

“Forecast Evaluation of Explanatory Models of Financial Return Variability”

by Genaro Sucarrat

4 June 2008.

This paper presents some ideas on how to evaluate models of volatility (conditional standard deviation) or “variability” (the conditional mean of the squared return). The author takes very seriously the fact that we cannot directly observe the conditional variance, and he presents some interesting simulation results for a reasonable data generating process, which examine the choice of loss function (MSE, MAE, R^2) and the choice of test (Diebold-Mariano 1995, White 2000, Hansen 2005) to compare competing models.

Some of the choices made and conclusions drawn by the author suggest to me that his framework for thinking about this problem is not ideal. Rather than presenting simulation evidence and hoping that the conclusions apply more generally, theory from the existing literature shows that some of the conclusions do hold, whilst others do not. Two papers in particular are relevant to the author’s study, Hansen and Lunde (2006) and Patton (2006); the author cites both of these papers, but he either does not agree with their results or does not see how they relate to this problem.

Moreover, the author appears to be very concerned about “methodological and practical” issues raised by using high-frequency data to evaluate forecasts of lower-frequency variables. A lot of theoretical and applied research is currently being done in this area, and so a serious concern about this may have wide-ranging impact. I got the impression that the author had thought hard about this problem, but I was not able to pinpoint the problems that the author was so concerned about. This is perhaps a fault in my own reasoning, but a simple example that clearly illustrated these problem(s) would have been helpful.

COMMENTS

On the use of MAE versus MSE loss functions: one of the author’s main conclusions is that the MAE loss function ($L(e) = |e|$) is preferable to the MSE loss function ($L(e) = e^2$) for small sample sizes. Whilst this is true for the particular simulations presented by the author this is not generally true. In showing this I think I also cover an important difference in how the author and I view “volatility” or “variability”.

The author’s model in his simulation study is:

$$\begin{aligned}r_t &= bx_t + \sigma_t z_t \\ \sigma_t^2 &= \omega + \beta \sigma_{t-1}^2 + \alpha \sigma_{t-1}^2 z_{t-1}^2 + cy_t \\ z_t, x_t &\sim iid N(0, 1) \\ y_t &\sim iid Bernoulli(p)\end{aligned}$$

where x_t, y_t, z_t are independent of each other. The author uses the competing models’ predictions for r_t^2 as a means of comparison. The process for r_t^2 from above is:

$$\begin{aligned}r_t^2 &= b^2 x_t^2 + \sigma_t^2 z_t^2 + 2b\sigma_t x_t z_t \\ &= \{b^2 x_t^2 + \sigma_t^2\} + \{2b\sigma_t x_t z_t + \sigma_t^2 (z_t^2 - 1)\} \\ &\equiv E[r_t^2 | I_t] + \eta_t, \text{ with } E[\eta_t | I_t] = 0\end{aligned}$$

The “variability forecasts” from a model, denoted $(\hat{r}_{mt}^2 + \hat{\sigma}_{mt}^2)$ by the author, can be interpreted as models for $E[r_t^2|I_t] = (E[r_t|I_t])^2 + V[r_t|I_t]$.

Rankings of these models based on MSE loss or the R^2 from Mincer-Zarnowitz regressions, and using r_t^2 as a proxy for $E[r_t^2|I_t]$, can be shown to be correct as $T \rightarrow \infty$, see Hansen and Lunde (2006b), and this is borne out in the author’s simulation results. Rankings of these models based on MAE or the “kurtosis loss function” can be shown to be *incorrect*, in general, see Patton (2006).

Consider MAE loss: the optimal variability forecast under this loss function is the one that minimises $E[|r_t^2 - f| | I_t]$, which is $Median[r_t^2|I_t]$. However the models considered by the author are models for the conditional *mean* and the conditional *variance*, and when combined as $(\hat{r}_{mt}^2 + \hat{\sigma}_{mt}^2)$ they are models for $E[r_t^2|I_t]$. If $Median[r_t^2|I_t] \neq E[r_t^2|I_t]$ then the model that performs best under MAE may not be the actual best model. In fact, even the *true* model may be rejected in favour of some other model if MAE is used.

When r_t^2 is close to symmetrically distributed we will find $Median[r_t^2|I_t] \approx E[r_t^2|I_t]$ and thus MAE will “work” and will provide approximately correct rankings of models. In Figure 1 below I plot $Median[r_t^2|I_t]$ as a function of x_t , holding σ_t^2 equal to its unconditional average, using the same parameter values as the author: $(b, \omega, \alpha, \beta, c, p) = (5^{-1/2}, 0.02, 0.1, 0.8, 2, 0.1)$. This figure shows that $Median[r_t^2|I_t] \neq E[r_t^2|I_t]$, but that it is “close”, in the sense that the error is small relative to the changes in $E[r_t^2|I_t]$ coming from changes in x_t . This explains why MAE gives approximately the right rankings, for small T . However, since strictly $Median[r_t^2|I_t] \neq E[r_t^2|I_t]$, MAE will *not* provide correct rankings as $T \rightarrow \infty$, since the deviation of MAE from MSE will be detected in large samples, and this is also what is found by the author in his simulations.

Thus, in general, MAE is *not* a good loss function for comparing forecasts of variability. This holds true regardless of whether high or low frequency data is used.

On sample kurtosis as a measure of goodness of fit: the author considers four methods to compare forecasts, the third of these is given in equation 12:

$$K_m = \frac{1}{T} \sum_{t=1}^T \left(\frac{r_t - \hat{r}_{mt}}{\hat{\sigma}_{mt}} \right)^4$$

where \hat{r}_{mt} and $\hat{\sigma}_{mt}$ are the mean and standard deviation forecasts from model m . This loss function turns out to perform terribly in the author’s empirical section. This can be explained theoretically: Ignore, for simplicity, the model for the conditional mean, and think about the volatility forecast that would perform best according to this metric. It is then clear that the optimal model under this metric is the one where $\hat{\sigma}_{mt} \rightarrow \infty \forall t$, which makes the average loss arbitrarily small. So, in the spirit of Patton (2006), the volatility model that will appear best according to this metric will simply be the one with the highest volatility. Moreover, any given sequence of volatility forecasts could be made to look better according to this metric by simply multiplying them by some large positive number. Thus this metric is terrible both theoretically and empirically, and the author should just cut it from the paper altogether.

An alternative metric that I expect would have good properties is the so-called QLIKE loss function:

$$L(r_t, \hat{r}_{mt}, \hat{\sigma}_{mt}^2) = \frac{1}{T} \sum_{t=1}^T \left\{ \log \hat{\sigma}_{mt}^2 + \frac{(r_t - \hat{r}_{mt})^2}{\hat{\sigma}_{mt}^2} \right\}$$

Using quasi-likelihood style arguments this loss function can be shown to work well even when the data are non-Normally distributed.

On the need for a model to think about volatility: In the final paragraph on page 4, and elsewhere, the author suggests discussion of the conditional volatility of a variable is necessarily a *model-based* discussion. I disagree. Using the author’s notation from page 4, I do agree that the conditional variance of e_t is model-based, as e_t is the residual from some parametric model. But the conditional variance of the dependent variable, r_t , given some information set, $V[r_t|\mathbf{x}_t, \mathbf{y}_t]$ or $V[r_t|\mathbf{x}_t]$, can be defined and discussed without writing down a model; these conditional variances depend on the data generating process (DGP) and not a model. Of course obtaining an *estimate* of the conditional variance usually requires a model, but a direct estimate is not always needed: Hansen and Lunde (2006b) and Patton (2006) present methods for evaluating and comparing volatility models that avoid having to directly model the conditional variance.

On the choice of data for the empirical illustration: the author uses weekly returns on the Norwegian kroner/Euro exchange rate, over the period October 2005 to January 2008. The author motivates this choice of data as it is “well suited to illustrate the methodological and practical issues that can arise in the forecast evaluation of explanatory models of financial variability”. In contrast, I found this a very strange choice: this series does not exhibit volatility clustering! The authors results at the top of page 21 reveal that this series does not exhibit ARCH (the p-value for the ARCH(1) test is 0.86), and thus using this series to examine models, and methods for comparing models, of financial volatility is difficult to understand. How will the author’s conclusions change when series with substantial volatility clustering are considered?

On the “methodological and practical issues” raised by this paper: the author writes in several places (middle of page 2, top of page 5, top of page 18) that there are complications when thinking about evaluating discrete-time models of volatility using volatility proxies constructed using high-frequency data, or when evaluating forecasts of volatility/variability generally. I agree with some parts of this discussion, but even by the end of the paper I was at a loss to name the issues that the author is so concerned about.

1. One issue I do understand is the possibility that high-frequency estimates of volatility are contaminated with market microstructure noise, and so are not valid volatility proxies. This is well-known, with several solutions already proposed in the literature, and the author cites several papers that consider this problem.
2. Another issue the author raises is the problem of a time-varying conditional mean, and this can indeed cause problems with some volatility proxies, if they are interpreted as proxies for the conditional variance only. The author has a neat solution to this for the case of squared returns: interpret the squared return as an unbiased proxy for the *sum of* the conditional variance (σ_t^2) and the square of the conditional mean (μ_t^2), the “variability” of the variable, in the author’s terminology. Whilst this is neat, I think it is not ideal: since the model for the variability is (generally) comprised of two models (one for the mean and another for the variance) a test of these two models separately would be more informative about model mis-specification than a test on the combination.

3. At the start of Section 3 (page 7) the author writes that the “evaluation of discrete models of volatility against high-frequency estimates of continuous-time analogs can be mis-leading”. The only possible sources of problems that I could see are the two points above, but I get the feeling that the author is thinking about something more profound (see perhaps the bottom half of page 6, which did not seem to me to be relevant to this discussion).

