Further Results on

Forecasting and Model Selection Under Asymmetric Loss

Peter F. Christoffersen and Francis X. Diebold

Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104

October 1995
This Print: June 27, 1996

Abstract: We make three related contributions. First, we propose a new technique for solving prediction problems under asymmetric loss using piecewise-linear approximations to the loss function, and we establish existence and uniqueness of the optimal predictor. Second, we provide a detailed application to optimal prediction of a conditionally heteroskedastic process under asymmetric loss, the insights gained from which are broadly applicable. Finally, we incorporate our results into a general framework for recursive prediction-based model selection under the relevant loss function.

Acknowledgements: Helpful discussion was provided by Adolf Buse, Hashem Pesaran, Dale Poirrier, Enrique Sentana, Jim Stock, Ken Wallis, participants at meetings of the Econometric Society World Congress, the NBER/NSF Forecasting Seminar, the UCSD Conference on Multivariate Financial Econometrics, and numerous university seminars. All remaining inadequacies are ours alone. We thank the National Science Foundation, the Sloan Foundation, the University of Pennsylvania Research Foundation and the Barbara and Edward Netter Fellowship for support.

1. Introduction

Proper specification of the loss function is crucial in empirical work (e.g., McCloskey, 1985). Nowhere is this more evident than in forecasting. It is widely acknowledged that textbook favorites like mean squared prediction error or mean absolute prediction error, although mathematically convenient, are not flexible enough to capture the loss structures that often face actual forecasters.

In spite of the need for a practical forecasting framework that incorporates realistic loss functions, until recently one was forced to favor mathematical convenience over realism -- quite simply, there was no alternative. But modern computing power has changed the situation dramatically, as computations that were infeasible not long ago are now done in a few seconds on a desktop computer.

Thus, we have three related objectives in this paper. First, we propose a forecasting framework that exploits modern computational capabilities to find optimal forecasts under general loss structures, in spite of the fact that the optimal predictor typically does not exist in closed form.[1] One approach, taken in Christoffersen and Diebold (1994), is to approximate the optimal predictor. Here we take a different and complementary approach -- instead of approximating the optimal predictor for the exact loss function, we find the exactly optimal predictor for an approximate loss function.

Second, we provide a detailed application to the optimal prediction of a GARCH(1,1)

[1] To the best of our knowledge, only two known loss functions produce closed-form optimal predictors -- the linlin and linex loss functions studied by Granger (1969), Varian (1974), Zellner (1986) and Christoffersen and Diebold (1994).

process under a prediction-error loss function linear on each side of the origin.[2] Conveniently for our illustrative application, the optimal predictor *does* have an analytic closed-form expression under that loss function, as shown by Christoffersen and Diebold (1994). But the insights gained are relevant for *any* attempt at optimal prediction under asymmetric loss, whether by the methods of this paper or our earlier paper.

Finally, we show how optimal prediction under asymmetric loss may be combined with related techniques for estimation and forecast accuracy comparison under asymmetric loss to produce a flexible framework for forecast model selection.

2. Closed-Form Optimal Predictors Typically Don't Exist

To see the difficulty associated with analytic solution, even for very simple loss functions, consider the following natural generalization of quadratic loss ("quadquad" loss), in which loss is quadratic on each side of the origin, but positive errors cost more than negative errors (or conversely),

$$L(y_{t+h}-\hat{y}_{t+h}) = \begin{cases} a(y_{t+h}-\hat{y}_{t+h})^2, \text{ if } y_{t+h}-\hat{y}_{t+h} > 0 \\ b(y_{t+h}-\hat{y}_{t+h})^2, \text{ if } y_{t+h}-\hat{y}_{t+h} \le 0. \end{cases}$$

Conditionally expected loss is

$$E_t\{L(y_{t+h}-\hat{y}_{t+h})\} = a\int_{\hat{y}_{t+h}}^{\infty}(y_{t+h}-\hat{y}_{t+h})^2 \, f(y_{t+h}|\Omega_t) \, dy_{t+h} + b\int_{-\infty}^{\hat{y}_{t+h}}(y_{t+h}-\hat{y}_{t+h})^2 \, f(y_{t+h}|\Omega_t) \, dy_{t+h}.$$

[2] A prediction-error loss function, L(•), is a loss function defined directly on the prediction error, $y-\hat{y}$.

Differentiating with respect to the predictor, we obtain the first order condition

$$a\int_{\hat{y}_{t+h}}^{\infty}(y_{t+h}-\hat{y}_{t+h})\ f(y_{t+h}|\Omega_t)\ dy_{t+h}\ +\ b\int_{-\infty}^{\hat{y}_{t+h}}(y_{t+h}-\hat{y}_{t+h})\ f(y_{t+h}|\Omega_t)\ dy_{t+h}\ =\ 0.$$

It is clear that analytic solution of this first-order condition is impossible in general. Moreover, even in cases as highly-structured as conditional normality, analytic solution remains impossible except for very special cases.[3] To see this, rewrite the first-order condition as

$$a(1-F(\hat{y}_{t+h}|\Omega_t))(E[y_{t+h}|(y_{t+h}>\hat{y}_{t+h})]\ -\ \hat{y}_{t+h})\ +\ bF(\hat{y}_{t+h}|\Omega_t)(E[y_{t+h}|(y_{t+h}<\hat{y}_{t+h})]\ -\ \hat{y}_{t+h})\ =\ 0.$$

Under conditional normality, expressions for the truncated expectations are available. Inserting these, using $F(\hat{y}_{t+h}|\Omega_t)\ =\ \Phi(\xi_{t+h|t})$, where $\xi_{t+h|t}\ =\ \frac{\hat{y}_{t+h}-\mu_{t+h|t}}{\sigma_{t+h|t}}$, and cancelling terms yields[4]

$$(a-b)\phi(\xi_{t+h|t})\sigma_{t+h|t}\ +\ (a-b)\Phi(\xi_{t+h|t})(\hat{y}_{t+h}-\mu_{t+h|t})\ -\ a(\hat{y}_{t+h}-\mu_{t+h|t})\ =\ 0.$$

Thus, although conditional normality does yield some simplification, closed-form analytic solution remains impossible.

Existence and uniqueness of the optimal predictor are easily established under conditional normality, however. Denote the first-order condition that defines the optimal

[3] Newey and Powell (1987) give an analytic solution in the uniform case.

[4] Notice that for a=b the conditional mean is of course optimal.

predictor by $g(\hat{y}_{t+h}) = 0$. Existence follows from $\lim_{\hat{y}_{t+h} \to \infty} g(\hat{y}_{t+h}) < 0$ and $\lim_{\hat{y}_{t+h} \to -\infty} g(\hat{y}_{t+h}) > 0$, together with continuity of the first-order condition. The two limits are easily verified; immediately, $\lim_{\hat{y}_{t+h} \to \infty} g(\hat{y}_{t+h}) = -\infty$ and $\lim_{\hat{y}_{t+h} \to -\infty} g(\hat{y}_{t+h}) = +\infty$. For uniqueness we need that $g'(\hat{y}_{t+h})$ be strictly negative everywhere. This too is easily verified; immediately,

$$g'(\hat{y}_{t+h}) = -a(1-\Phi(\xi_{t+h|t})) - b\Phi(\xi_{t+h|t}),$$

which is strictly negative everywhere, because a>0, b>0 and $\Phi(\cdot)$ is a cumulative density function.

When the optimal predictor exists and is unique (as is the case here), numerical algorithms (nonlinear equation solution algorithms in conjunction with numerical integration) may be used to compute the optimal predictor quickly and reliably. We now turn to a convenient and flexible class of loss functions for which it is easy to show that the optimal predictor exists and is unique, even in conditionally non-Gaussian cases.

3. Piecewise-Linear Approximation of the Loss Function

Consider a piecewise-linear loss function L(•) constructed by concatenating linear segments, such that the loss of zero is zero and loss is increasing on each side of the origin. This may actually *be* the relevant loss function, or it may be used as an approximation to any prediction-error loss function.[5]

Conditionally expected loss is

[5] Note that any desired level of approximation accuracy may be obtained by taking sufficiently many segments.

$$E_t\{L(y_{t+h}-\hat{y}_{t+h})\} = \sum_{i=1}^{I-1} \int_{\hat{y}_{t+h}+c_{i-1}}^{\hat{y}_{t+h}+c_i} (a_i(y_{t+h}-\hat{y}_{t+h})+b_i)f(y_{t+h}|\Omega_t)dy_{t+h} + \int_{\hat{y}_{t+h}+c_{I-1}}^{\infty} (a_I(y_{t+h}-\hat{y}_{t+h})+b_I)f(y_{t+h}|\Omega_t)dy_{t+h}$$

$$+ \sum_{j=1}^{J-1} \int_{\hat{y}_{t+h}+c^{j}}^{\hat{y}_{t+h}+c^{j-1}} (a^{j}(y_{t+h}-\hat{y}_{t+h})+b^{j})f(y_{t+h}|\Omega_t)dy_{t+h} + \int_{-\infty}^{\hat{y}_{t+h}+c^{J-1}} (a^{J}(y_{t+h}-\hat{y}_{t+h})+b^{J})f(y_{t+h}|\Omega_t)dy_{t+h},$$

for I, J $\geq$ 2. The first line denotes the pieces on the positive side of the origin and the second line the negative, i.e., $a_i \geq 0, \ \forall i$ and $a^j \leq 0, \ \forall j$. The c_i's and c^j's denote the breakpoints between segments, with $c^l < c^k < 0$ and $0 < c_k < c_l$, $\forall \ l > k$. To ensure zero loss at the origin we impose $b_1 = b^1 = c_0 = c^0 = 0$. To ensure that neighboring segments connect at the breakpoints we impose $b_i = b_{i-1} + (a_{i-1}-a_i)c_{i-1}, \ i = 2, 3, ..., I$, and similarly $b^j = b^{j-1} + (a^{j-1}-a^j)c^{j-1}, \ j = 2, 3, ..., J$.

Differentiating with respect to the predictor, $\hat{y}_{t+h}$, and using Leibniz's rule we obtain

$$\sum_{i=1}^{I-1} (a_i c_i+b_i)f(\hat{y}_{t+h}+c_i|\Omega_t) - \sum_{i=1}^{I-1} (a_i c_{i-1}+b_i)f(\hat{y}_{t+h}+c_{i-1}|\Omega_t) - (a_I c_{I-1}+b_I)f(\hat{y}_{t+h}+c_{I-1}|\Omega_t)$$

$$- \sum_{i=1}^{I-1} a_i(F((\hat{y}_{t+h}+c_i)|\Omega_t)-F(\hat{y}_{t+h}+c_{i-1})) - a_I(1-F((\hat{y}_{t+h}+c_{I-1})|\Omega_t))$$

$$+ \sum_{j=1}^{J-1} (a^{j}c^{j-1}+b^{j})f((\hat{y}_{t+h}+c^{j-1})|\Omega_t) + (a^{J}c^{J-1}+b^{J})f((\hat{y}_{t+h}+c^{J-1})|\Omega_t) - \sum_{j=1}^{J-1} (a^{j}c^{j}+b^{j})f((\hat{y}_{t+h}+c^{j})|\Omega_t)$$

$$- \sum_{j=1}^{J-1} a^{j}(F((\hat{y}_{t+h}+c^{j-1})|\Omega_t)-F((\hat{y}_{t+h}+c^{j})|\Omega_t)) - a^{J}F((\hat{y}_{t+h}+c^{J-1})|\Omega_t) = 0.$$

This first-order condition defines the optimal predictor. After some manipulation all pdf terms cancel, leaving

$$-\sum_{i=1}^{I-1} a_i(F((\hat{y}_{t+h}+c_i)|\Omega_t)-F((\hat{y}_{t+h}+c_{i-1})|\Omega_t)) \;-\; a_I(1-F((\hat{y}_{t+h}+c_{I-1})|\Omega_t))$$
$$-\sum_{j=1}^{J-1} a^{j}(F((\hat{y}_{t+h}+c^{j-1})|\Omega_t)-F((\hat{y}_{t+h}+c^{j})|\Omega_t)) \;-\; a^{J}F((\hat{y}_{t+h}+c^{J-1})|\Omega_t) \;=\; 0,$$

or equivalently (after a bit more manipulation),

$$\sum_{i=2}^{I}(a_i-a_{i-1})F((\hat{y}_{t+h}+c_{i-1})|\Omega_t) \;+\; \sum_{j=2}^{J}(a^{j-1}-a^{j})F((\hat{y}_{t+h}+c^{j-1})|\Omega_t) \;+\; (a_1-a^1)F((\hat{y}_{t+h})|\Omega_t) \;-\; a_I \;=\; 0.$$

This first-order condition cannot be solved analytically, but it is easy to solve numerically, given the conditional cumulative density function $F(y_{t+h}|\Omega_t)$. Sufficient conditions for existence and uniqueness of the solution are given in the following theorem.

Theorem If:

(1) $a_i \geq a_{i-1},\ i = 2, 3, ..., I$ and $a^{j-1} \geq a^{j},\ j = 2, 3, ..., J$

(2) $f(y|\Omega) > 0,\ \forall y$

(3) $a_i > a_{i-1}$ for some i, or $a^{j-1} > a^{j}$, for some j,

then a solution to the first-order condition exists and is unique.

Proof Denote the first-order condition by $g(\hat{y}_{t+h}) = 0$. We shall show that $\lim_{\hat{y}_{t+h}\to\infty} g(\hat{y}_{t+h}) > 0$ and $\lim_{\hat{y}_{t+h}\to-\infty} g(\hat{y}_{t+h}) < 0$, so that the first-order condition has at least one root, by continuity of $g(\cdot)$. Immediately, $\lim_{\hat{y}_{t+h}\to\infty} g(\hat{y}_{t+h}) = -a^{J}$ and $\lim_{\hat{y}_{t+h}\to-\infty} g(\hat{y}_{t+h}) = -a_I$. These limits are strictly positive and negative, respectively, by condition (3) in conjunction with the fact that the a_i's are all non-negative and the a^j's are all non-positive. Now we establish uniqueness by showing that $g^{/}(\hat{y}_{t+h}) > 0,\ \forall \hat{y}_{t+h}$. Immediately,

$$g^{/}(\hat{y}_{t+h}) \;=\; \sum_{i=2}^{I}(a_i-a_{i-1})f((\hat{y}_{t+h}+c_{i-1})|\Omega_t) \;+\; \sum_{j=2}^{J}(a^{j-1}-a^{j})f((\hat{y}_{t+h}+c^{j-1})|\Omega_t) \;+\; (a_1-a^1)f(\hat{y}_{t+h}|\Omega_t).$$

Notice that all terms are nonnegative from condition (1) in conjunction with the fact that the a_i's are all non-negative and the a^j's are all non-positive, and because $f(\cdot)$ is a density function. Conditions (2) and (3) are sufficient to guarantee strict positivity, by guaranteeing that at least one term is strictly positive, but of course they are not necessary. Q.E.D.

4. Forecasting a Conditionally Heteroskedastic Process Under Asymmetric Loss

Here we illustrate our methods by predicting a simple conditionally-Gaussian GARCH(1,1) process under linlin loss. The GARCH(1,1) process is

$$y_{t+1} = \varepsilon_{t+1}, \qquad \varepsilon_{t+1}|\Omega_t \sim N(0,\ \sigma^2_{t+1|t})$$

$$\sigma^2_{t+1|t} = \omega + \alpha\varepsilon^2_t + \beta\sigma^2_{t|t-1}, \qquad \omega,\ \alpha,\ \beta > 0,\ \alpha+\beta < 1,$$

Linlin loss, for which where there is only one linear piece on each side of the origin, is a special case of piecewise-linear loss ($a_i=a_1$ for all I, and $a^j=a^1$, for all j, which in turn implies $b_i=0$ and $b^j=0$ for all I and j). In Figure 1, we show various parameterizations of the linlin loss function superimposed for reference on a symmetric, quadratic loss function. The first order condition that defines the optimal predictor collapses to $(a_1 - a^1)F(\hat{y}_{t+h}|\Omega_t) - a_1 = 0$, which actually yields a closed form for the optimal predictor, $\hat{y}_{t+h} = F^{-1}(\frac{a_1}{a_1 - a^1}|\Omega_t)$.[6] Throughout, we normalize the unconditional variance to 1 by taking $\omega = (1-\alpha-\beta)$, and we set the GARCH parameters at α=.2 and β=.75, which are typical of estimates reported in the

[6] See Christoffersen and Diebold (1994) for more detailed discussion of optimal prediction under linlin loss.

literature. We set the linlin loss parameters at a_1=.95 and a^1=- .05, corresponding to high asymmetry, which is useful for pedagogical purposes.

For h=1 the conditional density is Gaussian so the optimal predictor is easily calculated as $\hat{y}_{t+h} = F^{-1}(\frac{a_1}{a_1 - a^1}|\Omega_t) = \sigma_{t+h|t}\ \Phi^{-1}(\frac{a_1}{a_1 - a^1}) = 1.65\sigma_{t+h|t}$. We will compare the conditionally expected linlin loss of the optimal predictor to that of two competitors. The first competitor is the pseudo-optimal predictor, $\hat{y}_{t+h} = \sigma_h\ \Phi^{-1}(\frac{a_1}{a_1 - a^1}) = 1.65,$ which ignores conditional heteroskedasticity, and the second is the conditional mean predictor, $\hat{y}_{t+h} = \mu_{t+h|t} = 0,$ which ignores both loss asymmetry and conditional heteroskedasticity.

Note that the optimal predictor acknowledges loss asymmetry and the possibility of conditional heteroskedasticity through a possibly *time-varying* adjustment to the conditional mean, thereby providing a direct link from conditional heteroskedasticity to optimal *point* prediction, rather than simply to *interval* prediction. The conditional mean, in contrast, is always suboptimal as it incorporates *no* adjustment. The pseudo-optimal predictor is intermediate in that it incorporates only a *constant* adjustment for asymmetry; thus, it is fully optimal only in the conditionally homoskedastic case $\sigma^2_{t+h|t} = \sigma^2_h,\ \forall\ t,\ h.$

In Figure 2, we show a realization of the GARCH(1,1) process, together with the real-time linlin-optimal, pseudo-optimal and conditional mean predictors. It is apparent that the optimal predictor injects more bias when conditional volatility is high, reflecting the fact that it accounts for both loss asymmetry and conditional heteroskedasticity. This conditionally optimal amount of bias is sometimes more and sometimes less than the constant bias associated with the pseudo-optimal predictor, which accounts for loss asymmetry but not conditional heteroskedasticity. Finally, of course, the conditional mean injects no bias, as it accounts neither for loss asymmetry nor conditional heteroskedasticity.

It is worth mentioning that the "optimal" predictor used here is truly optimal only for h = 1, because conditional normality holds only for h = 1. But, although the "optimal" predictor used in this example is in fact only an approximation to the optimal predictor when h > 1 (it is in fact an improved pseudo-optimal predictor), one expects it to perform better than the "constant adjustment" pseudo-optimal predictor, because it explicitly adapts to the time-varying conditional variance. Recognizing the abuse of language, we shall continue to refer to it as the "optimal predictor" and to use the predictor formula for h > 1.[7]

Computation of conditionally expected linlin loss requires conditioning on an initial value of $\sigma^2_{t+1|t}$, and the results will, of course, vary with the value adopted. We set the initial conditional variance equal to the unconditional variance plus one standard deviation of the conditional variance, $\sigma^2_{t+1|t} = \sigma^2_1 + \sqrt{\mathrm{var}(\sigma^2_{t+1|t})}$.[8] Calculation of $\mathrm{var}(\sigma^2_{t+1|t})$, the variance of the conditional variance, is straightforward but somewhat tedious. We have $\mathrm{var}\left(\sigma^2_{t+1|t}\right) = E\left[\left(\sigma^2_{t+1|t}\right)^2\right] - \left(\sigma^2_1\right)^2$, but recall that $E_t\varepsilon^4_{t+1} = 3(\sigma^2_{t+1|t})^2$, so that $(\sigma^2_{t+1|t})^2 = (E_t\varepsilon^4_{t+1})/3$. Thus, $\mathrm{var}\left(\sigma^2_{t+1|t}\right) = \dfrac{E\varepsilon^4_{t+1}}{3} - \left(\sigma^2_1\right)^2$, by the law of iterated expectations, and as shown by Bollerslev (1986) the requisite unconditional fourth moment is

$$E\varepsilon^4_{t+1} = \frac{3\omega^2(1+\alpha+\beta)}{(1-\alpha-\beta)(1-\beta^2-2\alpha\beta-3\alpha^2)} = \frac{3(1-\alpha-\beta)(1+\alpha+\beta)}{1-\beta^2-2\alpha\beta-3\alpha^2},$$

because we set $\omega = (1-\alpha-\beta)$. The normalization of ω implies that $\sigma^2_1 = 1$, and we get

[7] Baillie and Bollerslev (1992) suggest a Cornish-Fisher expansion to approximate the conditional distribution for h > 1, but such extensions are beyond the scope of the present example.

[8] Note that it would be uninformative to set $\sigma^2_{t+1|t}$ equal to the unconditional variance, σ^2_1, because that would obscure the difference between the optimal and pseudo-optimal predictors.

$$\mathrm{var}(\sigma^2_{t+1|t}) = \frac{\alpha}{1-\beta^2-2\alpha\beta-3\alpha^2}.$$

Computation of conditionally expected linlin loss also requires an expression for $\sigma_{t+h|t}$, which enters the expression for the optimal linlin predictor. Using results from Baillie and Bollerslev (1992), it is easy to show that for the GARCH(1,1) process,

$$\sigma_{t+h|t} = \sqrt{\sigma^2_h + (\sigma^2_{t+1|t}-\sigma^2_1)(\alpha+\beta)^{h-1}}\ .$$

Because of the conditional non-normality when h > 1, we do not rely on the formulas derived in Diebold and Christoffersen (1994) to compute the conditionally expected losses of the optimal, pseudo-optimal, and conditional-mean predictors. Instead, we compute them by simulation. At each of 20,000 replications, we draw a GARCH(1,1) realization of length 50, with the conditional variance initialized as discussed above, and we compute the loss of each of the three predictors at each of the 50 horizons. Finally, we average across replications.

In Figure 3, we show the conditionally expected linlin loss of the pseudo-optimal predictor relative to that of the optimal predictor, across prediction horizons. The increase in conditionally expected loss from ignoring the conditional variance dynamics--that is, the increase in conditionally expected loss from using the pseudo-optimal as opposed to the optimal predictor--is as high as 35% for short horizons. Of course, as the prediction horizon increases, the cost of ignoring the conditional variance dynamics decreases, and the ratio of conditionally expected losses converges to 1.

In Figure 4, we show the conditionally expected linlin loss of the conditional mean relative to that of the optimal predictor. Although the cost of ignoring the conditional variance dynamics still decreases with horizon, the ratio of conditionally expected losses does not approach 1, because the conditional mean predictor ignores loss asymmetry in addition to conditional heteroskedasticity. The failure of the conditional mean to acknowledge the loss asymmetry affects predictive performance at *all* horizons.

5. Model Selection Under the Relevant Loss Function

The prediction techniques developed here can be used in recursive prediction-based procedures for model selection under the relevant loss function. This also involves estimation under the relevant loss function, as in Weiss and Andersen (1984) and Weiss (1994). Important related work along those lines, under a Kullback-Liebler distance metric (one-step-ahead squared-error loss), is reported in Vuong (1989) and Phillips (1994).

First, assume prediction-error loss with known optimal predictor of the form

$$\hat{y}_{t+h} = \mu_{t+h|t}(\theta) + f(\delta, \gamma_{t+h|t}(\theta)),$$

where δ is the vector of loss function parameters, θ is the vector of model parameters, $\gamma_{t+h|t}(\theta)$ is the vector of higher order moments, and $f(\bullet)$ might be an explicit function or it might be given implicitly by a first order condition.[9]

Let the initial estimation sample run from t = 1, ..., T^*, so that the "holdout sample" used for comparing predictive performance runs from t = T^*+1, ..., T. We proceed as follows:

(1) Using a numerical optimization procedure, find for model j:

[9] The separation of conditional mean and higher order dynamics is guaranteed by a theorem in Christoffersen and Diebold (1994), who build on an earlier result of Granger (1969). The theorem follows from the loss function being defined directly on prediction errors.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{T^{*}-h} \sum_{t=1}^{T^{*}-h} L(y_{t+h} - \hat{y}^{j}_{t+h}(\theta)).$$

Such procedures are discussed in Weiss and Andersen (1984) and Weiss (1994).

(2) Calculate the loss of the h-step-ahead prediction error at time T^{*},

$$L^{j}_{1} = L(y_{T^{*}+h} - \hat{y}^{j}_{T^{*}+h}).$$

(3) Use terminal estimation date $T^{*}+1$. Repeat steps (1) and (2) to get

$$L^{j}_{2} = L(y_{T^{*}+1+h} - \hat{y}^{j}_{T^{*}+1+h}).$$

(4) Repeat steps (1)-(3) until the terminal estimation date is T-h. Then form the average loss for model j as

$$\bar{L}^{j} = \frac{1}{T-h-T^{*}+1} \sum_{i=1}^{T-h-T^{*}+1} L^{j}_{i}.$$

(5) Repeat steps (1)-(4) for all models j = 1, 2, ..., J. Use the appropriate Diebold-Mariano (1995) test to assess the significance of the difference between the models based on average loss.

Second, suppose the form of the optimal predictor is unknown. Hence the algorithm above must be augmented with a step that estimates the form of the predictor. This situation could be brought about by a non-tractable loss function, perhaps not defined on the prediction errors. In the conditionally Gaussian case, we form the predictor as an expansion in the first two conditional moments (here, for example, we adopt a second-order expansion, but higher-order terms could of course be included):

$$\hat{y}_{t+h}(\beta,\ \theta)\ =\ \beta_0+\beta_1\mu_{t+h|t}(\theta)+\beta_2\sigma_{t+h|t}(\theta)+\beta_3(\sigma_{t+h|t}(\theta))^2+\beta_4(\mu_{t+h|t}(\theta))^2+\beta_5(\mu_{t+h|t}(\theta)\sigma_{t+h|t}(\theta)).$$

Step (1) of the algorithm simply becomes more complicated; the others are unchanged. Step (1) becomes:

(1') Using a numerical optimization procedure, find for model j:

$$\{\hat{\beta},\ \hat{\theta}\}\ =\ \underset{\{\beta,\ \theta\}}{\text{argmin}}\ \frac{1}{T^*-h}\sum_{t=1}^{T^*-h} L(y_{t+h}\ -\ \hat{y}^j_{t+h}(\beta,\ \theta)).$$

Finally, if the form of the optimal predictor is unknown and the conditional density is non-Gaussian, again only step (1) changes. We form the predictor as an expansion in the conditional moments, but moments above the second will need to be included. Hansen's (1994) autoregressive conditional density approach may help to achieve parsimony.

6. Summary and Directions for Future Research

We have studied prediction under asymmetric loss and its role in a broader framework for model selection. The discussion consisted of three parts. First, we suggested a flexible yet tractable piecewise-linear approximation to the loss function, and we established existence and uniqueness of the optimal predictor. This approach to optimal prediction under asymmetric loss complements those proposed by Christoffersen and Diebold (1994).

Second, we provided a detailed application to prediction of a GARCH(1,1) process under linlin loss, which clearly illustrated the fact that higher-order conditional moments (that is, conditional moments beyond the conditional mean) are relevant for point prediction under asymmetric loss. Under conditional normality, for example, the conditional variance plays a key role in optimal point prediction. Thus, as in Granger (1981) (although for very different reasons), one can "forecast white noise" under asymmetric loss.

Third, we showed how our results on optimal prediction under asymmetric loss could be combined with results on estimation and forecast accuracy comparison under asymmetric loss to produce a unified and general framework for forecast model selection.

As for future work, it will be of interest to examine the usefulness of the parametric prediction and model selection procedures developed here in applied forecasting, and to compare the performance of our parametric predictors to White's (1992) nonparametric predictor.[10] We conjecture that our approach will perform well, as much of the literature suggests that simple, tightly parameterized -- but nevertheless sophisticated -- models tend to perform best in out-of-sample prediction.[11]

[10] White develops his nonparametric prediction procedure under linlin loss, but it is readily extended to other loss functions.

[11] See, for example, Zellner (1992).

References

Baillie, R.T. and T. Bollerslev (1992), "Prediction in Dynamic Models with Time-Dependent Conditional Variances," *Journal of Econometrics*, 52, 91-113.4,

Christoffersen, P. F. and F. X. Diebold (1994), "Optimal Prediction under Asymmetric Loss," *NBER Technical Working Paper No. 167*, revised September 1995.

Diebold, F.X. and R.S. Mariano (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253-265.

Granger, C.W.J. (1969), "Prediction with a Generalized Cost of Error Function," *Operational Research Quarterly*, 20,199-207.

Granger, C.W.J. (1981), "Forecasting White Noise," in A. Zellner (ed.), *Proceedings of the Conference on Applied Time Series Analysis of Economic Data*. ASA-Census-NBER.

Hansen, B.E. (1994), "Autoregressive Conditional Density Estimation," *International Economic Review*, 35, 705-730.

McCloskey, D.N. (1985), "The Loss Function has Been Mislaid: The Rhetoric of Significance Tests," *American Economic Review*, 75, 201-205.

Newey, W.K. and J.L. Powell (1987), "Asymmetric Least Squares Estimation and Testing," *Econometrica*, 55, 819-847.

Phillips, P.C.B. (1994), "Bayes Models and Macroeconomic Activity," Manuscript, Department of Economics, Yale University.

Varian, H. (1974), "A Bayesian Approach to Real Estate Assessment," in S.E. Fienberg and A. Zellner (eds.), *Studies in Bayesian Econometrics and Statistics in Honor of L.J. Savage*, 195-208. Amsterdam: North-Holland.

Vuong, Q. (1989), "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses," *Econometrica*, 57, 307-333.

Weiss, A.A. (1994), "Estimating Time Series Models Using the Relevant Cost Function," Manuscript, Department of Economics, University of Southern California.

Weiss, A.A. and Andersen, A.P. (1984), "Estimating Forecasting Models Using the Relevant Forecast Evaluation Criterion," *Journal of the Royal Statistical Society A*, 137, 484-487.

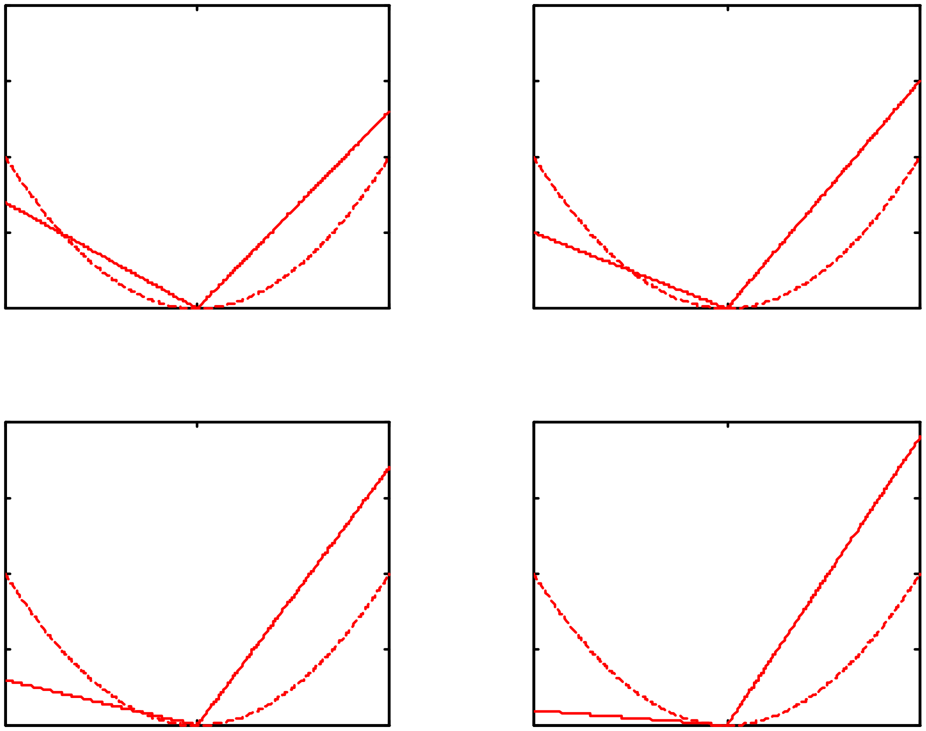
White, H. (1992) "Nonparametric Estimation of Conditional Quantiles Using Neural

Networks," in C. Page and R. LePage (eds.), *Computing Science and Statistics, Proceedings of the 22nd Symposium on the Interface, Statistics of Many Parameters: Curves, Images, Spatial Models*. New York: Springer Verlag.

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

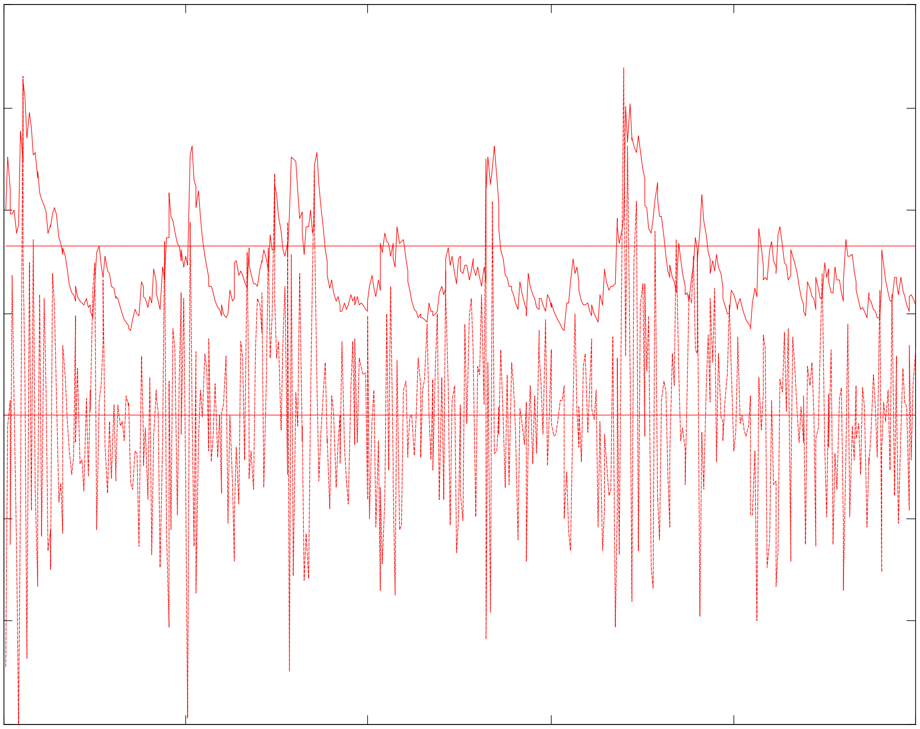
Zellner, A. (1992), "Statistics, Science and Public Policy," *Journal of the American Statistical Association*, 87, 1-6.

Figure 1
Various Linlin Loss Functions with Quadratic Loss Superimposed for Reference



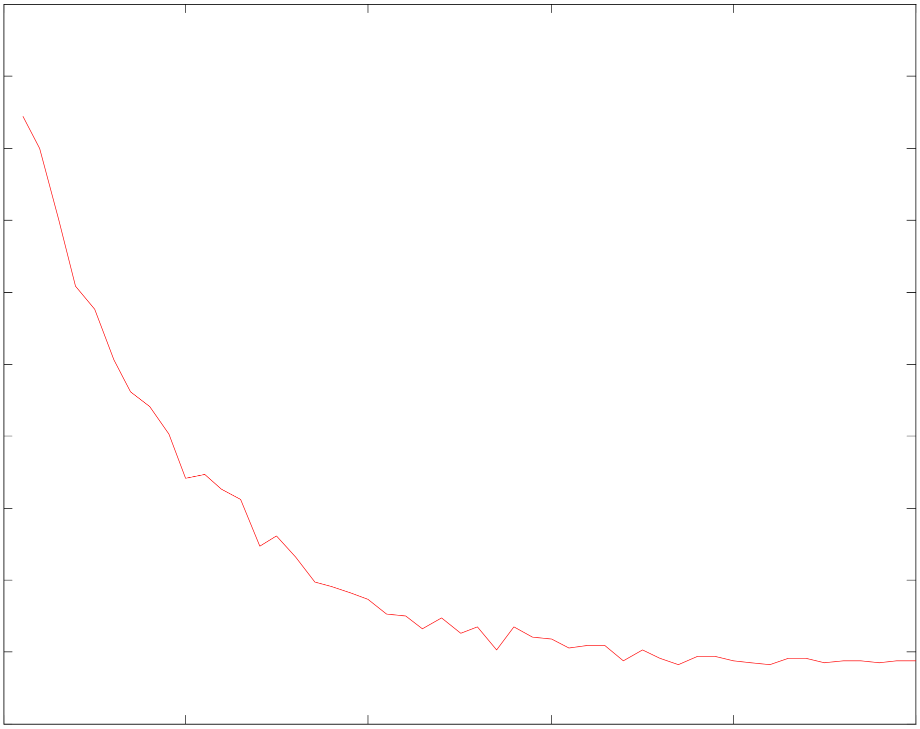
Notes to Figure: Quadratic loss appears as a dashed line and linlin loss appears as a solid line. Asym = $a_1 / (a_1 - a^1)$, where a_1 and a^1 are Linlin loss parameters such that $L(x) = a_1x$, if $x > 0$; and $L(x) = - a^1x$, if $x \leq 0$.

Figure 2
GARCH(1,1) Realization with
Linlin Optimal, Pseudo-Optimal, and Conditional Mean Predictors



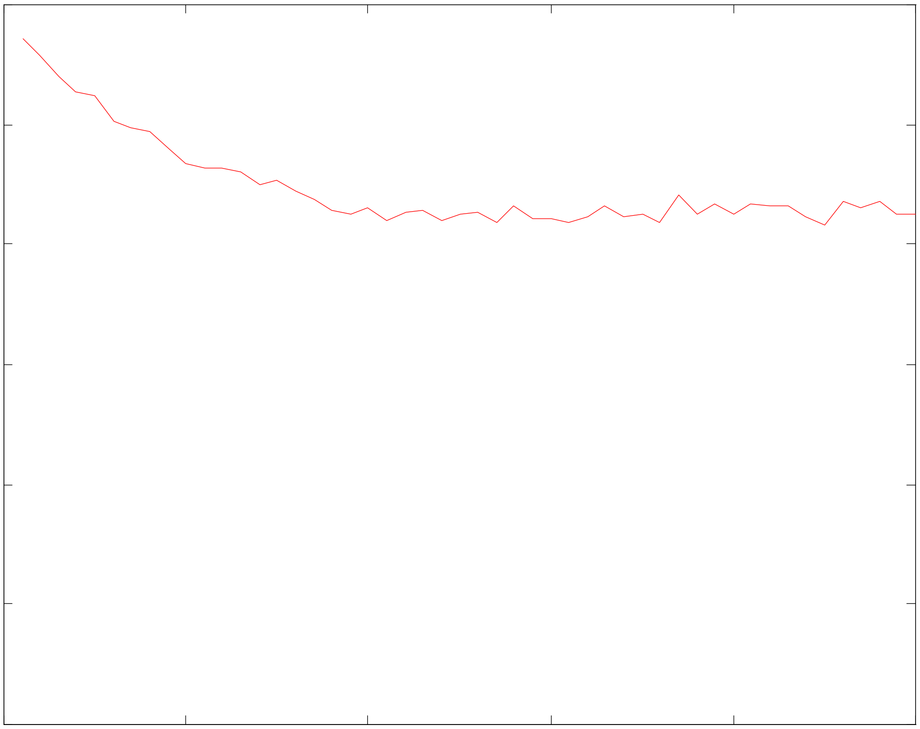
Notes to Figure: The linlin loss parameters are set to a_1 = .95 and a^1 = - .05, so that $a_1/(a_1 - a^1)$ = .95. The GARCH(1,1) parameters are set to α=.2 and β=.75. The dotted line is the GARCH(1,1) realization. The horizontal line at zero is the conditional mean predictor, the horizontal line at 1.65 is the pseudo-optimal predictor, and the time-varying solid line is the optimal predictor.

Figure 3
Ratio of Conditionally Expected Linlin Loss of Pseudo-Optimal and Optimal Predictors



Notes to Figure: The linlin loss parameters are set to a_1 = .95 and a^1 =- .05, so that $a_1/(a_1 - a^1)$ = .95. The GARCH(1,1) parameters are set to α = .2 and β = .75.

Figure 4
Ratio of Conditionally Expected Linlin Loss of Conditional Mean and Optimal Predictors



Notes to Figure: The linlin loss parameters are set to $a_1 = .95$ and $a^1 = -.05$, so that $a_1/(a_1 - a^1) = .95$. The GARCH(1,1) parameters are set to $\alpha = .2$ and $\beta = .75$.