FEDERAL RESERVE BANK OF PHILADELPHIA

Ten Independence Mall
Philadelphia, Pennsylvania 19106-1574
(215) 574-6428, www.phil.frb.org

Working Papers

Research Department

WORKING PAPER NO. 97-6

EVALUATING DENSITY FORECASTS

Francis X. Diebold
Department of Economics, University of Pennsylvania
Visiting Scholar, Federal Reserve Bank of Philadelphia

Todd A. Gunther
Department of Economics, University of Pennsylvania

Anthony S. Tay
Department of Economics, University of Pennsylvania
Department of Economics and Statistics
National University of Singapore

May 1997

WORKING PAPER NO. 97-6
EVALUATING DENSITY FORECASTS

Francis X. Diebold
Department of Economics, University of Pennsylvania
Visiting Scholar, Federal Reserve Bank of Philadelphia

Todd A. Gunther
Department of Economics, University of Pennsylvania

Anthony S. Tay
Department of Economics, University of Pennsylvania
Department of Economics and Statistics
National University of Singapore

May 1997

The views expressed here are those of the authors and not necessarily those of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. Helpful discussion was provided by seminar participants at Michigan, Penn, Princeton, and the UCSD Conference on Time Series Analysis of High-Frequency Financial Data. We are especially grateful to Clive Granger, Jin Hahn, Atsushi Inoue, Hashem Pesaran, Ken Wallis, and Tao Zha. All remaining inadequacies are ours alone. We thank the National Science Foundation, the Sloan Foundation, the University of Pennsylvania Research Foundation, and the National University of Singapore for support.

ABSTRACT

We propose methods for evaluating and improving density forecasts. We focus primarily on methods that are applicable regardless of the particular user's loss function, though we take explicit account of the relationships between density forecasts, action choices, and the corresponding expected loss throughout. We illustrate the methods with a detailed series of examples, and we discuss extensions to improving and combining suboptimal density forecasts, multistep-ahead density forecast evaluation, multivariate density forecast evaluation, monitoring for structural change and its relationship to density forecasting, and density forecast evaluation with known loss function.

Correspondence to:
F. X. Diebold
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104
fdiebold@mail.sas.upenn.edu

1. Introduction

Prediction occupies a distinguished position in econometrics; 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 treat interval forecasts (e.g., Chatfield, 1993; Christoffersen, 1997) and probability forecasts (e.g., Wallis, 1993; Clemen, Murphy and Winkler, 1995).

Remarkably 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 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.[1] In fact, improvements in computer technology have rendered the provision of credible density forecasts increasingly straightforward.

Second, until recently there was little demand for density forecasts; historically, point and interval forecasts proved 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 future

[1] See, for example, the discussion of construction of density forecasts using resampling techniques surveyed in Berkowitz and Kilian (1996).

portfolio values and to tracking certain aspects of the densities.[2] In fact, 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. Using, for example, Christoffersen's (1997) 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.

In this paper we treat density forecasting and density forecast evaluation. We explicitly account for the relationships between the density forecast, the action choice, and the resultant expected loss, and we evaluate density forecasts in their entirety. For the reasons discussed above, there is very little literature on the topic, although professional interest is increasing, as evidenced by the contemporaneous and independent work of Granger and Pesaran (1996) and Crnkovic and Drachman (1996), which is closely related and highly complementary. To the best of our knowledge, there is no other literature that bears directly on the issues we address. In section 2, we present a detailed statement and discussion of the problem, and we provide the theoretical underpinnings of the methods that we introduce subsequently. In section 3 we present

[2] Moreover, scalar distillations of densities into "risk measures," such as value at risk, are often inadequate for risk assessment. See, for example, Guthoff, Pfingsten and Wolf (1997).

methods of density forecast evaluation when the loss function is unknown, which is likely to be important in practice. In section 4, we provide a series of detailed examples to density forecasting in environments with time-varying volatility. In section 5 we discuss extensions to improving suboptimal density forecasts, evaluating multi-step and multivariate density forecasts, monitoring for structural change when density forecasting, and evaluating density forecasts when the loss function is known. We conclude in section 6.

2. Loss Functions, Action Choices, and Density Forecast Evaluation

Statement and Discussion of the Problem

Let $f_t(y_t|\Omega_t)$ be the data-generating process governing a series y_t, where Ω_t contains all relevant conditioning variables. Let $\{y_t\}_{t=1}^m$ denote the corresponding series of realizations.[3] Suppose that a series of 1-step-ahead density forecasts of y_t is available, made over m periods. The forecasts might be available in closed form, or they might be constructed from large sets of simulated draws from the densities. In any case, we represent the series of forecasts by $\{p_t(y_t|\Phi_t)\}_{t=1}^m$ where $\Phi_t \subseteq \Omega_t$ is the information set used by the forecaster. 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 and Φ_t should be understood.

We wish to evaluate these density forecasts by considering the historical performance of the forecaster. The solution would seem to lie in the application of the well-known fact that if a sequence of realizations corresponds to iid draws from a fixed density, then the probability integral transforms of the realizations with respect to the density are iid U(0,1). However, there

[3] 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.

are two complicating factors. First, we are not dealing with iid realizations from a fixed density. Instead, the realizations come from a series of generally time-varying and dependent densities. Second, we may need to account for the forecast user's loss function when evaluating density forecasts. We will address both complications.

The Decision Environment

The intimate relationship between density forecasts, action choices, and loss functions is relevant when evaluating density forecasts. In this section, we consider this relationship in a one-period context, and hence we drop the time subscripts for notational convenience.

Each forecast user has a loss function $L(a, y)$, where a refers to an action choice. The action choice need not be a prediction of y. For example, a may refer to the amount of insurance coverage to purchase, with y representing the realized loss. The user chooses an action to minimize expected loss computed using the density believed to be the data-generating process. If she believes that $p(y)$, the prediction from the forecaster for the current period, is the correct density, then she chooses an action a^* by solving[4]

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

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 distribution function, which we will call the loss distribution, and which depends only on the action choice. Expected loss with respect to the true data-generating process is

[4] We assume a unique minimizer. A sufficient condition is that A be compact and that L be convex in a.

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

The effect of the density forecast on the user's expected loss is easily seen. A density forecast translates into a loss distribution. Two different forecasts will, in general, lead to different action choices and, hence, different loss distributions. Different forecasts that lead to the same action choice are, to the user, equivalent. A "good" forecast will lead to an action choice that results in a comparatively low expected loss 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 forecast user will prefer the forecast $p_j(y)$ if the expected loss associated with using $p_j(y)$ is less than the expected loss associated with using $p_k(y)$ instead; that is, if

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

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

Ideally, we would like to find a way of assigning to each forecast a score $D(p_j)$, constructed from the history of density forecasts and realizations, which would measure the divergence of the realization from the forecast density, such that all users, *regardless of loss function*, would prefer the forecast with the lower divergence. This would allow us to rank the forecasts. Unfortunately, the following proposition shows that no such divergence score exists.

Proposition 1. Let $f(y)$ be the density of y, p_j be a density forecast of y, and $a \in A$ be the set of action choices. Let a_j^* be the optimal action based on forecast p_j. Then no score $D(\cdot)$ exists such that for arbitrary forecast densities p_j and p_k, both distinct from f , and for all possible loss functions $L(a, y)$,

$$D(p_j) \geq D(p_k) \Leftrightarrow \int L(a_j^*, y) f(y) dy \geq \int L(a_k^*, y) f(y) dy .$$

Proof. In order to establish the result, it is sufficient 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

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

while

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

That is, user 1 does better on average under forecast k, while user 2 does better under forecast j. It is straightforward to construct such an example. Suppose 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 $\int y\, p(y)\, dy$ and $\int y^2\, p(y)\, dy$ respectively. That is, user 1 bases her action choice on the mean, with higher expected loss occurring with larger errors in the forecast mean, while user 2's actions and expected losses 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 that, in turn, leads to a loss distribution with 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.[5] However, if a forecast coincides with the true data-generating process, it will be the forecast, among the class of forecasts that uses the same information set, that minimizes expected loss for all forecast users, regardless of their loss function. We show this in the following proposition.[6]

Proposition 2. Suppose that the forecast, $p_j(y)$, is identical to the data-generating process, i.e., $p_j(y) = f(y)$, and hence a_j^* minimizes the actual expected loss. Then for all possible density forecasts $p_k(y)$, choosing the action according to the true density gives the least expected loss among all forecasts with the same information set. That is,

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

Proof. The result follows immediately from the assumption that a_j^* minimizes expected loss over all possible actions, including those that might be chosen under alternative densities of y. □

The case where different information sets are used to produce the forecasts is somewhat less clear cut. For instance, it can be shown that a correct density forecast will be weakly preferred to another correct density based on a smaller information set, but not much can be said

[5] The result is analogous to Arrow's celebrated impossibility theorem. D(•) is effectively a social welfare function, which the proposition shows does not exist.

[6] Granger and Pesaran (1996) independently arrive at a similar result.

about two correct forecasts based on information sets where neither subsumes the other. Nonetheless, the preceding propositions do suggest a useful direction for evaluating density forecasts. Even without taking loss functions into consideration, we know that the correct density, correct relative to some information set, is weakly superior to all forecasts with either the same or smaller information set. This suggests evaluating forecasts by assessing whether the forecast densities, conditioned on some information set, are correct, i.e., whether $\{p_t(y_t|\Phi_t=\varphi_t)\}_{t=1}^m = \{f_t(y_t|\Phi_t=\varphi_t)\}_{t=1}^m$. We do so by assessing whether the realizations $\{y_t\}_{t=1}^m$, come from the forecast densities $\{p(y_t|\Phi_t=\varphi_t))\}_{t=1}^m$. If not, we know that some users, depending on their loss functions, could potentially be better served by a different density forecast.

3. Evaluating Density Forecasts

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 forecasts. The following lemma describes the distribution, $q_t(z_t)$, of the probability integral transform.

Lemma 1. Let y_t be the variable of interest with $f_t(y_t)$ as its density, and let $p_t(y_t)$ represent a density forecast of y_t. Let the variable z_t be the probability integral transform of y_t with respect to $p_t(y_t)$. That is,

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

$$= P_t(y_t).$$

Then assuming that $\frac{\partial P_t^{-1}(z_t)}{\partial z_t}$ is continuous and non-zero over the support of y_t, z_t will have support on the unit interval with density function

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

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

Proof: Use the facts that $p_t(y_t) = \frac{\partial P_t(y_t)}{\partial y_t}$ and $y_t = P_t^{-1}(z_t)$. □

Note that in the well-known special case where $p_t(y_t) = f_t(y_t)$, $q_t(z_t)$ is simply the U(0,1) density. While this special case is all we will need for the evaluation of density forecasts, the general form of the density $q_t(z_t)$ will be useful for the purpose of improving suboptimal density forecasts.

As pointed out earlier, the realizations are typically dependent draws from a sequence of dependent densities, not iid draws from a single density. We thus need a complete probabilistic characterization of the z series when the density forecasts are correct. We do so in the following proposition. Note that for ease of exposition, we restrict the information set at each time period to comprise only the history of past y_t; that is, $\Omega_t = \{ y_{t-1}, y_{t-2}, ... \}$.[7]

Proposition 3. Suppose a series $\{y_t\}_{t=1}^m$ is generated from $\{f_t(y_t|\Omega_t)\}_{t=1}^m$ where $\Omega_t = \{ y_{t-1}, y_{t-2}, ... \}$. If a sequence of density forecasts $\{p_t(y_t)\}_{t=1}^m$ coincides with

[7] The proposition also holds when $\Omega_t = \{ x_t, y_{t-1}, x_{t-1}, y_{t-2}, ... \}$, where x refers to some variable that can help in forecasting y, although the proof must be modified slightly.

$\{f_t(y_t|\Omega_t)\}_{t=1}^m$, then under the usual conditions of a non-zero Jacobian and the continuity of its partial derivatives, we have

$$\{z_t\}_{t=1}^m = \left\{\int_{-\infty}^{y_t} p_t(u)du\right\}_{t=1}^m \overset{iid}{\sim} U(0,1).$$

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

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

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

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

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

$$= \frac{\partial y_1}{\partial z_1}\frac{\partial y_2}{\partial z_2}\frac{\partial y_m}{\partial z_m} f_m(P_m^{-1}(z_m)|\Omega_m)\, f_{m-1}(P_{m-1}^{-1}(z_{m-1})|\Omega_{m-1})\ldots f_1(y_1^{-1}(z_1)|\Omega_1),$$

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

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

From Lemma 1, 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 iid U(0,1). □

Practical Application

The theory developed thus far suggests that we use the series $\{z_t\}_{t=1}^m$, derived from the history of realizations and density forecasts, to evaluate forecasts. We simply check whether $\{z_t\}_{t=1}^m$ is iid U(0,1). As noted earlier, we effectively check whether a random sample $\{y_t\}_{t=1}^m$ is drawn from density $p(y)$ by taking the probability integral transform of the sample with respect to $p(y)$. In this paper, we effectively consider a sample $\{y_t\}_{t=1}^m$ that is a realization from a sequence of time-varying and dependent densities, but Proposition 3 reveals that if a forecaster manages to capture the sequence of densities that forms the true data-generating process, the probability integral transforms are still iid U(0,1).

Simple tests of iid U(0,1) behavior are readily available. Recall, for example, that if $z_t \sim U(0,1)$ then $-2 \log z_t \sim \chi^2_2$.[8] Hence, if z is iid U(0,1), then $\sum_{t=1}^{m} -2 \log z_t \sim \chi^2_{2m}$, which yields a simple significance test. Alternatively, we could perform any of the various well-known tests for uniformity, such as a runs test or a Kolmogorov-Smirnov test, all of which are actually joint tests of uniformity and iid.

Such tests, however, are not likely to be of much value in practical applications, because they are not constructive; that is, when rejection occurs, the tests generally provide no guidance as to *why*. If, for example, such a statistic rejects the hypothesis of iid U(0,1) behavior, is it because of violation of unconditional uniformity, violation of iid, or both? Moreover, even if we

[8] See Johnson and Kotz (1970).

know that rejection comes from violation of uniformity, we'd 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 iid, 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 iid an adequate approximation, even if strictly false?

The nonconstructive nature of tests of iid U(0,1) behavior, and the nonconstructive nature of related separate tests of iid and U(0,1), make us eager to adopt more revealing methods of exploratory data analysis. First, as regards evaluating 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. The estimated density can be visually compared to a U(0,1).

Second, as regards evaluating whether z is iid, we again suggest visual assessment using the obvious graphical tool, the correlogram. Because we're interested in potentially sophisticated forms of dependence—not just linear dependence—we examine not only the correlogram of $(z-\bar{z})$, but also those of powers of $(z-\bar{z})$. In practice, examination of the correlograms of $(z-\bar{z})$, $(z-\bar{z})^2$, $(z-\bar{z})^3$ and $(z-\bar{z})^4$ is usually adequate; it will reveal dependence operative through the conditional mean, conditional variance, conditional skewness, or conditional kurtosis.

The presence of particular forms of dependence in z can be informative in guiding forecasters and users about how to improve density forecasts. For instance, serial correlation in the z series may indicate that the mean dynamics have been inadequately modeled by the forecaster. A caveat, however, is that there is in general no one-to-one correspondence between the type of dependence found in z and the dependence in y missed by the forecasts. For example, assume that the true data-generating process is GARCH(p,q). Even if a forecaster correctly specifies the conditional variance function and perfectly estimates its parameters, there may be dependence in z if the forecaster assumes the wrong conditional density.

In closing this section, we note that our methods of density forecast evaluation have interesting parallels with well-known methods of evaluating point and interval forecasts. It is well known, for example, that under certain conditions optimal point forecasts have corresponding 1-step-ahead errors that are iid with zero mean. In addition, Christoffersen (1997) shows that the hit series corresponding to a series of correctly calibrated (1-α)% interval forecasts is iid Bernoulli(1-α).[9] Our methods, in parallel, are based on the fact that the z series corresponding to a series of correct density forecasts is iid U(0,1).

4. Examples

Here we illustrate our methods with four examples. Throughout, we use a t-GARCH(1,1) data-generating process,

$$y_t | \Omega_{t-1} \sim h_t^{1/2} t(v)$$

$$h_t = \omega + \alpha y_{t-1}^2 + \beta h_{t-1},$$

[9] Christoffersen defines the "hit" at any time to be 1 if the realization is in the forecasted interval, and zero otherwise.

with parameters $\omega = 0.01$, $\alpha=0.13$, $\beta=0.86$ and v=6 chosen to mimic values typically obtained when fitting GARCH models to asset returns. We simulate a series of length 8000, again chosen to mimic typical analyses of high-frequency financial data, and we plot it in Figure 1. The persistence in conditional variance is visually obvious.

First, we evaluate forecasts that are based on an incorrect assumption that the process is iid N(0,1).[10] We make forecasts from period 4001 through 8000. In Figure 2 we show the two histograms of z (one with 20 bins and the other with 40 bins) and the correlograms of $(z-\bar{z})$, $(z-\bar{z})^2$, $(z-\bar{z})^3$ and $(z-\bar{z})^4$.[11] The histograms and correlograms indicate poor density forecasts. In particular, note the obvious non-uniformity of the histogram and the strong autocorrelations in squares and fourths of $(z-\bar{z})$. Notice that the peaks at the ends of the histogram are even more accentuated as we take a larger number of bins.

Second, we evaluate forecasts that are based on the incorrect assumption that the process is iid, but we base the forecasts on an estimate of the correct unconditional distribution rather than a N(0,1). We estimate the unconditional distribution as the empirical distribution of observations 1-4000. Figure 3 contains the results. The histogram is, of course, almost perfect, but the correlograms, which show strong persistence in the squares and fourths of z, correctly continue to indicate neglected volatility dynamics.

Third, we evaluate forecasts that are based on an estimated GARCH(1,1) model, but with the incorrect assumption that the conditional density is Gaussian. We use observations 1-4000

---

[10] The process as specified does have mean zero and variance 1, but it is neither iid 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 iid U(0,1).

for estimation, and we forecast from 4001 through 8000. The estimated conditional variance function is

$$\underset{}{h_t} = \underset{(0.0012)}{0.0182} + \underset{(0.0050)}{0.1245}\, y_{t-1}^2 + \underset{(0.0040)}{0.8586}\, h_{t-1}.$$

Figure 4 contains the z histogram and correlograms. The correlograms show no evidence of neglected conditional volatility dynamics. The histogram is improved, but it still displays slight peaks at either end and a hump in the middle. This pattern is expected, because allowance for conditionally Gaussian GARCH effects should account for some, but not all, unconditional leptokurtosis, given that the data-generating process is conditionally fat-tailed.

Finally, we forecast with an estimated GARCH(1,1) model based on the correct specification. The estimated conditional variance function is

$$\underset{}{h_t} = \underset{(0.0017)}{0.0125} + \underset{(0.0062)}{0.0850}\, y_{t-1}^2 + \underset{(0.0090)}{0.8569}\, h_{t-1},$$

and the estimated degrees of freedom are 6.2466. In Figure 5 we show the z histogram and correlograms. Because we are forecasting with a correctly specified model, estimated using a large sample, we expect both the histogram and the correlograms to look good, and they do.

5. Extensions

Improving Density Forecasts

We have approached forecast evaluation from an historical perspective, evaluating the ability of a forecaster based on past realizations. The intent, of course, is to gauge the likely

future accuracy of the forecaster based on past performance, assuming that the relationship between the correct density and the forecaster's predictive density remains fixed. Given that we observe systematic errors in the historical forecasts, we may wish to simply reject the forecast. It may also turn out that the errors are irrelevant to the user, a case we further examine when we explicitly account for the user's loss function. Nevertheless, it is possible to take the errors into consideration when using the current forecast, just as it is possible to do so in the point forecast case. In the point forecast case, for example, we can regress the y's on the $\hat{y}$'s, the predicted values, and use the estimated relationship to construct an adjusted point forecast.[12]

In the context of density forecasts, we can construct a similar procedure by rewriting the relationship in Lemma 1. Suppose that the user is in period m and possesses a density forecast of y_{m+1}. From Lemma 1, we have

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

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, an estimate $\hat{q}_{m+1}(z_{m+1})$ can be formed using the historical series of $\{z_t\}_{t=1}^{m}$, and an estimate of the true distribution $\hat{f}_{m+1}(y_{m+1})$ can then be constructed. If the sample $\{z_t\}_{t=1}^{m}$ turns out to be iid, then standard density estimation techniques can be used to produce the estimate $\hat{q}_{m+1}(z_{m+1})$. Otherwise, the estimation of $q_{m+1}(z_{m+1})$ becomes a non-trivial matter, which we defer to future research.

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

Multi-Step-Ahead Density Forecasts

The evaluation of h-step ahead forecasts can also be evaluated using our methods, except that provisions must be made for autocorrelation in z. This is analogous to expecting MA(h-1) autocorrelation structures for optimal h-step ahead point forecast errors. In this case, it will probably be easier to partition the z series into groups for which we expect iid uniformity if the 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 iid U(0,1), although the full series would not be iid 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 iid: $\{z_1, z_{1+h}, z_{1+2h}, \ldots\}$, $\{z_2, z_{2+h}, z_{2+2h}, \ldots\}$, ..., $\{z_h, z_{2h}, z_{3h}, \ldots\}$. 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 z sub-series, and rejecting the null hypothesis of iid uniformity if the null is rejected for *any* of the h sub-series. With the huge high-frequency data sets now available in finance, such sample splitting, although inefficient, is not likely to cause important power deterioration.

Multivariate Density Forecasts

The principle that governs the univariate techniques in this paper readily extends to the multivariate case, as shown in Tay (1997). Suppose that the variable of interest y is now an (N x 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, i.e., for each period's forecasts we can write

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

Then for each period we can transform each element of the multivariate observation $(y_{1t}, y_{2t},..., y_{Nt})$ by its corresponding conditional distribution. This procedure will produce a set of N z-series that will be iid U(0,1) individually, and also when taken as a whole. 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

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

Monitoring for Structural Change When Density Forecasting

Real-time monitoring of adequacy of density forecasts using CUSUM techniques is a simple matter, because under the adequacy hypothesis the z series is iid U(0,1), which is free of nuisance parameters. In particular, if $z_t \overset{iid}{\sim} U(0,1)$, then $z_t \overset{iid}{\sim} \left(\frac{1}{2}, \frac{1}{12}\right)$, so that asymptotically in m,

$$\sum_{t=1}^{m} z_t \sim N\left(\frac{m}{2}, \frac{m}{12}\right),$$

which yields the approximate 95% confidence interval for the CUSUM,

$$\sum_{t=1}^{m} z_t \in \left[\frac{m}{2} \pm 1.96\sqrt{\frac{m}{12}} \right].$$

Similar calculations hold for the CUSUM of squares. Trivial calculations reveal that under the adequacy hypothesis $z_t^2 \overset{iid}{\sim} \left(\frac{1}{3}, \frac{4}{45} \right)$, so that asymptotically in m,

$$\sum_{t=1}^{m} z_t^2 \sim N\left(\frac{m}{3}, \frac{4m}{45} \right),$$

which yields the approximate 95% confidence interval for the CUSUM of squares,

$$\sum_{t=1}^{m} z_t^2 \in \left[\frac{m}{3} \pm 1.96\sqrt{\frac{4m}{45}} \right].$$

Evaluating Density Forecasts Using a Specific Loss Function

If a series of density forecasts has been systematically in error, it may still be the case that for a particular user, depending on her loss function, the systematic errors may be irrelevant. To be precise, the forecast may be such that the action choice induced by the forecast, a_p^*, minimizes the user's actual expected loss.[13] In such cases, which we now consider, the user's

[13] Because we have assumed a unique optimal action choice, $a_p^* = a_f^*$.

loss function can be incorporated into the evaluation process, as is done in other forecasting contexts by Diebold and Mariano (1995) and Christoffersen and Diebold (1996, 1997a, 1997b).

Consider a density forecast series, $\{p_t(y_t)\}_{t=1}^{m}$, and the corresponding action series, $\{a_{p,t}^{*}\}_{t=1}^{m}$, of a particular user. The series of action choices results in a series of potential losses, $L(a_{p,t}^{*}, y_t)$. We would like to compare each period's realized loss with that period's expected loss under the optimal action choice $E_{f,t}[L(a_{f,t}^{*}, y_t)]$. The expected difference will be positive unless $a_{p,t}^{*} = a_{f,t}^{*}$.

Unfortunately, we are unable to evaluate $E_{f,t}[L(a_{f,t}^{*}, y_t)]$. Instead, we will have to use an estimate of $E_{p,t}[L(a_{p,t}^{*}, y_t)]$ as a proxy for $E_{f,t}[L(a_{f,t}^{*}, y_t)]$. We can then compute the difference,

$$d_t = L(a_{p,t}^{*}, y_t) - \frac{1}{m}\sum_{t=1}^{m} L(a_{p,t}^{*}, y_t).$$

Under the joint null hypothesis that the series of density forecasts is optimal relative to the user's loss function and that the forecaster correctly specifies the expected loss in each period, i.e., $E_{f,t}[L(a_{f,t}^{*}, y_t)] = E_{p,t}[L(a_{p,t} = a_{f,t}^{*}, y_t)]$, we have $E[d_t]=0$, which can be tested in the same way that Diebold and Mariano (1995) test whether two point forecasts are equally accurate under the relevant loss function.

6. Summary and Concluding Remarks

We have provided a characterization of optimal density forecasts, and we have proposed methods for evaluating whether reported density forecasts coincide with the true sequence of conditional densities. In addition to studying the decision problem associated with density forecasting and showing how to use the series of probability integral transforms to judge the

adequacy of a series of density forecasts, we also discussed how to improve a suboptimal density forecast by using information on previously issued density forecasts and subsequent realizations, how to evaluate multistep and multivariate density forecasts, and how to monitor for structural change when density forecasting. We did all of this in a framework not requiring specification of the loss function, but when information on the relevant loss function is available, we also showed how to evaluate a density forecast with respect to that loss function.

In closing, we note that the methods used to produce the density forecasts being evaluated are inconsequential for application of our methods; the density forecasts could be produced by any method, including Bayesian methods. This is so in spite of the fact that our methods have a classical feel. Superficially, it would seem that strict Bayesians would have little interest in our evaluation methods; conditional on a particular sample path and specification of the prior and likelihood, the predictive density simply is what it is, and there's 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 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, Carlin, Stern and Rubin (1995), and our methods are very much in that spirit.

References

Berkowitz, J. and L. Kilian (1996), "Recent Developments in Bootstrapping Time Series," Manuscript, Department of Economics, University of Pennsylvania.

Campbell, B. and Ghysels, E. (1995), "Federal Budget Projections: A Nonparametric Assessment of Bias and Efficiency," *Review of Economics and Statistics*, 77, 71-31.

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

Christoffersen, P.F. (1997), "Evaluating Interval Forecasts," *International Economic Review*, Forthcoming.

Christoffersen, P.F. and Diebold, F.X. (1996), "Further Results on Forecasting and Model Selection Under Asymmetric Loss," *Journal of Applied Econometrics*, 11, 561-572.

Christoffersen, P.F. and Diebold, F.X. (1997a), "Optimal Prediction Under Asymmetric Loss," *Econometric Theory*, forthcoming.

Christoffersen, P.F. and Diebold, F.X. (1997b), "Cointegration and Long-Horizon Forecasting," manuscript, Department of Economics, University of Pennsylvania.

Clemen, R.T., A.H. Murphy and R.L. Winkler (1995), "Screening Probability Forecasts: Contrasts Between Choosing and Combining," *International Journal of Forecasting*, 11, 133-146.

Crnkovic, C. and Drachman, J. (1996), "A Universal Tool to Discriminate Among Risk Measurement Techniques," Manuscript, J.P. Morgan & Co.

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

Diebold, F.X. and R.S. Mariano (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253-263.

Gelman, A, Carlin, J.B., Stern, H.S., Rubin, D.B. (1995), *Bayesian Data Analysis*. London: Chapman and Hall.

Granger, C.W.J. and M.H. Pesaran (1996), "A Decision Theoretic Approach to Forecast Evaluation," Manuscript, Departments of Economics, University of California, San Diego and Cambridge University.

Guthoff, A., Pfingsten, A. and Wolf, J. (1997), "On the Compatibility of Value at Risk, Other Risk Concepts, and Expected Utility Maximization," Discussion Paper 97-01, Institute for Kreditwesen, University of Münster, Germany.

Johnson, N.L. and S. Kotz (1970), *Continuous Univariate Distributions*, Volume Two. Boston: Houghton-Mifflin.

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

Tay, A.S. (1997), "Evaluating Multivariate Density Forecasts," Manuscript, Department of Economics, University of Pennsylvania.

Wallis, K.F. (1993), 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, 153-156.

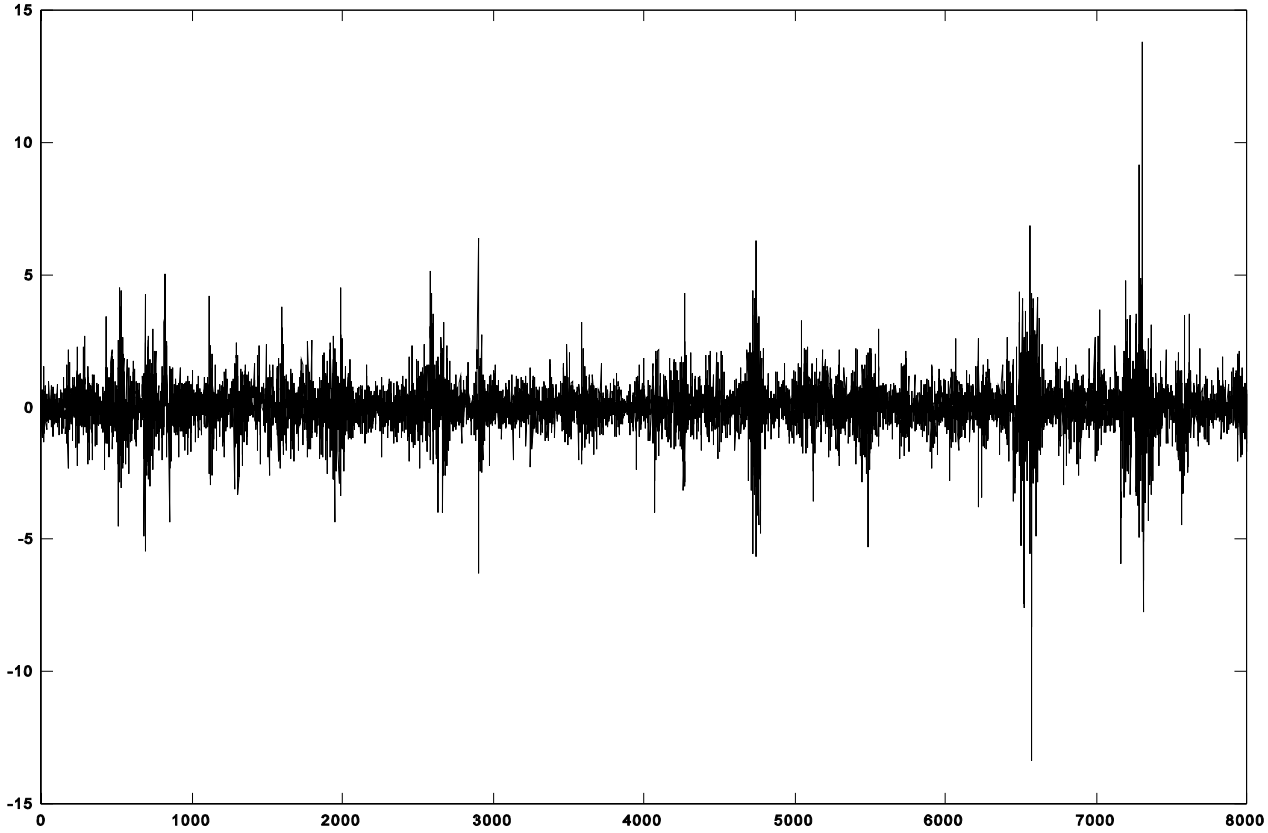

Figure 1. Plot of simulated t-GARCH(1,1) series, 8000 observations. The parameters are ω=0.01, α=0.13, and β=0.86. The standardized realizations are distributed t(6).

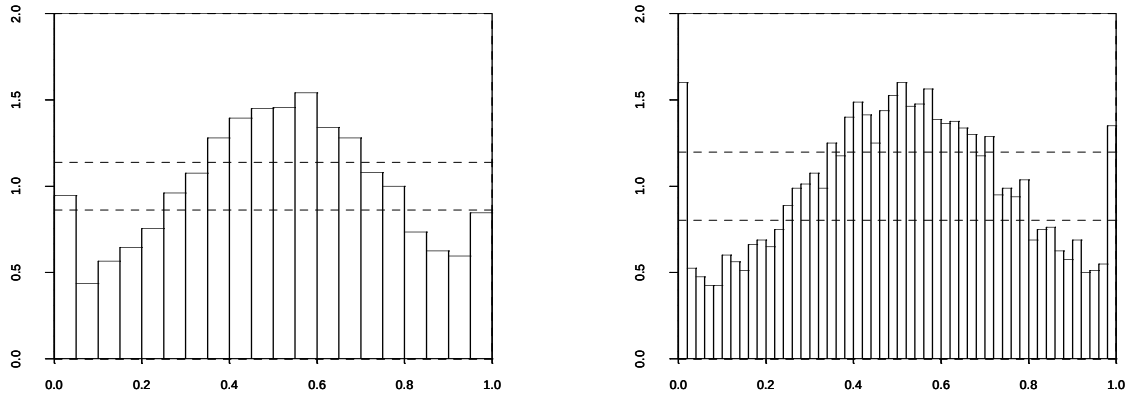

Figure 2a. Histograms (with twenty and forty bins) of z series produced from forecasts of simulated t-GARCH(1,1) series based on an incorrect assumption that the series iid N(0,1).

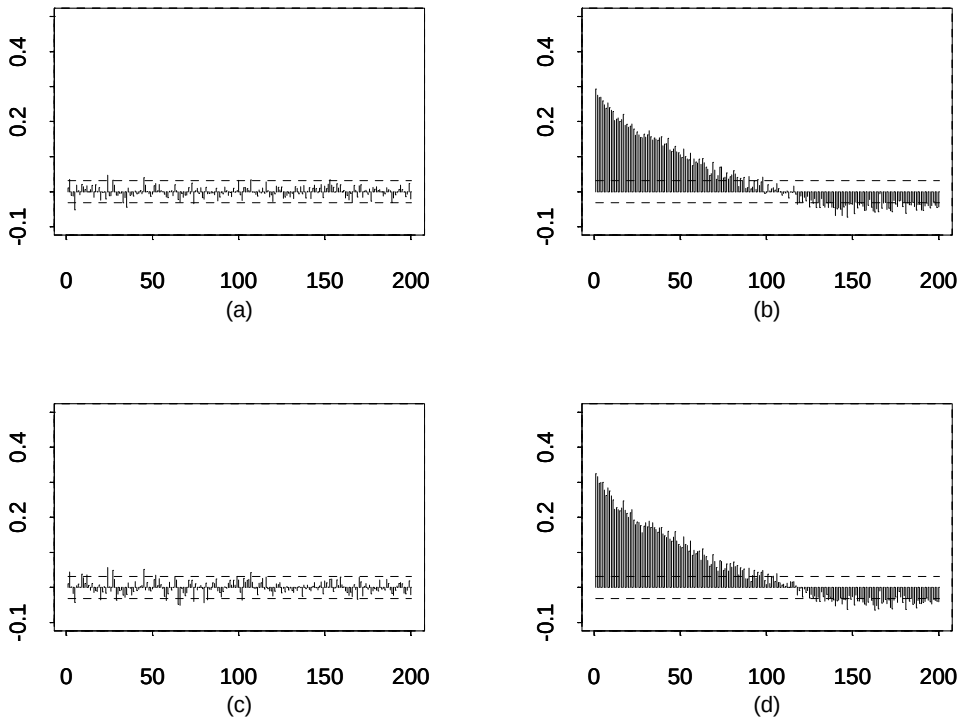

Figure 2b. Autocorrelations of various z series produced from forecasts of simulated t-GARCH(1,1) series based on an incorrect assumption that the series is iid N(0,1). Panels (a) to (d) show autocorrelations of $(z-\bar{z})$, $(z-\bar{z})^2$, $(z-\bar{z})^3$ and $(z-\bar{z})^4$ respectively.

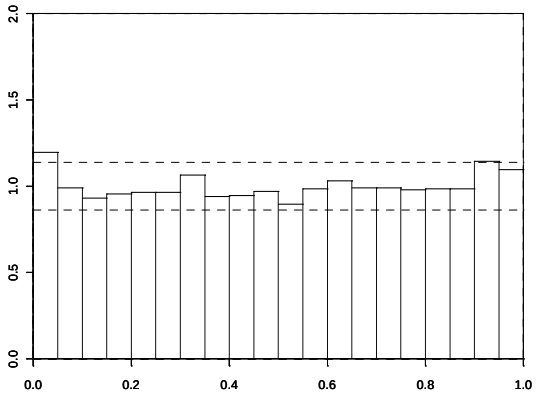

Figure 3a. Histogram of z series produced from forecasts based on an incorrect assumption that the process is iid with density equal to the empirical unconditional density estimated over periods 1-4000.

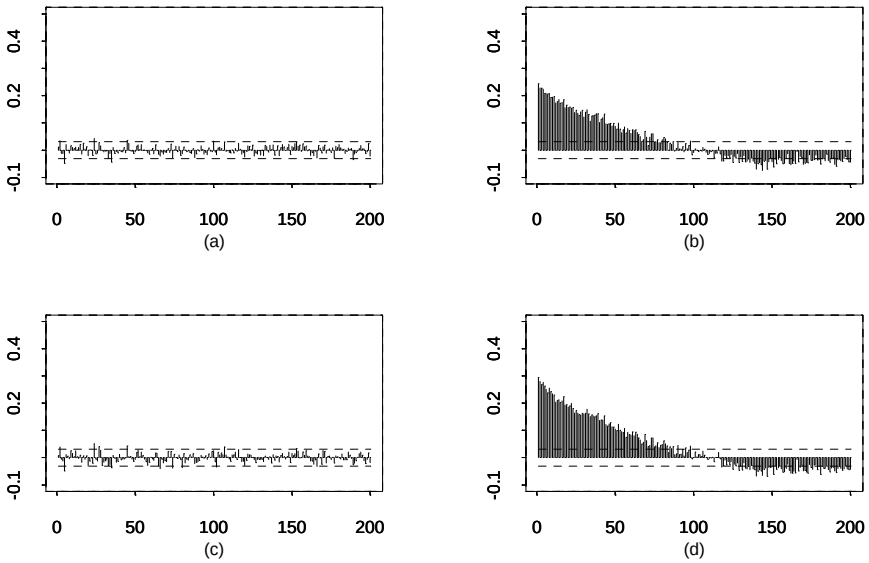

Figure 3b. Autocorrelations of various z-series produced from forecasts based on an incorrect assumption that the process is iid with the empirical unconditional density estimated over periods 1-4000. Panels (a) to (d) show autocorrelations of $(z-\bar{z})$, $(z-\bar{z})^2$, $(z-\bar{z})^3$ and $(z-\bar{z})^4$ respectively.

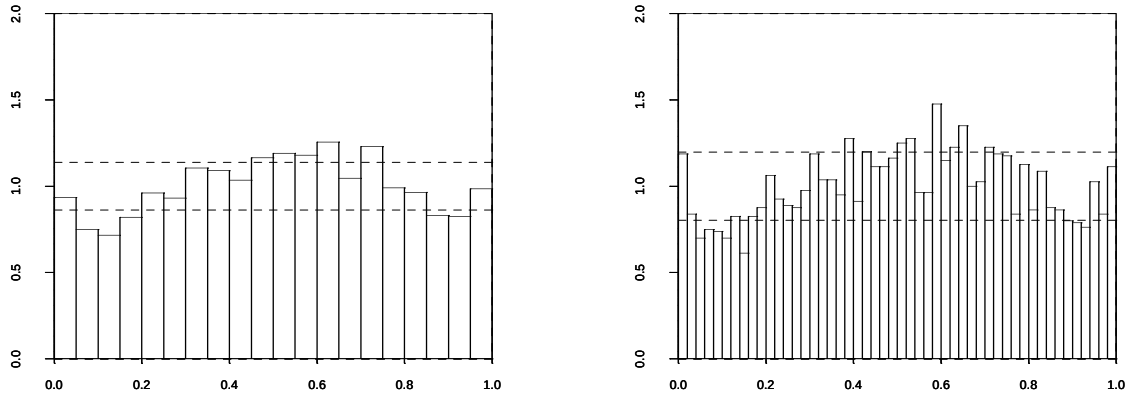

Figure 4a. Histograms of z series produced from forecasts of simulated t-GARCH(1,1) series based on estimated conditionally Gaussian GARCH(1,1). We estimate parameters over 1-4000 and forecast over 4001-8000.

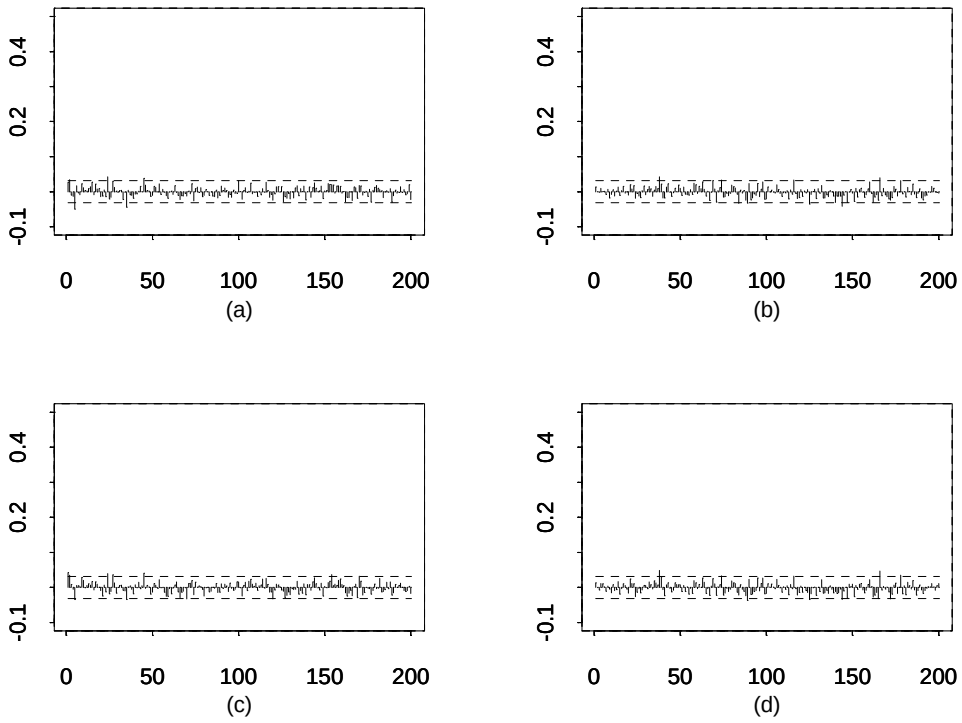

Figure 4b. Autocorrelations of various z series produced from forecasts of simulated t-GARCH(1,1) series based on estimated Gaussian GARCH(1,1) model. Panels (a) to (d) show autocorrelations of $(z-\bar{z})$, $(z-\bar{z})^2$, $(z-\bar{z})^3$ and $(z-\bar{z})^4$ respectively.

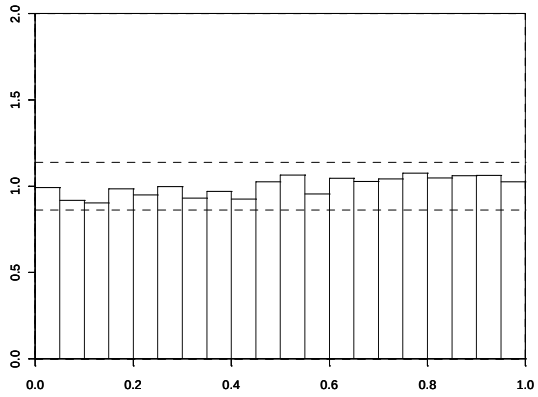

Figure 5a. Histogram of z series produced from forecasts of simulated t-GARCH(1,1) series based on estimated t-GARCH model. We estimate parameters over 1-4000 and forecast over 4001-8000.

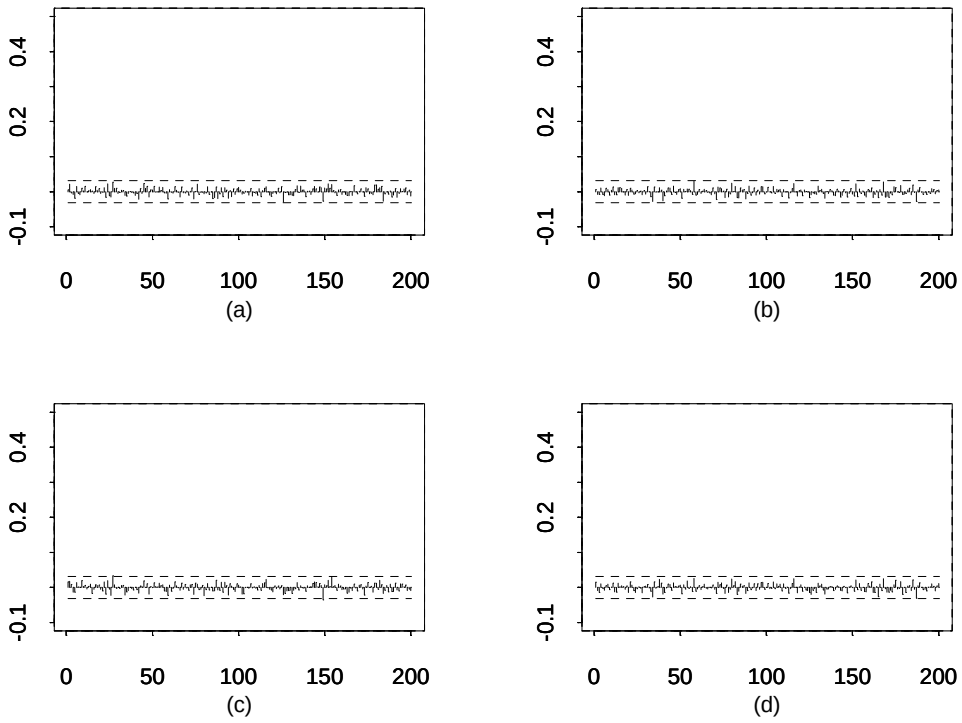

Figure 5b. Autocorrelations of various z-series produced from forecasts of simulated t-GARCH(1,1) series based on estimated t-GARCH model. Panels (a) to (d) show autocorrelations of $(z-\bar{z})$, $(z-\bar{z})^2$, $(z-\bar{z})^3$ and $(z-\bar{z})^4$ respectively.