X}_t))$, and the expected utility of $G_{t,1}(\mathbf{X}_t) = 0$ is $u_{10}\pi(\mathbf{X}_t) + u_{00}(1 - \pi(\mathbf{X}_t))$. Hence, to maximize utility, conditional on the values of $\mathbf{X}_t$, the prediction $G_{t,1}(\mathbf{X}_t) = 1$ will be made if

$$u_{11}\pi(\mathbf{X}_t) + u_{01}(1 - \pi(\mathbf{X}_t)) > u_{10}\pi(\mathbf{X}_t) + u_{00}(1 - \pi(\mathbf{X}_t)),$$

or

$$\pi(\mathbf{X}_t) > \frac{(u_{00} - u_{01})}{(u_{11} - u_{10}) + (u_{00} - u_{01})} \equiv 1 - \alpha.$$

By making correct prediction, our net utility gain is $(u_{00} - u_{01})$ when $G_{t+1} = 0$, and $(u_{11} - u_{10})$ when $G_{t+1} = 1$. We can put it in another way, our opportunity cost (in the sense that you lose the gain) of wrong prediction is $(u_{00} - u_{01})$ when $G_{t+1} = 0$ and $(u_{11} - u_{10})$ when $G_{t+1} = 1$. Since a multiple of a utility function represents the same preference, $(1 - \alpha)$ can be viewed as the utility-gain from correct prediction

when $G_{t+1} = 0$, or the opportunity cost of a false-alert. Similarly,

$$\alpha \equiv \frac{(u_{11} - u_{10})}{(u_{11} - u_{10}) + (u_{00} - u_{01})}$$

can be treated as the utility-gain from correct prediction when $G_{t+1} = 1$ is realized, or the opportunity cost of a failure-to-alert. We thus can define a cost function $c(e_{t+1})$ with $e_{t+1} = G_{t+1} - G_{t,1}(\mathbf{X}_t)$:

Cost	$G_{t+1} = 1$	$G_{t+1} = 0$
$G_{t,1}(\mathbf{X}_t) = 1$	0	$1 - \alpha$
$G_{t,1}(\mathbf{X}_t) = 0$	α	0

That is

$$c(e_{t+1}) = \begin{cases} \alpha & \text{if} \quad e_{t+1} = \quad 1 \\ 1-\alpha & \text{if} \quad e_{t+1} = -1 \\ 0 & \text{if} \quad e_{t+1} = \quad 0 \end{cases},$$

which can be equivalently written as $c(e_{t+1}) = \rho_\alpha(e_{t+1})$, where $\rho_\alpha(e) \equiv [\alpha - \mathbf{1}(e < 0)]e$ is the check function.

Hence, the optimal binary predictor $G^\dagger_{t,1}(X_t) = 1(\pi(X_t) > 1 - \alpha)$ maximizing the expected utility minimizes the expected cost $E(\rho_\alpha(e_{t+1})|X_t)$. A general result on the utility functions and the loss functions is derived in Granger and Machina (2007).

The optimal binary prediction that minimizes $\mathbb{E}_{Y_{t+1}}(\rho_\alpha(e_{t+1})|\mathbf{X}_t)$ is the conditional α-quantile of G_{t+1}, denoted as

$$G^\dagger_{t,1}(\mathbf{X}_t) = Q^\dagger_\alpha(G_{t+1}|\mathbf{X}_t) = \arg \min_{G_{t,1}(\mathbf{X}_t)} \mathbb{E}_{Y_{t+1}}(\rho_\alpha(G_{t+1} - G_{t,1}(\mathbf{X}_t))|\mathbf{X}_t).$$

This is a maximum score problem of Manski (1975).

Also, as noted by Powell (1986), using the fact that for any monotonic function $h(\cdot)$, $Q_\alpha(h(Y_{t+1})|\mathbf{X}_t) = h(Q_\alpha(Y_{t+1}|\mathbf{X}_t))$, which follows immediately from observing that $\Pr(Y_{t+1} < y|\mathbf{X}_t) = \Pr[h(Y_{t+1}) < h(y)|\mathbf{X}_t]$, and noting that the indicator function is monotonic, $Q_\alpha(G_{t+1}|\mathbf{X}_t) = Q_\alpha(\mathbf{1}(Y_{t+1} > 0)|\mathbf{X}_t) = \mathbf{1}(Q_\alpha(Y_{t+1}|\mathbf{X}_t) > 0)$. Hence,

$$G^\dagger_{t,1}(\mathbf{X}_t) = \mathbf{1}(Q^\dagger_\alpha(Y_{t+1}|\mathbf{X}_t) > 0).$$

where $Q_\alpha(Y_{t+1}|\mathbf{X}_t)$ is the α-quantile function of Y_{t+1} conditional on $\mathbf{X}_t$. Note that $Q^\dagger_\alpha(G_{t+1}|\mathbf{X}_t) = \arg\min \mathbb{E}_{Y_{t+1}}(\rho_\alpha(e_{t+1})|\mathbf{X}_t)$ with $e_{t+1} \equiv G_{t+1} - Q_\alpha(G_{t+1}|\mathbf{X}_t)$, and $Q^\dagger_\alpha(Y_{t+1}|\mathbf{X}_t) = \arg\min \mathbb{E}_{Y_{t+1}}(\rho_\alpha(u_{t+1})|\mathbf{X}_t)$ with $u_{t+1} \equiv Y_{t+1} - Q_\alpha(Y_{t+1}|\mathbf{X}_t)$. Therefore, the optimal binary prediction can be made from binary quantile regression for G_{t+1}. Binary prediction can also be made from a binary function of the α-quantile for Y_{t+1}.

9 Loss functions for probability forecasts

Diebold and Rudebush (1989) consider the probability forecasts for business cycle turning points. To measure the accuracy of predicted probabilities, that is the average distance between the predicted probabilities and observed realization (as measured by a zero-one dummy variable). Suppose we have time series of P

probability forecast $\{p_t\}_{t=R+1}^{T}$ where p_t is the probability of the occurrence of a turning point at date t. Let $\{d_t\}_{t=R+1}^{T}$ be the corresponding realization with $d_t = 1$ if a business cycle turning point (or any defined event) occurs in period t and $d_t = 0$ otherwise. The loss function analogous to the squared error is Brier's score based on quadratic probability score (QPS):

$$QPS = P^{-1} \sum_{t=R}^{T} 2(p_t - d_t)^2.$$

The QPS ranges from 0 to 2, with 0 for perfect accuracy. As noted by Diebold and Rudebush (1989), the use of the symmetric loss function may not be appropriate as a forecaster may be penalized more heavily for missing a call (making a type II error) than for signaling a false alarm (making a type I error). Another loss function is given by the log probability score (LPS)

$$LPS = -P^{-1} \sum_{t=R}^{T} \ln \left(p_t^{d_t} (1 - p_t)^{(1-d_t)} \right),$$

which is similar to the loss for the interval forecast. A large mistakes are penalized more heavily under LPS than under QPS. More loss functions are discussed in Diebold and Rudebush (1989).

Another loss function useful in this context is the Kuipers score (KS), which is defined by

$$KS = \text{Hit Rate} - \text{False Alarm Rate},$$

where Hit Rate is the fraction of the bad events that were correctly predicted as good events (power, or 1−probability of type II error), and False Alarm Rate is the fraction of good events that had been incorrectly as bad events (probability of type I error).

10 Loss function for interval forecasts

Suppose Y_t is a stationary series. Let the one-period ahead conditional interval forecast made at time t from a model be denoted as

$$J_{t,1}(\alpha) = (L_{t,1}(\alpha), U_{t,1}(\alpha)), \qquad t = R, \ldots, T,$$

where $L_{t,1}(\alpha)$ and $U_{t,1}(\alpha)$ are the lower and upper limits of the ex ante interval forecast for time $t+1$ made at time t with the coverage probability α. Define the indicator variable $X_{t+1}(\alpha) = \mathbf{1}[Y_{t+1} \in J_{t,1}(\alpha)]$. The sequence $\{X_{t+1}(\alpha)\}_{t=R}^{T}$ is IID Bernoulli (α). The optimal interval forecast would satisfy $\mathbb{E}(X_{t+1}(\alpha)|I_t) = \alpha$, so that $\{X_{t+1}(\alpha) - \alpha\}$ will be an MD. A better model has a larger expected Bernoulli log-likelihood

$$\mathbb{E}\alpha^{X_{t+1}(\alpha)}(1-\alpha)^{[1-X_{t+1}(\alpha)]}.$$

Hence, we can choose a model for interval forecasts with the smallest out-of-sample mean of the negative predictive log-likelihood defined by

$$-P^{-1}\sum_{t=R}^{T}\ln\left(\alpha^{x_{t+1}(\alpha)}(1-\alpha)^{[1-x_{t+1}(\alpha)]}\right).$$

11 Loss function for density forecasts

Consider a financial return series $\{y_t\}_{t=1}^T$. This observed data on a univariate series is a realization of a stochastic process $\mathbf{Y}^T \equiv \{Y_\tau : \Omega \to \mathbb{R},\ \tau = 1, 2, \ldots, T\}$ on a complete probability space $(\Omega, \mathcal{F}_T, P_0^T)$, where $\Omega = \mathbb{R}^T \equiv \times_{\tau=1}^T \mathbb{R}$ and $\mathcal{F}_T = \mathcal{B}(\mathbb{R}^T)$ is the Borel σ-field generated by the open sets of $\mathbb{R}^T$, and the *joint* probability measure $P_0^T(B) \equiv P_0[\mathbf{Y}^T \in B]$, $B \in \mathcal{B}(\mathbb{R}^T)$ completely describes the stochastic process. A sample of size T is denoted as $\mathbf{y}^T \equiv (y_1, \ldots, y_T)'$.

Let σ-finite measure ν^T on $\mathcal{B}(\mathbb{R}^T)$ be given. Assume $P_0^T(B)$ is absolutely continuous with respect to ν^T for all $T = 1, 2, \ldots$, so that there exists a measurable Radon-Nikodým density $g^T(\mathbf{y}^T) = dP_0^T/d\nu^T$, unique up to a set of zero measure-ν^T.

Following White (1994), we define a probability model $\mathcal{P}$ as a collection of distinct probability measures on the measurable space $(\Omega, \mathcal{F}_T)$. A probability model $\mathcal{P}$ is said to be correctly specified for $\mathbf{Y}^T$ if $\mathcal{P}$ contains P_0^T. Our goal is to evaluate and compare a set of parametric probability models $\{P_\theta^T\}$, where $P_\theta^T(B) \equiv P_\theta[Y^T \in B]$. Suppose there exists a measurable Radon-Nikodým density $f^T(\mathbf{y}^T) = dP_\theta^T/d\nu^T$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\theta}$ is a finite-dimensional vector of parameters and is assumed to be identified on $\boldsymbol{\Theta}$, a compact subset of $\mathbb{R}^k$. See White (1994, Theorem 2.6).

In the context of forecasting, instead of the joint density $g^T(\mathbf{y}^T)$, we consider forecasting the *conditional* density of $\mathbf{Y}^t$, given the information $\mathcal{F}_{t-1}$ generated by $\mathbf{Y}^{t-1}$. Let $\varphi_t(y_t) \equiv \varphi_t(y_t|\mathcal{F}_{t-1}) \equiv g^t(\mathbf{y}^t)/g^{t-1}(\mathbf{y}^{t-1})$ for $t = 2, 3, \ldots$ and $\varphi_1(y_1) \equiv \varphi_1(y_1|\mathcal{F}_0) \equiv g^1(\mathbf{y}^1) = g^1(y_1)$. Thus the goal is to forecast the (true, unknown) conditional density $\varphi_t(y_t)$.

For this, we use an one-step-ahead conditional density forecast model $\psi_t(y_t;\boldsymbol{\theta}) \equiv \psi_t(y_t|\mathcal{F}_{t-1};\boldsymbol{\theta}) \equiv f^t(\mathbf{y}^t)/f^{t-1}(\mathbf{y}^{t-1})$ for $t = 2, 3, \ldots$ and $\psi_1(y_1) \equiv \psi_1(y_1|\mathcal{F}_0) \equiv f^1(\mathbf{y}^1) = f^1(y_1)$. If $\psi_t(y_t;\boldsymbol{\theta}_0) = \varphi_t(y_t)$ almost surely for some $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, then the one-step-ahead density forecast is correctly specified, and it is said to be optimal because it dominates all other density forecasts for any loss functions as discussed in the previous section (see Granger and Pesaran, 2000a, 2000b; Diebold *et al.*, 1998; Granger 1999).

In practice, it is rarely the case that we can find an optimal model. As it is very likely that "the true distribution is in fact too complicated to be represented by a simple mathematical function" (Sawa, 1978), all the models proposed by different researchers can be possibly misspecified and thereby we regard each

model as an approximation to the truth. Our task is then to investigate which density forecast model can approximate the true conditional density most closely. We have to first define a metric to measure the distance of a given model to the truth, and then compare different models in terms of this distance.

The adequacy of a density forecast model can be measured by the conditional Kullback-Leibler (1951) Information Criterion (KLIC) divergence measure between two conditional densities,

$$\mathbb{I}_t\left(\varphi:\psi,\boldsymbol{\theta}\right)=\mathbb{E}_{\varphi_t}[\ln\varphi_t\left(y_t\right)-\ln\psi_t\left(y_t;\boldsymbol{\theta}\right)],$$

where the expectation is with respect to the true conditional density $\varphi_t\left(\cdot|\mathcal{F}_{t-1}\right)$, $\mathbb{E}_{\varphi_t}\ln\varphi_t\left(y_t|\mathcal{F}_{t-1}\right)<\infty$, and $\mathbb{E}_{\varphi_t}\ln\psi_t\left(y_t|\mathcal{F}_{t-1};\boldsymbol{\theta}\right)<\infty$. Following White (1994), we define the distance between a density model and the true density as the minimum of the KLIC

$$\mathbb{I}_t\left(\varphi:\psi,\boldsymbol{\theta}_{t-1}^*\right)=\mathbb{E}_{\varphi_t}\left[\ln\varphi_t\left(y_t\right)-\ln\psi_t\left(y_t;\boldsymbol{\theta}_{t-1}^*\right)\right],$$

where $\boldsymbol{\theta}_{t-1}^*=\arg\min\mathbb{I}_t\left(\varphi:\psi,\boldsymbol{\theta}\right)$ is the pseudo-true value of $\boldsymbol{\theta}$ (Sawa, 1978). We assume that $\boldsymbol{\theta}_{t-1}^*$ is an interior point of $\boldsymbol{\Theta}$. The smaller this distance is, the closer the density forecast $\psi_t\left(\cdot|\mathcal{F}_{t-1};\boldsymbol{\theta}_{t-1}^*\right)$ is to the true density $\varphi_t\left(\cdot|\mathcal{F}_{t-1}\right)$.

However, $\mathbb{I}_t\left(\varphi:\psi,\boldsymbol{\theta}_{t-1}^*\right)$ is unknown since $\boldsymbol{\theta}_{t-1}^*$ is not observable. We need to estimate $\boldsymbol{\theta}_{t-1}^*$. If our purpose is to compare the out-of-sample predictive abilities among competing density forecast models, we split the data into two parts, one for estimation and the other for out-of-sample validation. At each period t in the out-of-sample period ($t=R+1,\ldots,T$), we estimate the unknown parameter vector $\boldsymbol{\theta}_{t-1}^*$ and denote the estimate as $\hat{\boldsymbol{\theta}}_{t-1}$. Using $\{\hat{\boldsymbol{\theta}}_{t-1}\}_{t=R+1}^T$, we can obtain the out-of-sample estimate of $\mathbb{I}_t\left(\varphi:\psi,\boldsymbol{\theta}_{t-1}^*\right)$ by

$$\mathbb{I}_P(\varphi:\psi)\equiv\frac{1}{P}\sum_{t=R+1}^{T}\ln[\varphi_t(y_t)/\psi_t(y_t;\hat{\boldsymbol{\theta}}_{t-1})]$$

where $P=T-R$ is the size of the out-of-sample period. Note that

$$\mathbb{I}_P(\varphi:\psi)=\frac{1}{P}\sum_{t=R+1}^{T}\ln\left[\varphi_t(y_t)/\psi_t\left(y_t;\boldsymbol{\theta}_{t-1}^*\right)\right]+\frac{1}{P}\sum_{t=R+1}^{T}\ln[\psi_t\left(y_t;\boldsymbol{\theta}_{t-1}^*\right)/\psi_t(y_t;\hat{\boldsymbol{\theta}}_{t-1})],$$

where the first term in $\mathbb{I}_P(\varphi:\psi)$ measures model uncertainty (the distance between the optimal density $\varphi_t(y_t)$ and the model $\psi_t\left(y_t;\boldsymbol{\theta}_{t-1}^*\right)$) and the second term measures parameter estimation uncertainty due to the distance between $\boldsymbol{\theta}_{t-1}^*$ and $\hat{\boldsymbol{\theta}}_{t-1}$.

Since the KLIC measure takes on a smaller value when a model is closer to the truth, we can regard it as a loss function and use $\mathbb{I}_P(\varphi:\psi)$ to formulate the loss-differential. The out-of-sample average of the loss-differential between model 1 and model 2 is

$$\mathbb{I}_P(\varphi:\psi^1)-\mathbb{I}_P(\varphi:\psi^2)=\frac{1}{P}\sum_{t=R+1}^{T}\ln[\psi_t^2(y_t;\hat{\boldsymbol{\theta}}_{t-1}^2)/\psi_t^1(y_t;\hat{\boldsymbol{\theta}}_{t-1}^1)],$$

which is the ratio of the two predictive log-likelihood functions. With treating model 1 as a benchmark model (for model selection) or as the model under the null hypothesis (for hypothesis testing), $\mathbb{I}_P(\varphi : \psi^1) - \mathbb{I}_P(\varphi : \psi^2)$ can be considered as a loss function to minimize. To sum up, the KLIC differential can serve as a *loss* function for density forecast evaluation as discussed in Bao, Lee, and Saltoglu (2007).

12 Loss functions for volatility forecasts

González-Rivera, Lee, and Santosh Mishra (2004) analyze the predictive performance of various volatility models for stock returns. To compare the performance, they choose loss functions for which volatility estimation is of paramount importance. They deal with two economic loss functions (an option pricing function and an utility function) and two statistical loss functions (the check loss for a Value-at-Risk calculation and a predictive likelihood function of the conditional variance).

13 Loss functions for testing Granger-causality

In time series forecasting, a concept of causality is due to Granger (1969), who defined in terms of conditional distribution. Lee and Yang (2007) use loss functions to test for Granger-causality in conditional mean, in conditional distribution, and in conditional quantiles. The causal relationship between money and income (output) has been an important topic that has been extensively studied. However, those empirical studies are almost entirely on Granger-causality in the conditional mean. Compared to conditional mean, conditional quantiles give a broader picture of a variable in various scenarios. Lee and Yang (2007) explore whether forecasting the conditional quantile of output growth may be improved using money. They compare the check (tick) loss functions of the quantile forecasts of output growth with and without using the past information on money growth, and assess the statistical significance of the loss-differential of the unconditional and conditional predictive abilities. As conditional quantiles can be inverted to the conditional distribution, they also test for Granger-causality in the conditional distribution (via using a nonparametric copula function). Using U.S. monthly series of real personal income and industrial production for income, and M1 and M2 for money, for 1959-2001, they find that out-of-sample quantile forecasting for output growth, particularly in tails, is significantly improved by accounting for money. On the other hand, money-income Granger-causality in the conditional mean is quite weak and unstable. Their results have an important implication on monetary policy, showing that the effectiveness of monetary policy has been underestimated by merely testing Granger-causality in mean. Money does Granger-cause income more strongly than it has been known and therefore the information on money growth can (and should) be more utilized in implementing monetary policy.

14 References

Bao, Y., T.-H. Lee, and B. Saltoglu (2007), "Comparing Density Forecast Models", *Journal of Forecasting*, forthcoming.

Diebold, F. X., T.A. Gunther and A.S. Tay (1998). Evaluating Density Forecasts with Applications to Financial Risk Management, *International Economic Review,* 39, 863-883.

Diebold, F.X. and G.D. Rudebusch (1989), "Scoring the Leading Indicators", *Journal of Business,* 62(3) 369-391.

Ding, Z., C.W.J. Granger and R. Engle (1993), "A Long Memory Property of Stock Market Returns and a New Model", *Journal of Empirical Finance*, 1, 83-106.

González-Rivera, G., T.-H. Lee, and S. Mishra (2004), "Forecasting Volatility: A Reality Check Based on Option Pricing, Utility Function, Value-at-Risk, and Predictive Likelihood", *International Journal of Forecasting*, 20(4), 629-645

González-Rivera, G., T.-H. Lee, and E. Yoldas (2007), "Optimality of the RiskMetrics VaR Model", UCR.

Granger, C. W. J. (1969), "Investigating Causal Relations by econometric Models and Cross-Spectral Methods", *Econometrica,* 37, 424-438.

Granger, C.W.J. (1999), "Outline of Forecast Theory Using Generalized Cost Functions", *Spanish Economic Review,* 1, 161-173.

Granger, C.W.J. (2002), "Some Comments on Risk", *Journal of Applied Econometrics*, 17, 447-456.

Granger, C.W.J. and M.J. Machina (2007), "Forecasting and Decision Theory", *Handbook of Econometric Forecasting*

Granger, C.W.J. and M.H. Pesaran (2000a), "A Decision Theoretic Approach to Forecasting Evaluation", in *Statistics and Finance: An Interface*, W. S. Chan, W. K. Li, and Howell Tong (eds.), London: Imperial College Press.

Granger, C.W.J. and M.H. Pesaran (2000b), "Economic and Statistical Measures of Forecast Accuracy", *Journal of Forecasting,* 19, 537-560.

Harter, H.L. (1977), "Nonuniqueness of Least Absolute Values Regression", *Communications in Statistics - Theory and Methods,* A6, 829-838.

Hong, Y. and T.-H. Lee (2003), "Inference on Predictability of Foreign Exchange Rates via Generalized Spectrum and Nonlinear Time Series Models", *Review of Economics and Statistics*, 85(4), 1048-1062

Koenker, R. and G. Bassett (1978). Regression Quantiles, *Econometrica*, 46(1), 33-50.

Kullback, L. and R. A. Leibler (1951), "On Information and Sufficiency", *Annals of Mathematical Statistics* 22, 79-86.

Lee, T.-H. and Y. Yang (2006), "Bagging Binary and Quantile Predictors for Time Series", *Journal of Econometrics* 135, 465-497

Lee, T.-H. and W. Yang (2007), "Money-Income Granger-Causality in Quantiles" UCR.

Manski, C.F. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice", *Journal of Econometrics*, 3(3), 205-228.

Money, A.H., J.F. Affleck-Graves, M.L. Hart, and G.D.I. Barr (1982), "The Linear Regression Model and the Choice of p", *Communications in Statistics - Simulations and Computations,* 11(1), 89-109.

Nyguist, H. (1983), "The Optimal L_p-norm Estimation in Linear Regression Models", *Communications in Statistics - Theory and Methods,* 12, 2511-2524.

Powell, J.L. (1986), "Censored Regression Quantiles", *Journal of Econometrics,* 32, 143-155.

Sawa, T. (1978), "Information Criteria for Discriminating among Alternative Regression Models", *Econometrica* 46, 1273-1291

Varian, H.R. (1975), "A Bayesian Approach to Real Estate Assessment", in *Studies in Bayesian Econometrics and Statistics in Honor of L.J. Savage,* eds. S.E. Fienberg and A. Zellner, Amsterdam: North Holland, pp 195-208.

West, K. (1996), "Asymptotic Inference about Prediction Ability", *Econometrica* 64, 1067-1084.

White, H. (1994), *Estimation, Inference, and Specification Analysis,* Cambridge University Press.

Zellner, A. (1986). "Bayesian Estimation and Prediction Using Asymmetric Loss Functions", *Journal of the American Statistical Association* 81, 446-451