Evaluating Density Forecasts with Applications to Financial Risk Management

Francis X. Diebold; Todd A. Gunther; Anthony S. Tay

International Economic Review, Vol. 39, No. 4, Symposium on Forecasting and Empirical Methods in Macroeconomics and Finance. (Nov., 1998), pp. 863-883.

Stable URL:
http://links.jstor.org/sici?sici=0020-6598%28199811%2939%3A4%3C863%3AEDFWAT%3E2.0.CO%3B2-%23

International Economic Review is currently published by Economics Department of the University of Pennsylvania.

INTERNATIONAL ECONOMIC REVIEW
Vol. 39, No. 4, November 1998

EVALUATING DENSITY FORECASTS WITH APPLICATIONS TO FINANCIAL RISK MANAGEMENT*

BY FRANCIS X. DIEBOLD, TODD A. GUNTHER, AND ANTHONY S. TAY[†1]

Department of Economics, University of Pennsylvania, and NBER, U.S.A.
Department of Economics, University of Pennsylvania, U.S.A.
Department of Economics and Statistics, National University of Singapore, Singapore

Density forecasting is increasingly more important and commonplace, for example in financial risk management, yet little attention has been given to the evaluation of density forecasts. We develop a simple and operational framework for density forecast evaluation. We illustrate the framework with a detailed application to density forecasting of asset returns in environments with time-varying volatility. Finally, we discuss several extensions.

1. INTRODUCTION

Prediction occupies a distinguished position in econometrics, as it does in all the sciences. Hence, evaluating predictive ability is a fundamental concern. Reviews of the forecast evaluation literature, such as Diebold and Lopez (1996), reveal that most attention has been paid to evaluating *point* forecasts. In fact, the bulk of the literature focuses on point forecasts, while conspicuously smaller sub-literatures interval forecasts (Chatfield 1993, Christoffersen 1998) and probability forecasts (Wallis 1993, Clemen et al., 1995).

Particularly little attention has been given to evaluating *density forecasts*. At least three factors explain this neglect. First, analytic construction of density forecasts has historically required restrictive and sometimes dubious assumptions, such as linear

* Manuscript received October 1996.
† E-mail: *fdiebold@mail.sas.upenn.edu*
[1] Thorough and repeated readings and comments from two referees and Ken West drastically improved this paper, but remaining inadequacies are ours alone. Helpful discussion was also provided by participants at the University of California San Diego Conference on Time Series Analysis of High-Frequency Financial Data, the NBER/NSF Time Series Seminar, the LSE Financial Markets Group Conference on Empirical Finance, the European and North American meetings of the Econometric Society, and Computational Finance '97, as well as seminar participants at Harvard/MIT, Michigan, Penn, Princeton, NYU; the Federal Reserve Bank of Kansas City, and the Federal Reserve Bank of Atlanta. We are especially grateful for helpful comments from Gary Chamberlain, John Geweke, Eric Ghysels, Clive Granger, Jin Hahn, Bruce Hansen, Andrew Harvey, Jerry Hausman, Hashem Pesaran, Gleb Sandmann, Neil Shephard, Jim Stock, Ian Tonks, Casper De Vries, Ken Wallis, Mark Watson, and Tao Zha. For support we thank the National Science Foundation, the Sloan Foundation, the University of Pennsylvania Research Foundation, and the National University of Singapore.

dynamics, Gaussian innovations and no parameter estimation uncertainty. Recent work using numerical and simulation techniques to construct density forecasts, however, has reduced our reliance on such assumptions. In fact, improvements in computer technology have rendered the provision of credible density forecasts increasingly straightforward, in both classical and Bayesian frameworks.[2]

Second, until recently there was little demand for density forecasts; historically, point and interval forecasts seemed adequate for most users' needs. Again, however, recent developments have changed the status quo, particularly in quantitative finance. The booming area of financial risk management, for example, is effectively dedicated to providing density forecasts of portfolio values and to tracking certain aspects of the densities, such as value at risk. The day will soon arrive in which risk management will routinely entail nearly real-time issuance and evaluation of such density forecasts.

Finally, the problem of density forecast evaluation appears difficult. Although it is possible to adapt techniques developed for the evaluation of point, interval and probability forecasts to the evaluation of density forecasts, such approaches lead to incomplete evaluation of density forecasts. For example, using Christoffersen's (1998) method for evaluating interval forecasts, we can evaluate whether the series of 90% prediction intervals corresponding to a series of density forecasts is correctly conditionally calibrated but that leaves open the question of whether the corresponding prediction intervals at other confidence levels are correctly conditionally calibrated. Correct conditional calibration of density forecasts corresponds to the simultaneous correct conditional calibration of all possible interval forecasts, the assessment of which seems a daunting task.

In light of the increasing importance of density forecasts, and lack of attention paid to them in the literature, we propose methods for evaluating density forecasts. Our evaluation methods are based on an integral transform that turns out to have a long history, dating at least to Rosenblatt (1952). Independent work by Crnkovic and Drachman (1996) is also closely related, as is that of Granger and Pesaran (1996), who study decision making guided by probability forecasts defined over discrete outcomes.

We proceed as follows. In Section 2, we present a statement and discussion of the problem, and we provide decision-theoretic motivation for the density forecast evaluation methods that we introduce subsequently in Section 3. In Section 4, we provide a detailed simulation example of density forecast evaluation in an environment with time-varying volatility. In Section 5, we use our tools to evaluate density forecasts of U.S. S&P 500 daily stock returns. We conclude in Section 6.

2. DENSITY FORECASTS, LOSS FUNCTIONS AND ACTION CHOICES: IMPLICATIONS FOR DENSITY FORECAST EVALUATION

Studying the relationships among density forecasts, loss functions and action choices will help to clarify what can and cannot be hoped for when evaluating density forecasts, and it will also suggest productive directions for density forecast

[2] See, for example, Efron and Tibshirani (1993), and Gelman et al. (1995).

evaluation. We first show that the problem of density forecast evaluation is intrinsically linked to the forecast user's loss function, which would appear to bode poorly for our quest for a universally applicable approach to density forecast evaluation. We then show that, contrary to first impressions, all is not lost: the analysis suggests an important approach to density forecast evaluation, which we pursue in subsequent sections.

The Decision Environment. Let $\{f_t(y_t|\Omega_t)\}_{t=1}^m$ be the sequence of conditional densities governing a series y_t, where $\Omega_t = \{y_{t-1}, y_{t-2}, \ldots\}$, and let $\{p_t(y_t|\Omega_t)\}_{t=1}^m$ be a corresponding sequence of 1-step-ahead density forecasts.[3] Finally, let $\{y_t\}_{t=1}^m$ denote the corresponding series of realizations.[4] The forecast user has a loss function $L(a, y)$, where a refers to an action choice, and chooses an action to minimize expected loss computed using the density believed to be the data generating process. If the user believes that the density forecast $p(y)$ is the correct density, then he chooses an action a^* such that[5]

$$a^*(p(y)) = \underset{a \in A}{\operatorname{argmin}} \int L(a, y) p(y)\, dy.$$

The action choice defines the loss $L(a^*, y)$ faced for every realization of the process $y \sim f(y)$. This loss is a random variable and possesses a probability distribution that depends only on the action choice.

Expected loss with respect to the true data generating process is

$$E[L(a^*, y)] = \int L(a^*, y) f(y)\, dy.$$

The effect of the density forecast on the user's expected loss is easily seen. Different density forecasts will, in general, lead to different action choices and hence different distributions of loss. The better a density forecast, the lower its expected loss, computed with respect to the true data generating process.

Ranking Two Forecasts. Suppose the user has the option of choosing between two forecasts in a given period, denoted by $p_j(y)$ and $p_k(y)$, where the subscript refers to the forecast. The user will weakly prefer forecast $p_j(y)$ to forecast $p_k(y)$ if

$$\int L(a_j^*, y) f(y)\, dy \leq \int L(a_k^*, y) f(y)\, dy,$$

[3] For notational convenience, we will often not indicate the information set and simply write $f_t(y_t)$ and $p_t(y_t)$, but the dependence on Ω_t should be understood. Moreover, because in this section we consider the relationships among density forecasts, loss functions and actions in a one-period context, we temporarily drop the time subscripts for notational convenience.

[4] We indulge in the standard abuse of notation, which favors convenience over precision, by failing to distinguish between random variables and their realizations. The meaning will be clear from context.

[5] We assume a unique minimizer, a sufficient condition for which is that A be compact and that L be strictly convex in a.

where a_j^* denotes the action that minimizes expected loss when the user bases the action choice on forecast j.

Ideally, we would like to find a ranking of forecasts with which all users agree, *regardless of their loss function*. Unfortunately, such a ranking does not exist. More precisely, there does not exist a ranking r of arbitrary density forecasts p_j and p_k, both distinct from f, such that for all loss functions $L(a, y)$,

$$r_j \geq r_k \Leftrightarrow \int L(a_j^*, y) f(y)\, dy \geq \int L(a_k^*, y) f(y)\, dy.$$

To see why, simply notice that it is easy to find a pair of loss functions L_1 and L_2, a density function f governing y, and a pair of forecasts, p_j and p_k, such that

$$\int L_1(a_k^*, y) f(y)\, dy < \int L_1(a_j^*, y) f(y)\, dy,$$

while

$$\int L_2(a_k^*, y) f(y)\, dy > \int L_2(a_j^*, y) f(y)\, dy.$$

That is, user 1 does better on average under forecast k, while user 2 does better under forecast j. Suppose, for example, that the true density function is $N(0, 1)$, and suppose that user 1's loss function is $L_1(a, y) = (y - a)^2$ and user 2's loss function is $L_2(a, y) = (y^2 - a)^2$. The optimal action choices are then $\int yp(y)\, dy$ and $\int y^2 p(y)\, dy$. That is, user 1 bases his action choice on the mean, with higher expected loss occurring with larger errors in the forecast mean, while the actions and expected losses of user 2 depend on the error in the forecast of the uncentered second moment. In this context, consider two forecasts: forecast j is $N(0, 2)$ and forecast k is $N(1, 1)$. User 1 prefers forecast j, because it leads to an action choice implying lower expected loss, but user 2 prefers forecast k for the same reason.

To repeat: there is no way to rank two incorrect density forecasts such that all users will agree with the ranking.[6] However, it is easy to see that if a forecast coincides with the true data generating process, then it will be preferred by all forecast users, regardless of loss function.[7] More formally, suppose that $p_j(y) = f(y)$, so that a_j^* minimizes the expected loss with respect to the true distribution. Then

$$\int L(a_j^*, y) f(y)\, dy \leq \int L(a_k^*, y) f(y)\, dy, \forall k,$$

which follows immediately from the fact that a_j^* minimizes expected loss over all possible actions, including those which might be chosen under alternative density forecasts.

[6] The result is analogous to Arrow's celebrated impossibility theorem. The ranking effectively reflects a social welfare function, which does not exist.

[7] Granger and Pesaran (1996) independently arrive at a similar result in the context of probability forecasting.

Although simple, the insight that $f(y)$ dominates all other forecasts for all users regardless of loss function is not vacuous. In particular, it suggests a useful direction for evaluating density forecasts. Regardless of loss function, we know that the correct density is weakly superior to all forecasts, which suggests that we evaluate forecasts by assessing whether the forecast densities are correct, that is, whether $\{p_t(y_t|\Omega_t)\}_{t=1}^m = \{f_t(y_t|\Omega_t)\}_{t=1}^m$. If not, we know that some users, depending on their loss functions, could potentially be better served by a different density forecast. We now develop that idea in detail.

3. EVALUATING DENSITY FORECASTS

The task of determining whether $\{p_t(y_t|\Omega_t)\}_{t=1}^m = \{f_t(y_t|\Omega_t)\}_{t=1}^m$ appears difficult, perhaps hopeless, because $\{f_t(y_t|\Omega_t)\}_{t=1}^m$ is never observed, even after the fact. Moreover, and importantly, the true density $f_t(y_t|\Omega_t)$ may exhibit structural change, as indicated by its time subscript. As it turns out, the challenges posed by these subtleties are not insurmountable.

The Probability Integral Transform. Our methods are based on the relationship between the data generating process, $f_t(y_t)$, and the sequence of density forecasts, $p_t(y_t)$, as related through the probability integral transform, z_t, of the realization of the process taken with respect to the density forecast. The probability integral transform is simply the cumulative density function corresponding to the density $p_t(y_t)$ evaluated at y_t,

$$z_t = \int_{-\infty}^{y_t} p_t(u)\,du$$

$$= P_t(y_t).$$

The density of z_t, $q_t(z_t)$, is of particular significance. Assuming that $\partial P_t^{-1}(z_t)/\partial z_t$ is continuous and nonzero over the support of y_t, then, because $p_t(y_t) = \partial P_t(y_t)/\partial y_t$ and $y_t = P_t^{-1}(z_t)$, z_t has support on the unit interval with density

$$q_t(z_t) = \left|\frac{\partial P_t^{-1}(z_t)}{\partial z_t}\right| f_t\left(P_t^{-1}(z_t)\right)$$

$$= \frac{f_t\left(P_t^{-1}(z_t)\right)}{p_t\left(P_t^{-1}(z_t)\right)}.$$

Note, in particular, that if $p_t(y_t) = f_t(y_t)$, then $q_t(z_t)$ is simply the $U(0,1)$ density.

Now we go beyond the one-period characterization of the density of z when $p_t(y_t) = f_t(y_t)$, and characterize both the density and dependence structure of the entire z sequence when $p_t(y_t) = f_t(y_t)$.

PROPOSITION. *Suppose* $\{y_t\}_{t=1}^m$ *is generated from* $\{f_t(y_t|\Omega_t)\}_{t=1}^m$ *where* $\Omega_t = \{y_{t-1}, y_{t-2}, \ldots\}$. *If a sequence of density forecasts* $\{p_t(y_t)\}_{t=1}^m$ *coincides with* $\{f_t(y_t|\Omega_t)\}_{t=1}^m$, *then under the usual condition of a nonzero Jacobian with continuous*

partial derivatives, the sequence of probability integral transforms of $\{y_t\}_{t=1}^m$ *with respect to* $\{p_t(y_t)\}_{t=1}^m$ *is i.i.d.* $U(0,1)$. *That is,*

$$\{z_t\}_{t=1}^m \overset{\text{i.i.d.}}{\sim} U(0,1).$$

PROOF. The joint density of $\{y_t\}_{t=1}^m$ can be decomposed as

$$f(y_m, \ldots, y_1|\Omega_1) = f_m(y_m|\Omega_m) f_{m-1}(y_{m-1}|\Omega_{m-1}) \cdots f_1(y_1|\Omega_1).$$

We therefore compute the joint density of $\{z_t\}_{t=1}^m$ using the change of variables formula:

$$\begin{aligned} q(z_1, z_2, \ldots, z_m) &= \begin{vmatrix} \frac{\partial y_1}{\partial z_1} & \cdots & \frac{\partial y_1}{\partial z_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial z_1} & \cdots & \frac{\partial y_m}{\partial z_m} \end{vmatrix} f_m\left(P_m^{-1}(z_m)|\Omega_m\right) f_{m-1}\left(P_{m-1}^{-1}(z_{m-1})|\Omega_{m-1}\right) \cdots \\ &\quad \cdots \times f_1\left(P_1^{-1}(z_1)|\Omega_1\right) \\ &= \frac{\partial y_1}{\partial z_1} \frac{\partial y_2}{\partial z_2} \cdots \frac{\partial y_m}{\partial z_m} f_m\left(P_m^{-1}(z_m)|\Omega_m\right) f_{m-1}\left(P_{m-1}^{-1}(z_{m-1})|\Omega_{m-1}\right) \cdots \\ &\quad \cdots \times f_1\left(P_1^{-1}(z_1)|\Omega_1\right), \end{aligned}$$

because the Jacobian of the transformation is lower triangular. Thus we have

$$\begin{aligned} q(z_m, \ldots, z_1|\Omega) = \frac{f_m\left(P_m^{-1}(z_m)|\Omega_m\right)}{p_m\left(P_m^{-1}(z_m)\right)} \cdot \frac{f_{m-1}\left(P_{m-1}^{-1}(z_{m-1})|\Omega_{m-1}\right)}{p_{m-1}\left(P_{m-1}^{-1}(z_{m-1})\right)} \cdots \\ \cdots \times \frac{f_1\left(P_1^{-1}(z_1)|\Omega_1\right)}{p_1\left(P_1^{-1}(z_1)\right)}. \end{aligned}$$

Under the assumed conditions, each of the ratios above is a $U(0,1)$ density, the product of which yields an m-variate $U(0,1)$ distribution for $\{z_t\}_{t=1}^m$. Because the joint distribution is the product of the marginals, we have that $\{z_t\}_{t=1}^m$ is distributed i.i.d. $U(0,1)$. □

The intuition for the above result may perhaps be better understood from the perspective of Christoffersen (1998), who shows that a correctly conditionally calibrated interval forecast will provide a hit sequence that is distributed i.i.d. Bernoulli, with the desired success probability.[8] If a sequence of density forecasts is correctly conditionally calibrated, then *every* interval will be correctly conditionally calibrated

[8] The 'hit' series is 1 if the realization is contained in the forecast interval, and 0 otherwise.

and will generate an i.i.d. Bernoulli hit sequence. This fact manifests itself in the i.i.d. uniformity of the corresponding probability integral transforms.

Practical Application. The theory developed thus far suggests that we evaluate density forecasts by assessing whether the probability integral transform series, $\{z_t\}_{t=1}^n$, is i.i.d. $U(0,1)$. Simple tests of i.i.d. $U(0,1)$ behavior are readily available, such as those of Kolmogorov–Smirnov and Cramer–vonMises. Alone, however, such tests are not likely to be of much value in the practical applications that we envision, because they are not constructive; that is, when rejection occurs, the tests generally provide no guidance as to *why*. If, for example, a Kolmogorov–Smirnov test rejects the hypothesis of i.i.d. $U(0,1)$ behavior, is it because of violation of unconditional uniformity, violation of i.i.d., or both? Moreover, even if we know that rejection comes from violation of uniformity, we would like to know more: What, precisely, is the nature of the violation of uniformity, and how important is it? Similarly, even if we know that rejection comes from a violation of i.i.d., what precisely is its nature? Is z heterogeneous but independent, or is z dependent? If z is dependent, is the dependence operative primarily through the conditional mean, or are higher ordered conditional moments, such as the variance, relevant? Is the dependence strong and important, or is i.i.d. an economically adequate approximation, even if strictly false?

Hence we adopt less formal, but more revealing, graphical methods, which we *supplement* with more formal tests. First, as regards unconditional uniformity, we suggest visual assessment using the obvious graphical tool, a density estimate. Simple histograms are attractive in the present context because they allow straightforward imposition of the constraint that z has support on the unit interval, in contrast to more sophisticated procedures such as kernel density estimates with the standard kernel functions. We visually compare the estimated density to a $U(0,1)$, and we compute confidence intervals under the null hypothesis of i.i.d. $U(0,1)$ exploiting the binomial structure, bin-by-bin.

Second, as regards evaluating whether z is i.i.d., we again suggest visual assessment using the obvious graphical tool, the correlogram, supplemented with the usual Bartlett confidence intervals. The correlogram assists with the detection of particular dependence patterns in z and can provide useful information about the deficiencies of density forecasts. For example, serial correlation in the z series indicates that conditional mean dynamics have been inadequately captured by the forecasts. Because we are interested in potentially sophisticated nonlinear forms of dependence, not simply linear dependence, we examine not only the correlogram of $(z-\bar{z})$, but also those of powers of $(z-\bar{z})$. Examination of the correlograms of $(z-\bar{z})$, $(z-\bar{z})^2$, $(z-\bar{z})^3$, and $(z-\bar{z})^4$ should be adequate; it will reveal dependence operative through the conditional mean, conditional variance, conditional skewness, or conditional kurtosis.

4. APPLICATION TO A SIMULATED GARCH PROCESS

Before proceeding to apply our density forecast evaluation methods to real data, it is useful to examine their efficacy on simulated data, for which we know the true data-generating process, We examine a simulated sample of length 8000 from the

t-GARCH(1, 1) process (Bollerslev 1987):

$$y_t = \sqrt{\frac{2h_t}{3}}\, t(6)$$

$$h_t = 0.01 + 0.13y_{t-1}^2 + 0.86h_{t-1}.$$

Both the sample size and the parameter values are typical for financial asset returns.[9] Throughout, we split the sample in half and use the 'in-sample' observations 1 through 4000 for estimation, and the 'out-of-sample' observations 4001 through 8000 for density forecast evaluation.

We will examine the usefulness of our density forecast evaluation methods in assessing four progressively better density forecasts. To establish a benchmark, we first evaluate forecasts based on the naive and incorrect assumption that the process is i.i.d. $N(0, 1)$.[10] That is, in each of the periods 4001–8000, we simply issue the forecast '$N(0, 1)$.'

In Figure 1a we show two histograms of z, one with 20 bins and one with 40 bins.[11] The histograms have a distinct, nonuniform 'butterfly' shape—a hump in the middle and two wings on the sides—indicating that too many of the realizations fall in the middle and tails of the forecast densities relative to what we would expect if the data were really i.i.d. normal. This is exactly what we hope the histograms would reveal, given that the data-generating process is known to be unconditionally leptokurtic.

In Figure 1b we show the correlograms of $(z - \bar{z})$, $(z - \bar{z})^2$, $(z - \bar{z})^3$ and $(z - \bar{z})^4$.[12] The strong serial correlation in $(z - \bar{z})^2$ (and hence $(z - \bar{z})^4$) makes clear another key deficiency of the $N(0, 1)$ forecasts—they fail to capture the volatility dynamics operative in the process. Again, this is what we hope the correlograms would reveal, given our knowledge of the true data-generating process.

Second, we evaluate forecasts produced under the incorrect assumption that the process is i.i.d. but not necessarily Gaussian. We estimate the unconditional distribution from observations 1 through 4000, freeze it, and then issue it as the density forecast in each of the periods 4001 through 8000. Figures 2a and 2b contain the results. The z histogram is now almost perfect (as it must be, apart from estimation error, which is small in a sample of size 4000), but the correlograms correctly continue to indicate neglected volatility dynamics.

Third, we evaluate forecasts that are based on a GARCH(1, 1) model estimated under the incorrect assumption that the conditional density is Gaussian. We use observations 1 through 4000 to estimate the model, freeze the estimated model, and

[9] The conditional variance function intercept of 0.01 is arbitrary but inconsequential; it simply amounts to a normalization of the unconditional variance to 1 $(0.01/(1 - 0.13 - 0.86))$.

[10] The process as specified does have mean zero and variance 1, but it is neither i.i.d. nor unconditionally Gaussian.

[11] The dashed lines superimposed on the histogram are approximate 95% confidence intervals for the individual bin heights under the null that z is i.i.d. $U(0, 1)$.

[12] The dashed lines superimposed on the correlograms are Bartlett's approximate 95% confidence intervals under the null that z is i.i.d.

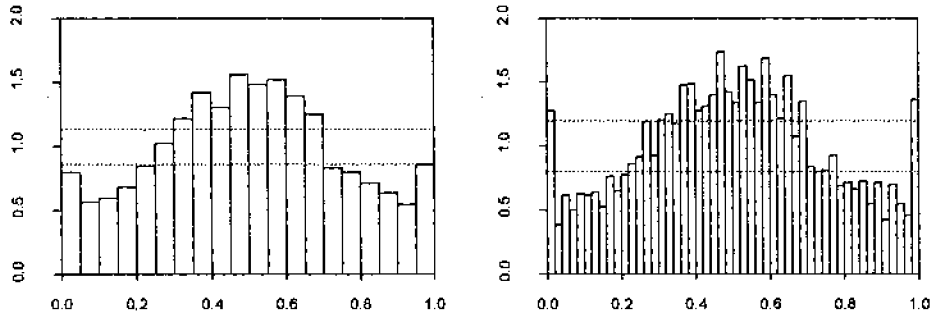

FIGURE 1A

ESTIMATES OF THE DENSITY OF z^*

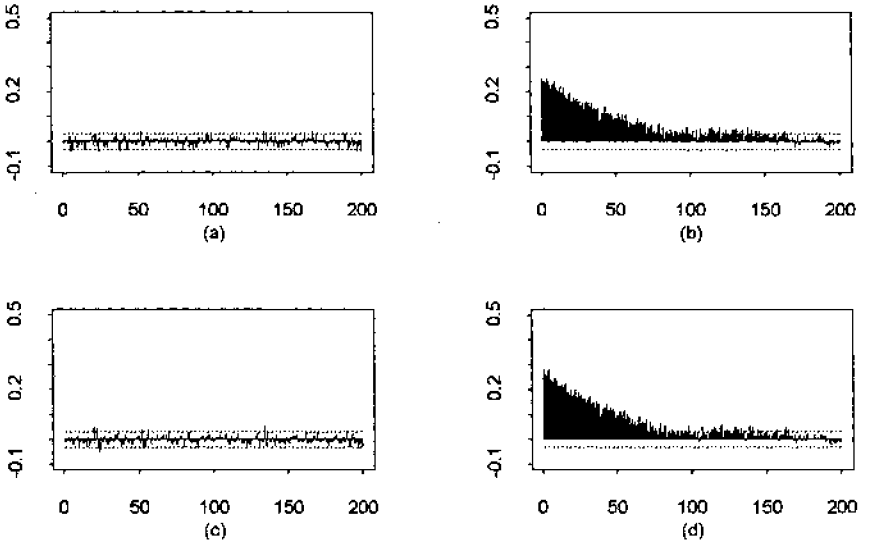

FIGURE 1B

ESTIMATES OF THE AUTOCORRELATION FUNCTIONS OF POWERS OF $z^\dagger$

* Figure 1a: z is the probability integral transform of y with respect to density forecasts produced under the incorrect assumption that y is i.i.d. $N(0,1)$. See text for details.

† Figure 1b: Panels (a) to (d) show sample autocorrelations of $(z-\bar{z})$, $(z-\bar{z})^2$, $(z-\bar{z})^3$ and $(z-\bar{z})^4$.

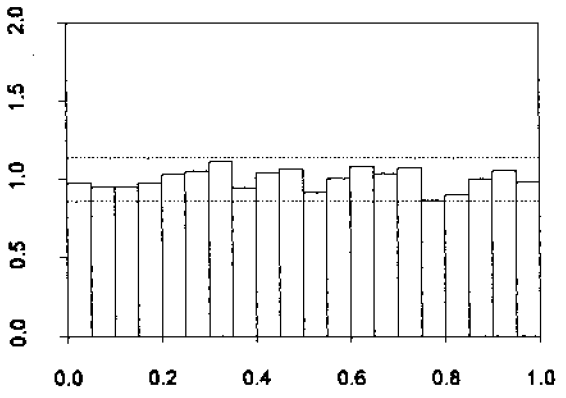

FIGURE 2A

ESTIMATE OF THE DENSITY OF z^*

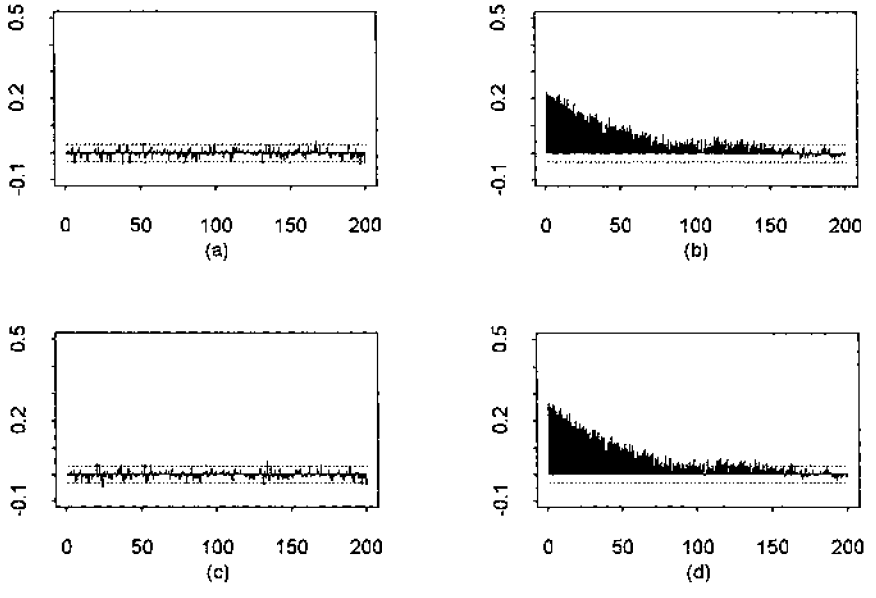

FIGURE 2B

ESTIMATES OF THE AUTOCORRELATION FUNCTIONS OF POWERS OF $z^\dagger$

* Figure 2a: z is the probability integral transform of y with respect to density forecasts produced under the incorrect assumption that y is i.i.d. with density equal to the unconditional density estimated over periods 1–4000. See text for details.

† Figure 2b: Panels (a) to (d) show sample autocorrelations of $(z-\bar{z})$, $(z-\bar{z})^2$, $(z-\bar{z})^3$ and $(z-\bar{z})^4$.

then use it to make (time-varying) density forecasts from 4001 through 8000. Figures 3a and 3b contain the z histograms and correlograms. The histograms are closer to uniform than those of Figure 1a, but they still display slight peaks at either end and a hump in the middle. We would expect to see such a reduction, but not elimination, of the butterfly pattern, because allowance for conditionally Gaussian GARCH effects should account for some, but not all, unconditional leptokurtosis.[13] The correlograms now show no evidence of neglected conditional volatility dynamics, again as expected because the conditionally Gaussian GARCH model delivers consistent estimates of the conditional variance parameters, in spite of the fact that the conditional density is misspecified (Bollerslev and Wooldridge, 1992), so that the estimated model tracks the volatility dynamics well.

Finally, we forecast with an estimated correctly-specified t-GARCH(1, 1) model. We show the z histogram and correlograms in Figures 4a and 4b. Because we are forecasting with a correctly specified model, estimated using a large sample, we would expect that the histogram and correlograms would fail to find flaws with the density forecasts, which is the case.

In closing this section, we note that at each step of the above simulation exercise, our density forecast evaluation procedures clearly and correctly revealed the strengths and weaknesses of the various density forecasts. The results, as with all simulation results, are specific to the particular data-generating process examined, but the process and the sample size were chosen to be realistic for the leading applications in high-frequency finance. This gives us confidence that the procedures will perform well on real financial data, to which we now turn, and for which we do not have the luxury of knowing the true data-generating process.

5. APPLICATION TO DAILY S&P 500 RETURNS

We study density forecasts of daily value-weighted S&P 500 returns, with dividends, from 02/03/62 through 12/29/95. As before, we split the sample into in-sample and out-of-sample periods for model estimation and density forecast evaluation. There are 4133 in-sample observations (07/03/62–12/29/78) and 4298 out-of-sample observations (01/02/79–12/29/95). As before, we assess a series of progressively more sophisticated density forecasts.

As in the simulation example, we begin with an examination of $N(0, 1)$ density forecasts, in spite of the fact that high-frequency financial data are well-known to be unconditionally leptokurtic and conditionally heteroskedastic.[14] In Figures 5a and 5b we show the histograms and correlograms of z. The histograms have the now-familiar butterfly shape, indicating that the S&P realizations are leptokurtic relative to the $N(0, 1)$ density forecasts, and the correlograms of $(z - \bar{z})^2$ and $(z - \bar{z})^4$ indicate that the $N(0, 1)$ forecasts are severely deficient, because they neglect strong conditional volatility dynamics.

Next, we generate density forecasts using an apparently much more sophisticated model. Both the Akaike and Schwarz information criteria select an MA(1)-

[13] Recall that the data generating process is *conditionally*, a well as unconditionally, fat-tailed.

[14] See, among many others, Bollerslev et al. (1992).

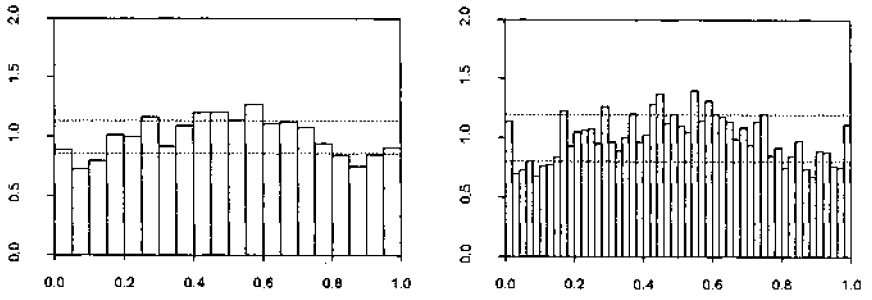

FIGURE 3A

ESTIMATES OF THE DENSITY OF z^*

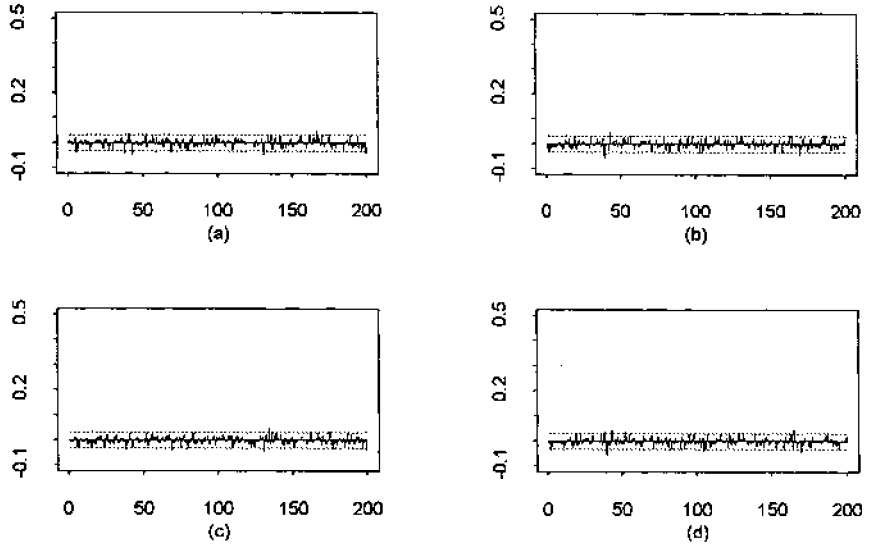

FIGURE 3B

ESTIMATES OF THE AUTOCORRELATION FUNCTIONS OF POWERS OF $z^\dagger$

* Figure 3a: z is the probability integral transform of y with respect to density forecasts produced under the incorrect assumption that y is a conditionally Gaussian GARCH(1,1) process with parameters equal to those estimated over periods 1–4000. See text for details.

† Figure 3b: Panels (a) to (d) show sample autocorrelations of $(z-\bar{z})$, $(z-\bar{z})^2$, $(z-\bar{z})^3$ and $(z-\bar{z})^4$.

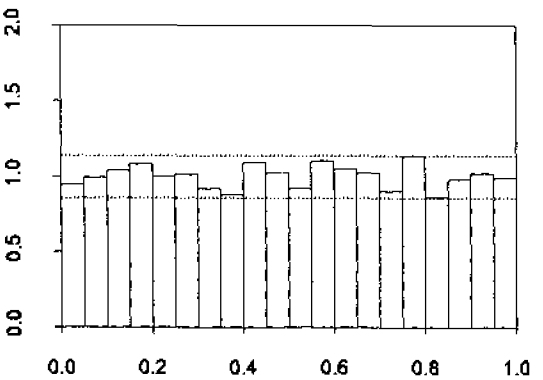

FIGURE 4A

ESTIMATE OF THE DENSITY OF z^*

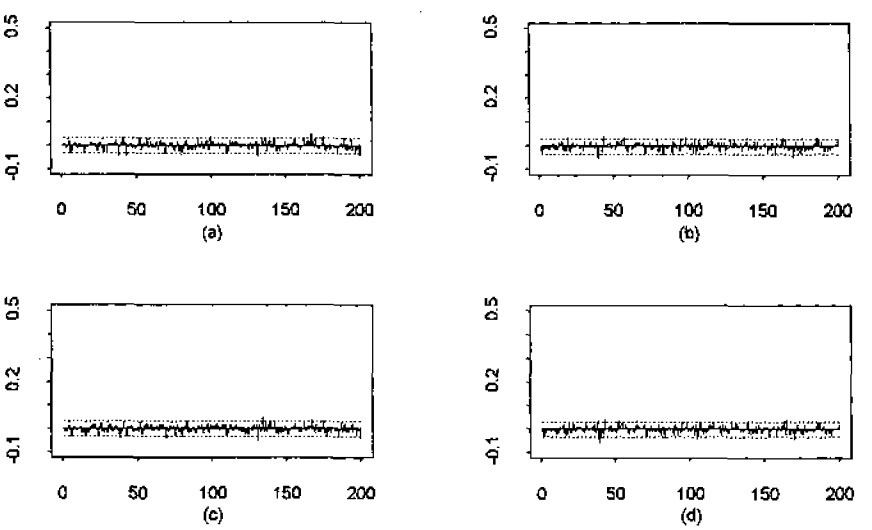

FIGURE 4B

ESTIMATES OF THE AUTOCORRELATION FUNCTIONS OF POWERS OF $z^\dagger$

*Figure 4a: Histogram of z series produced from forecasts of simulated t-GARCH(1, 1) series based on estimated t-GARCH model. We estimate parameters over 1–4000 and forecast over 4001–8000.

†Figure 4b: Panels (a) to (d) show sample autocorrelations of $(z-\bar{z})$, $(z-\bar{z})^2$, $(z-\bar{z})^3$ and $(z-\bar{z})^4$.

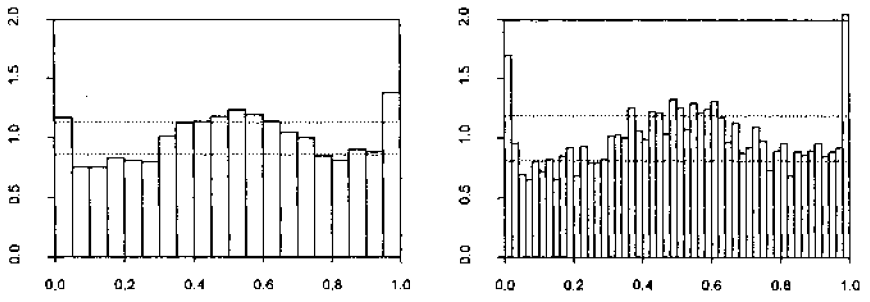

FIGURE 5A

ESTIMATES OF THE DENSITY OF z^*

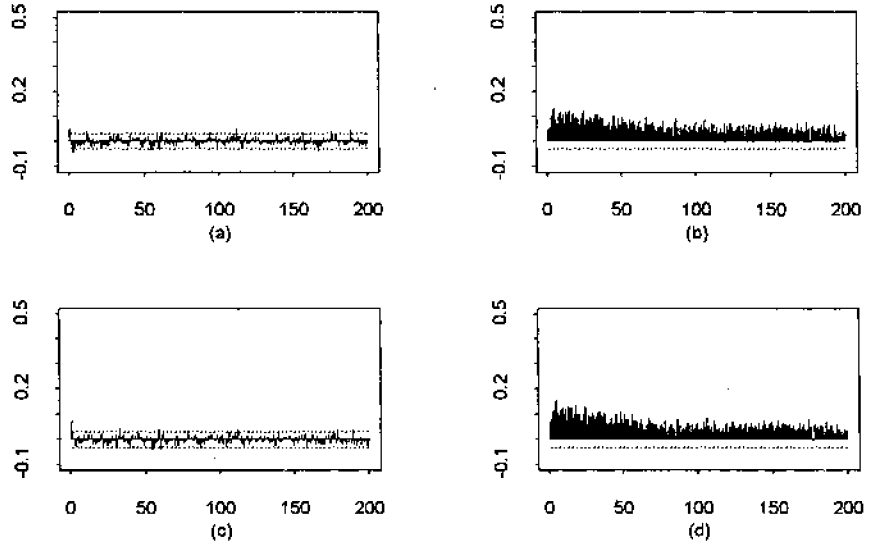

FIGURE 5B

ESTIMATES OF THE AUTOCORRELATION FUNCTIONS OF POWERS OF $z^\dagger$

* Figure 5a: z is the probability integral transform of y with respect to density forecasts produced under the assumption that y is i.i.d. normal. See text for details.

† Figure 5b: Panels (a) to (d) show sample autocorrelations of $(z-\bar{z})$, $(z-\bar{z})^2$, $(z-\bar{z})^3$ and $(z-\bar{z})^4$.

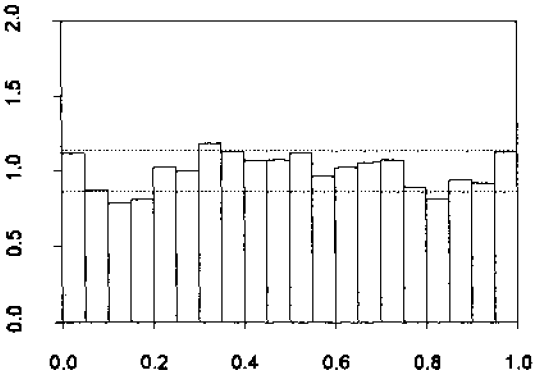

FIGURE 6A

ESTIMATE OF THE DENSITY OF z^*

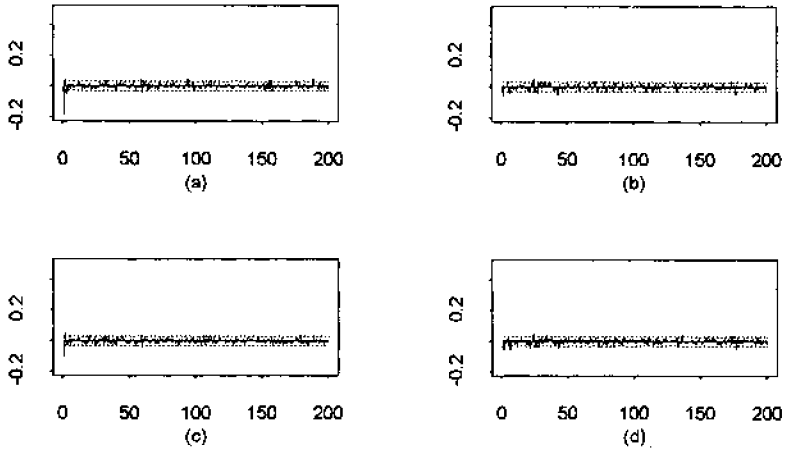

FIGURE 6B

ESTIMATES OF THE AUTOCORRELATION FUNCTIONS OF POWERS OF $z^\dagger$

* Figure 6a: z is the probability integral transform of y with respect to density forecasts produced under the assumption that y is a conditionally Gaussian MA(1)-GARCH(1, 1) process with parameters equal to those estimated from 07/03/62 to 12/29/78. See text for details.

† Figure 6b: Panels (a) to (d) show autocorrelations of $(z-\bar{z})$, $(z-\bar{z})^2$, $(z-\bar{z})^3$ and $(z-\bar{z})^4$.

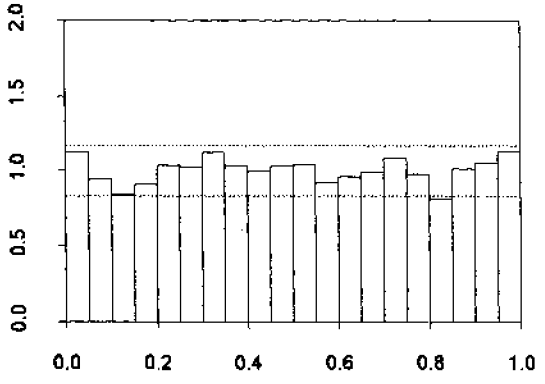

FIGURE 7A

ESTIMATE OF THE DENSITY OF z^*

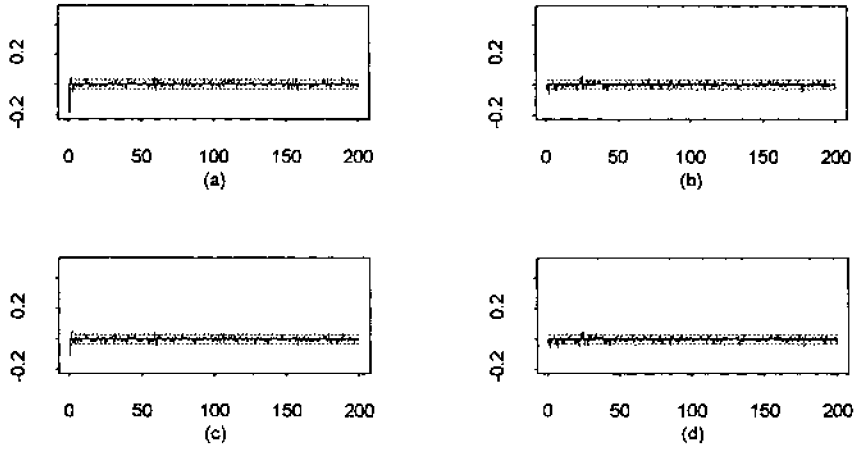

FIGURE 7B

ESTIMATES OF THE AUTOCORRELATION FUNCTIONS OF POWERS OF $z^\dagger$

* Figure 7a: z is the probability integral transform of y with respect to density forecasts produced under the assumption that y is a conditionally Student's t MA(1)-GARCH(1, 1) process with parameters equal to those estimated from 07/03/62 to 12/29/78. See text for details.

† Figure 7b: Panels (a) to (d) show sample autocorrelations of $(z-\bar{z})$, $(z-\bar{z})^2$, $(z-\bar{z})^3$ and $(z-\bar{z})^4$.

GARCH(1,1) model for the in-sample data, which we estimate, freeze, and use to generate out-of-sample density forecasts. Figures 6a and 6b contain the z histograms and correlograms. The histograms are closer to uniform and therefore improved, although they still display a slight butterfly pattern. The correlograms look even better; all evidence of neglected conditional volatility dynamics has vanished.

Finally, we estimate and then forecast with an MA(1)–t-GARCH(1,1) model. We show the z histogram and correlograms in Figures 7a and 7b. The histogram is improved, albeit slightly, and the correlograms remain good.

6. CONCLUDING REMARKS

Let us begin by tying up a couple of loose ends. First, note that notwithstanding the classical feel of most of our discussion, our methods are equally applicable to Bayesian forecasts issued as predictive probability densities. Superficially, it might appear that strict Bayesians would have little interest in our evaluation methods, on the grounds that conditional on a particular sample path and specification of the prior and likelihood, the predictive density simply is what it is, so that there is nothing to evaluate. But such is not the case. A misspecified likelihood, for example, can lead to poor forecasts, whether classical or Bayesian, and density forecast evaluation can help us to flag misspecified likelihoods. It comes as no surprise, therefore, that model checking by comparing predictions to data is emerging as an integral part of modern Bayesian data analysis and forecasting, as highlighted for example in Gelman et al. (1995), and our methods are very much in that spirit.

Second, we wish to emphasize that our decision to ignore parameter estimation uncertainty was intentional. In our framework, the forecasts are the primitives, and we do not require that they be based on a model. This is useful because many density forecasts of interest do not come from models. Such is the case, for example, with the survey density forecasts of inflation recorded in the Survey of Professional Forecasters since 1968; for a description of those forecasts and evaluation using our methods, see Diebold et al. (1998a).[15] A second and very important example of model-free density forecasts is provided by the recent finance literature, which shows how to use options written at different strike prices to extract a model-free estimate of the market's risk-neutral density forecast of returns on the underlying asset (e.g., Aït-Sahalia and Lo, 1998; Soderlind and Svensson, 1997). Moreover, many density forecasts based on estimated models already incorporate the effects of parameter estimation uncertainty; Bayesian predictive density forecasts are a leading example, as are classical density forecasts computed using appropriate bootstrap techniques. Finally, it would seem that samples of the size typically available in high-frequency finance are often so large as to render negligible the effects of parameter estimation uncertainty, as for example in our simulation study. At the

[15] Diebold et al. (1998a) also augment the methods proposed here with resampling procedures to approximate better the finite-sample distributions of the test statistics of interest in small macroeconomic, as opposed to financial, samples.

same time, we readily acknowledge that many model-based density forecasts do not explicitly account for parameter estimation uncertainty, and the sample size sometimes *is* small; for such situations it may be useful to extend our methods to account for parameter estimation uncertainty, in a fashion precisely analogous to West's (1996), and West and McCracken's (1998) extensions of Diebold and Mariano (1995).[16]

Now let us sketch several promising directions for future research. First, it is apparent that our methods can be used to improve defective density forecasts, in a fashion parallel to standard procedures for improving defective point forecasts. Recall that in the case of defective point forecasts we can regress the y's on the $\hat{y}$'s (the point forecasts), and use the estimated relationship to construct improved point forecasts.[17] Similarly, in the context of density forecasts that are defective in that they produce an i.i.d. but nonuniform z sequence, we can exploit the fact that (in period $m+1$, say)

$$\begin{aligned} f_{m+1}(y_{m+1}) &= p_{m+1}(y_{m+1})\, q_{m+1}(P(y_{m+1})) \\ &= p_{m+1}(y_{m+1})\, q_{m+1}(z_{m+1}). \end{aligned}$$

Thus, if we know $q_{m+1}(z_{m+1})$, we would know the actual distribution $f_{m+1}(y_{m+1})$. Because $q_{m+1}(z_{m+1})$ is unknown, we can estimate $\hat{q}_{m+1}(z_{m+1})$ using the historical series of $\{z_t\}_{t=1}^{m}$, and we can use that estimate to construct an improved estimate, $\hat{f}_{m+1}(y_{m+1})$, of the true distribution. Standard density estimation techniques can be used to produce the estimate $\hat{q}_{m+1}(z_{m+1})$.[18]

Second, our methods may be generalized to handle multi-step-ahead density forecasts, so long as we make provisions for serial correlation in z, in a fashion to the usual MA($h-1$) structure for optimal h-step ahead point forecast errors. It may prove most effective to partition the z series into groups for which we expect i.i.d. uniformity if the density forecasts were indeed correct. For instance, for correct 2-step ahead forecasts, the sub-series $\{z_1, z_3, z_5, \ldots\}$ and $\{z_2, z_4, z_6, \ldots\}$ should each be i.i.d. $U(0,1)$, although the full series would not be i.i.d. $U(0,1)$. If a formal test is desired, it may be obtained via Bonferroni bounds, as suggested in a different context by Campbell and Ghysels (1995). Under the assumption that the z series is $(h-1)$-dependent, each of the following h sub-series will be i.i.d.: $\{z_1, z_{1+h}, z_{1+2h}, \ldots\}$, $\{z_2, z_{2+h}, z_{2+2h}, \ldots\}, \ldots, \{z_h, z_{2h}, z_{3h}, \ldots\}$. Thus, a test with size bounded by α can be obtained by performing h tests, each of size α/h, on each of the h sub-series of z, and rejecting the null hypothesis of i.i.d. uniformity if the null is

[16] It would be similarly interesting to see whether and how the decision-theoretic background that we sketch, which requires that agents use density forecasts as if they were known to be the true conditional density, in a fashion similar to West et al. (1993), would change if parameter estimation uncertainty were acknowledged.

[17] Such a regression is sometimes called a Mincer–Zarnowitz regression, after Mincer and Zarnowitz (1969).

[18] In finite samples, of course, there is no guarantee that the 'improved' forecast will actually be superior to the original, because it is based on an estimate of q rather than the true q, and the estimate could be very poor. In the large samples typical in high-frequency finance, however, very precise estimation should be possible.

rejected for *any* of the h sub-series. With the huge high-frequency datasets now available in finance, such sample splitting, although inefficient, is not likely to cause important power deterioration.

Third, the principle that governs the univariate techniques in this paper extends to the multivariate case, as shown in Diebold et al., (1996). Suppose that the variable of interest y is now an $(N \times 1)$ vector, and that we have on hand m multivariate forecasts and their corresponding multivariate realizations. Further suppose that we are able to decompose each period's forecasts into their conditionals, that is, for each period's forecasts we can write

$$p(y_{1t}, y_{2t}, \ldots, y_{Nt} | \Phi_{t-1}) = p(y_{Nt} | y_{N-1,t}, \ldots, y_{1t}, \Phi_{t-1}) \cdots p(y_{2t} | y_{1t}, \Phi_{t-1}) p(y_{1t} | \Phi_{t-1}),$$

where Φ_{t-1} now refers to the past history of $(y_{1t}, y_{2t}, \ldots, y_{Nt})$. Then for each period we can transform each element of the multivariate observation $(y_{1t}, y_{2t}, \ldots, y_{Nt})$ by its corresponding conditional distribution. This procedure will produce a set of N z series that will be i.i.d. $U(0,1)$ individually, and also when taken as a whole, if the multivariate density forecasts are correct. Note that we will have $N!$ sets of z series, depending on how the joint density forecasts are decomposed, giving us a wealth of information with which to evaluate the forecasts. In addition, the univariate formula for the adjustment of forecasts, discussed above, can be applied to each individual conditional, yielding

$$f(y_{1t}, y_{2t}, \ldots, y_{Nt} | \Phi_{t-1}) = \prod_{i=1}^{N} \left[p(y_{it} | y_{i-1,t}, \ldots, y_{1t}, \Phi_{t-1}) q(P(y_{it} | y_{i-1,t}, \ldots, y_{1t}, \Phi_{t-1})) \right] = p(y_{1t}, y_{2t}, \ldots, y_{Nt} | \Phi_{t-1}) q(z_{1t}, z_{2t}, \ldots, z_{Nt} | \Phi_{t-1}).$$

Fourth, we note that our methods may be related to the idea of predictive likelihood, which is based not on the joint density of the sample (the likelihood), but rather the joint density of *future* observations, *conditional* upon the sample (the predictive likelihood).[19,20] Moreover, Clements and Hendry (1993) establish a close link between predictive likelihood and a measure of the accuracy of *point* forecasts that they propose, the generalized forecast error second moment. Investigation of the relationships among such methods and ours is beyond the scope of this paper but appears to be a promising direction for future research.

Fifth, real-time monitoring of adequacy of density forecasts using CUSUM and other recursive techniques should be a simple matter, because under the adequacy hypothesis the z series is i.i.d. $U(0,1)$, which is free of nuisance parameters, thereby enabling trivial calculation of CUSUM bounds.

Finally, if we have information regarding the user's loss function, we should be able to evaluate density forecasts under the relevant loss function, as done in other

[19] For a concise introduction to predictive likelihood, see Bjørnstad (1990).

[20] We thank a referee for making this observation.

forecasting contexts by Diebold and Mariano (1995) and Christoffersen and Diebold (1996, 1997, 1998).

REFERENCES

AÏT-SAHALIA, Y. AND A. LO, "Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices," *Journal of Finance* 53 (1998), 499–547.

BJØRNSTAD, J.F., "Predictive Likelihood: A Review," *Statistical Science* 5 (1990), 242–265.

BOLLERSLEV, T., "A Conditional Heteroskedastic Time Series Model for Speculative Prices and Rates of Return," *Review of Economics and Statistics* 69 (1987), 542–547.

———, R.Y. CHOU, AND K.F. KRONER, "ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence," *Journal of Econometrics* 52 (1992), 5–59.

——— AND J.M. WOOLDRIDGE, "Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariances," *Econometric Reviews* 11 (1992), 143–179.

CAMPBELL, B. AND E. GHYSELS, "Federal Budget Projections: A Nonparametric Assessment of Bias and Efficiency," *Review of Economics and Statistics* 77 (1995), 17–31.

CHATFIELD, C., "Calculating Interval Forecasts," *Journal of Business and Economics Statistics* 11 (1993), 121–135.

CHRISTOFFERSEN, P.F., "Evaluating Interval Forecasts," *International Economic Review* (1998), this issue, pp. 841–862.

——— AND F.X. DIEBOLD, "Further Results on Forecasting and Model Selection Under Asymmetric Loss," *Journal of Applied Econometrics* 11 (1996), 561–572.

——— AND ———, "Optimal Prediction Under Asymmetric Loss," *Econometric Theory* 13 (1997a), 808–817.

——— AND ———, "Cointegration and Long-Horizon Forecasting," *Journal of Business and Statistics* 16 (1998), 450–458.

CLEMEN, R.T., A.H. MURPHY, AND R.L. WINKLER, "Screening Probability Forecasts: Contrasts Between Choosing and Combining," *International Journal of Forecasting* 11 (1995), 133–146.

CLEMENTS, M.P. AND D.F. HENDRY, "On the Limitations of Comparing Mean Square Forecast Errors" (with Discussion), *Journal of Forecasting* 12 (1993), 617–637.

CRNKOVIC, C. AND J. DRACHMAN, *A Universal Tool to Discriminate Among Risk Measurement Techniques* (New York: J.P. Morgan, 1996).

DIEBOLD, F.X. AND LOPEZ, J.A., "Forecast Evaluation and Combination," in G.S. Maddala and C.R. Rao, eds., *Handbook of Statistics* (Amsterdam: North-Holland, 1996), 241–268.

——— AND R.S. MARIANO, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13 (1995), 253–263.

———, A.S. TAY, AND K.D. WALLIS, "Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters," prepared for R.F. Engle and H. White, eds., *Festschrift in Honor of C.W.J. Granger*, http://www.ssc.upenn.edu/~diebold/, 1998a.

———, J. HAHN, AND A. TAY, "Real-Time Multivariate Density Forecast Evaluation and Calibration: Monitoring the Risk of High-Frequency Returns on Foreign Exchange," Manuscript, Department of Economics, University of Pennsylvania (1998b).

EFRON, B. AND R.J. TIBSHIRANI, *An Introduction to the Bootstrap* (New York: Chapman and Hall, 1993).

GELMAN, A., J.B. CARLIN, H.S. STERN, AND D.B. RUBIN, *Bayesian Data Analysis* (London: Chapman and Hall, 1995).

GRANGER, C.W.J. AND M.H. PESARAN, "A Decision Theoretic Approach to Forecast Evaluation," Manuscript, Department of Economics, UCSD and Cambridge University, 1996.

MINCER, J. AND V. ZARNOWITZ, "The Evaluation of Economic Forecasts," in J. Mincer, ed., *Economic Forecasts and Expectations* (New York: National Bureau of Economic Research, 1969).

ROSENBLATT, M., "Remarks on a Multivariate Transformation," *Annals of Mathematical Statistics* 23 (1952), 470–472.

SODERLIND, P. AND L.E.O. SVENSSON, "New Techniques to Extract Market Expectations from Financial Instruments," National Bureau of Economic Research Working Paper 5877, 1997.

WALLIS, K.F., Comment on J.H. Stock and M.W. Watson, "A Procedure for Predicting Recessions with Leading Indicators," in J.H. Stock and M.W. Watson, eds., *Business Cycles, Indicators and Forecasting* (Chicago: University of Chicago Press for NBER, 1993, 153–156).

WEST, K.D., "Asymptotic Inference About Predictive Ability," *Econometrica* 64 (1996), 1067–1084.

——— AND M.W. MCCRACKEN, "Regression-Based Tests of Predictive Ability," *International Economic Review* (1998), this issue, pp. 817–840.

———, H.J. EDISON, AND D. CHO, "A Utility-Based Comparison of Some Models of Exchange Rate Volatility," *Journal of International Economics* 35 (1993), 23–45.