Chapter 11

Interval and Density Forecast Evaluation

11.1 Interval Forecast Evaluation

Interval forecast evaluation is largely, but not entirely, subsumed by density forecast evaluation. There is a simple method for absolute interval forecast evaluation that must be mentioned. It is of great practical use, and moreover is establishes the proper notion of a 1-step-ahead interval forecast error (which should be unforecastable), and which then translates into the proper notion of a 1-step-ahead density forecast error (which should also be unforecastable).

11.1.1 Absolute Standards

On Correct Unconditional vs. Conditional Coverage

A $(1-\alpha)\%$ interval is correctly *unconditionally* calibrated if it brackets the truth $(1-\alpha)\%$ of the time, on average over the long run. But an interval can be correctly unconditionally calibrated and still poorly *conditionally* calibrated insofar as it's poorly calibrated at any given time, despite being correct on average. In environments of time-varying conditional variance, for example, constant-width intervals may be correctly unconditionally calibrate, but they cannot be correctly conditionally calibrated, because they

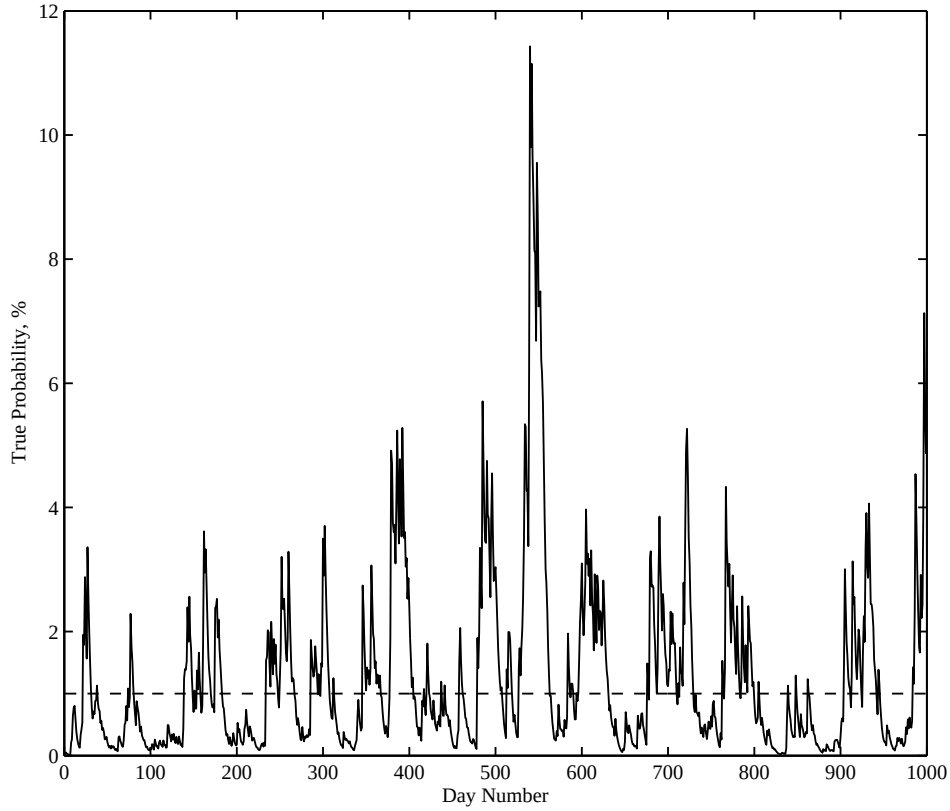

Figure 11.1: True Exceedance Probabilities of Nominal one-sideed 1% Interval When Volatility is Persistent. We simulate returns from a realistically-calibrated dynamic volatility model. We plot the series of true conditional exceedance probabilities, which we infer from the model. For visual reference we include a horizontal line at the desired 1% probability level. Adapted from Andersen et al. 2013.

fail to tighten appropriately in low-volatility times and widen appropriately in high-volatility times. Intervals can be completely mis-calibrated, correctly calibrated unconditionally but not conditionally, or correctly conditionally calibrated (which automatically implies correct conditional calibration). Figure 11.1 says it all

Christoffersen's Absolute Interval Forecast Evaluation

Christoffersen (1998) considers likelihood-ratio tests of correct $(1-\alpha)\%$ conditional coverage. Define the sequence of hit indicators of a 1-step-ahead forecast interval (the "hit series") as

$$I_t^{(1-1\alpha)} = 1\{\text{realized } y_t \text{ falls inside the interval}\}$$

Under the null hypothesis of correct conditional calibration,

$$I_t^{(1-\alpha)} \sim iid\ Bernoulli(1-\alpha).$$

Note well the two-part characterization. The hit series must have the correct mean, $(1-\alpha)$, which corresponds to correct unconditional calibration. But there's more: the hit series must also be *iid*.[1] When both hold, we have correct conditional calibration. Conversely, rejection of the *iid Bernoulli* null could be due to rejection of *iid*, rejection of the *Bernoulli* mean of $(1-\alpha)$, or both. Hence it is advisable to use constructive procedures, which, when rejections occur, convey information as to *why* rejections occur.

On Testing *iid* in Forecast Evaluation

Note that in (1-step) forecast evaluation we're always testing some sort of 1-step error for *iid* (or at least white noise) structure.

For point forecasts the forecast errors are immediately at hand. If they're dependent, then, in general, today's error is informative regarding tomorrow's likely error, and we could we could generally use that information to adjust today's point forecast to make it better, which means something is wrong.

For interval forecasts, the correct notion of "error" is the hit sequence, which is readily constructed. If the hit sequence is dependent, then, in general, today's hit value (0 or 1) is informative regarding tomorrow's likely hit value, and we could we could generally use that information to adjust today's interval forecast to make it better conditionally calibrated, which means something is wrong.

Soon in section 11.2.1 we will introduce yet another generalized "forecast error" series for *density* forecasts, which again should be *iid* if all is well.

[1] In h-step-ahead contests the hit sequence need not be *iid* but should have h-dependent structure.

11.1.2 Relative Standards

Little studied. It seems clear that for two correctly conditionally calibrated interval forecasts, one should prefer the one with shorter average length. But, just as with bias-variance tradeoffs for point forecast evaluation, presumably one should willing to accept a little mis-calibration in exchange for a big length reduction. One would have to define a loss function over miscalibration and length.

11.2 Density Forecast Evaluation

11.2.1 Absolute Standards

Theory

We seek to characterize the properties of a density forecast that is optimal with respect to an information set, that is, a density forecast that coincides with the true conditional expectation.

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.

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,

$$\begin{aligned} z_t &= \int_{-\infty}^{y_t} p_t(u)du \\ &= P_t(y_t). \end{aligned}$$

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

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

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

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 $\Omega_t = \{y_{t-1}, y_{t-2}, ..$ 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 non-zero 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 *iid* $U(0,1)$. That is,

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

The intuition for the above result may perhaps be better understood from the perspective of Christoffersen's method for interval forecast evaluation, If a sequence of density forecasts is correctly conditionally calibrated, then *every* interval will be correctly conditionally calibrated and will generate an *iid* Bernoulli hit sequence. This fact manifests itself in the *iid* 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, $\{z_t\}_{t=1}^{m}$, is *iid* $U(0,1)$. Simple tests of *iid* $U(0,1)$ 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 *iid* $U(0,1)$ behavior, is it because of violation of unconditional uniformity, violation of *iid*, 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 *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 dependance strong and important, or is *iid* 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 *iid* $U(0,1)$ exploiting the binomial structure, bin-by-bin.

Second, as regards evaluating whether z is *iid*, 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 modeled 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.

11.2.2 Additional Discussion

Parameter Estimation Uncertainty

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, such as those from surveys, do not come from models. 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. Moreover, many density forecasts based on estimated models already incorporate the effects of parameter estimation uncertainty, for example by using simulation techniques. Finally, sample sizes are often so large as to render negligible the effects of parameter estimation uncertainty, as for example in our simulation study.

Improving Mis-Calibrated Density Forecasts

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 case we can regress the y's on the $\hat{y}$'s (the point forecasts) and use the estimated relationship to construct improved point forecasts. Similarly, in the context of density forecasts that are defective in that they produce an *iid* but non-uniform z sequence, we can exploit the fact that (in period $m+1$, say)

$$f_{m+1}(y_{m+1}) = p_{m+1}(y_{m+1})\ q_{m+1}(P(y_{m+1}))$$

$$= p_{m+1}(y_{m+1})\ q_{m+1}(z_{m+1}).$$

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 obtain an estimate $\hat{q}_{m+1}(z_{m+1})$ using the historical series of $z_{t\,t=1}^{m}$, and we 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 $\hat{q}_{m+1}(z_{m+1})$.[2]

Multi-Step Density Forecasts

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 *iid* uniformity if the density forecasts were indeed correct. For instance, for correct 2-step ahead forecasts, the sub-series $z_1, z_3, z_5, ...$ and $z_2, z_4, z_6, ...$ 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

[2] 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 large samples, however, very precise estimation should be possible.

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*: $\{z_1, z_{1+h}, z_{1+2h}, ...\}$, $\{z_2, z_{2+h}, z_{2+2h}, ...\}$, ..., $\{z_h, z_{2h}, z_{3h}, ...\}$. 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 *iid* uniformity if the null is 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.

11.2.3 Relative Standards

The time-t one-step-ahead point predictive likelihood is

$$P_t = p_{t,t-1}(y_t)$$

It is simply the height of the earlier-made density forecast, $p_{t,t-1}(\cdot)$ at the realized value, y_t. The full predictive likelihood is then

$$P = \prod_{i=1}^{N} P_t.$$

We can rank density forecasts using P. The sequence of density forecasts with the largest P is the the sequence for which the subsequently-observed realizations were most likely.

11.3 Stock Return Density Forecasting

11.3.1 A Preliminary GARCH Simulation

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:

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

$$h_t = .01 + .13y_{t-1}^2 + .86h_{t-1}.$$

Both the sample size and the parameter values are typical for financial asset returns.[3] 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 *iid* $N(0,1)$.[4] That is, in each of the periods 4001-8000, we simply issue the forecast "$N(0,1)$."

In Figure *** we show two histograms of z, one with 20 bins and one with 40 bins.[5] The histograms have a distinct non-uniform "butterfly" shape – a hump in the middle and two wings on the sides – indicating that too many of the realizations fall in middle and in the tails of the forecast densities relative to what we would expect if the data were really iid normal. This is exactly what we hope the histograms would reveal, given that the data-generating process known to be unconditionally leptokurtic.

In Figure *** we show the correlograms of $(z - \bar{z})$, $(z - \bar{z})^2$, $(z - \bar{z})^3$ and $(z - \bar{z})^4$.[6] The strong serial correlation in $(z - \bar{z})^2$ (and hence $(z - \bar{z})^4$)

[3]The conditional variance function intercept of .01 is arbitrary but inconsequential; it simply amounts to a normalization of the unconditional variance to 1 (.01/(1-.13-.86)).

[4]The process as specified does have mean zero and variance 1, but it is neither *iid* nor unconditionally Gaussian.

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

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

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 *iid* 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 *** and ***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 then use it to make (time-varying) density forecasts from 4001 through 8000. Figures *** and *** contain the z histograms and correlograms. The histograms are closer to uniform than those of Figure ***, 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.[7] 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, despite the fact that the conditional density is misspecified, so that the estimated model tracks the volatility dynamics well.

Finally, we forecast with an estimated correctly-specified $t-GARCH(1,1)$

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

model. We show the z histogram and correlograms in Figures *** and ***. 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.

11.3.2 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. In Figures *** and *** 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 $MA(1)-GARCH(1,1)$ model for the in-sample data, which we estimate, freeze, and use to generate out-of-sample density forecasts.

Figures *** and *** contain the z histograms and correlograms. The histograms are closer to uniform and therefore improved, although they still display 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 *** and ***. The histogram is improved, albeit slightly, and the correlograms remain good.

11.4 Exercises, Problems and Complements

1. xxx

11.5 Notes