Volatility Forecast Evaluation and Comparison

Andrew Patton

Department of Economics
Duke University

OMI-SoFiE Summer School, 2013
Papers to be covered


Evaluating volatility forecasts

- The standard definition of an **optimal forecast** for loss function $L$ is:

  $$\hat{Y}_{t+h, t}^* \equiv \arg\min_{\hat{y} \in \mathcal{Y}} E \left[ L \left( Y_{t+h}, \hat{y} \right) | \mathcal{F}_t \right]$$

- Traditional evaluation and comparison tests are usually based on **Mincer-Zarnowitz** type regressions:

  $$Y_{t+h} = \beta_0 + \beta_1 \hat{Y}_{t+h, t} + u_{t+h}$$

  $H_0 : \beta_0 = 0 \cap \beta_1 = 1$

- And **Diebold-Mariano-West** tests

  $$d_{t+h} \equiv L \left( Y_{t+h}, \hat{Y}_{t+h, t}^a \right) - L \left( Y_{t+h}, \hat{Y}_{t+h, t}^b \right)$$

  $$H_0 : E \left[ d_t \right] = 0$$

- Both of these rely on the **observability** of $Y_{t+h}$, which does not hold when the object of interest is $\sigma^2_{t+1} \equiv V \left[ r_{t+1} | \mathcal{F}_t \right]$. 
Latent variables in macroeconomics: GDP growth

Related to Clark and McCracken, 2009, JBES.
Using “volatility proxies” in MZ and DMW tests

- The standard method of overcoming the latent nature of \( \sigma^2_{t+1} \) (or \( IV_{t+1} \) or \( E_t[IV_{t+1}] \) in modern work) is to find some volatility proxy.

- A common example is the squared daily return, which is conditionally unbiased for \( \sigma^2_{t+1} \) if the conditional mean is zero:

\[
\hat{\sigma}^2_{t+1} = r_{t+1}^2 \\
E_t \left[ \hat{\sigma}^2_{t+1} \right] = E_t \left[ r_{t+1}^2 \right] = V_t \left[ r_{t+1} \right] \equiv \sigma^2_{t+1}
\]

- “Realized volatility” may also be conditionally unbiased for the expected integrated variance or the conditional variance (under some conditions)

\[
\hat{\sigma}^2_{t+1} = RV_{t+1}^{(m)} = \sum_{j=1}^{m} r_{jt}^2 \\
E_t \left[ \hat{\sigma}^2_{t+1} \right] = \sigma^2_{t+1} ? \\
E_t \left[ \hat{\sigma}^2_{t+1} \right] = E_t \left[ IV_{t+1} \right] ?
\]
Using “volatility proxies” in MZ and DMW tests

- Many previous papers ignored the estimation error in the volatility proxy
  - In some applications this is OK
  - In other applications, ignoring the estimation error can lead to undesirable outcomes.

- Hansen and Lunde (2006) and Patton (2011) derive results that show using a conditionally unbiased, but noisy, proxy

\[ E_{t-1} \left[ \hat{\sigma}_t^2 \right] = \sigma_t^2 \]

but
\[ E_{t-1} \left[ \left( \hat{\sigma}_t^2 - \sigma_t^2 \right)^2 \right] > 0 \]

may lead to problems, and suggest methods that are robust to noise in the proxy.
Hansen and Lunde (2006)

- HL set out to obtain conditions which ensure that the ranking of volatility forecasts is *robust* to noise:

\[
E \left[ L \left( \sigma^2_t, h_{1t} \right) \right] \geq E \left[ L \left( \sigma^2_t, h_{2t} \right) \right] \iff E \left[ L \left( \hat{\sigma}^2_t, h_{1t} \right) \right] \geq E \left[ L \left( \hat{\sigma}^2_t, h_{2t} \right) \right]
\]

- where

\[
\sigma^2_t \quad \text{is the object of interest (latent)} \quad \in \mathcal{F}_{t-1}
\]

\[
\hat{\sigma}^2_t \quad \text{is the proxy for } \sigma^2_t \quad \text{(observable)} \quad \in \mathcal{F}_{t}
\]

\[
h_{it} \quad \text{is the volatility forecast from model } i \quad \in \mathcal{F}_{t-1}
\]
Three “pre-orderings” of volatility models

- Hansen and Lunde distinguish three “pre-orderings” (rankings) of volatility models:

  **True** pre-ordering: \[ E \left[ L \left( \sigma_t^2, \cdot \right) \right] \]

  **Approximate** pre-ordering: \[ E \left[ L \left( \hat{\sigma}_t^2, \cdot \right) \right] \]

  **Empirical** pre-ordering: \[ \frac{1}{T} \sum_{t=1}^{T} L \left( \hat{\sigma}_t^2, \cdot \right) \]

- Only the “empirical pre-ordering” is directly observable. Under basic assumptions the empirical pre-ordering limits to the approximate pre-ordering.

- The main question is: under what conditions are the “true” and “approximate” pre-orderings equivalent?
Key assumptions in the analysis of Hansen and Lunde I

- Assumption 1(i): \( L \left( \sigma^2_t, h \right) \) and \( L \left( \hat{\sigma}^2_t, h \right) \) are “integrable” \( \forall t \)

\[ \Rightarrow E \left[ \left\| L \left( \sigma^2_t, h \right) \right\| \right] < \infty \text{ and } E \left[ \left\| L \left( \hat{\sigma}^2_t, h \right) \right\| \right] < \infty \]

- Assumption 1(ii+iii):

\[
\lim_{T \to \infty} T^{-1} \sum_{t=1}^{T} E \left[ L \left( \sigma^2_t, h_t \right) \right], \quad \lim_{T \to \infty} T^{-1} \sum_{t=1}^{T} E \left[ L \left( \hat{\sigma}^2_t, h_t \right) \right],
\]

and \( \lim_{T \to \infty} T^{-1} \sum_{t=1}^{T} L \left( \hat{\sigma}^2_t, h_t \right) \) exist and are finite
Key assumptions in the analysis of Hansen and Lunde II

- Assumption 2(i): Define $\eta_t \equiv \hat{\sigma}^2_t - \sigma^2_t$. Let $\mathcal{F}_t$ be some filtration such that for all $h_{it}$ we have $(\sigma^2_t, h_{it}) \in \mathcal{F}_{t-1}$

- Assumption 2(ii): (a) $\frac{\partial L(\sigma^2, h)}{\partial \sigma^2}$ exists and does not depend on $h$, OR

- Assumption 2(ii): (b) $\frac{\partial^2 L(\sigma^2, h)}{\partial (\sigma^2)^2}$ exists and does not depend on $h$, AND $\{\eta_t, \mathcal{F}_t\}$ is a martingale difference sequence $\Rightarrow E[\eta_t | \mathcal{F}_{t-1}] = 0$

- Note: Assumption 2(ii)(a) is unlikely to hold in practice – 2(ii)(b) is a more relevant assumption.
The main result in Hansen and Lunde is Theorem 2: under assumptions 1 and 2 the “true” and “approximate” pre-orderings are equivalent.

**Proof:** take a second-order mean-value expansion of $L \left( \hat{\sigma}_t^2, h_t \right)$:

$$ L \left( \hat{\sigma}_t^2, h_t \right) = L \left( \sigma_t^2, h_t \right) + \frac{\partial L \left( \sigma_t^2, h_t \right)}{\partial \sigma^2} (\hat{\sigma}_t^2 - \sigma_t^2) $$

$$ + \frac{1}{2} \frac{\partial^2 L \left( \sigma_t^2, h_t \right)}{\partial (\sigma^2)^2} (\hat{\sigma}_t^2 - \sigma_t^2)^2 $$

where $\hat{\sigma}_t^2 = \alpha \hat{\sigma}_t^2 + (1 - \alpha) \sigma_t^2$, for some $\alpha \in [0, 1]$, and

$\partial^2 L \left( \sigma_t^2, h_t \right) / \partial (\sigma^2)^2 \equiv \lambda \left( \ddot{\sigma}_t^2 \right)$, since $\partial^2 L / \partial (\sigma^2)^2$ does not depend on $h$ by assumption 2(ii) (b). Then:
Take unconditional expectation of \( L(\hat{\sigma}_t^2, h_t) \):

\[
E_{t-1} \left[ L(\hat{\sigma}_t^2, h_t) \right] = E_{t-1} \left[ L(\sigma_t^2, h_t) \right] + \frac{\partial L(\sigma_t^2, h_t)}{\partial \sigma^2} E_{t-1} \left[ \hat{\sigma}_t^2 - \sigma_t^2 \right] \\
+ \frac{1}{2} E_{t-1} \left[ \lambda (\ddot{\sigma}_t) (\hat{\sigma}_t^2 - \sigma_t^2)^2 \right] \\
= E_{t-1} \left[ L(\sigma_t^2, h_t) \right] + \frac{1}{2} E_{t-1} \left[ \lambda (\ddot{\sigma}_t) (\hat{\sigma}_t^2 - \sigma_t^2)^2 \right]
\]

so

\[
E \left[ L(\hat{\sigma}_t^2, h_t) \right] = E \left[ L(\sigma_t^2, h_t) \right] + \frac{1}{2} E \left[ \lambda (\ddot{\sigma}_t) (\hat{\sigma}_t^2 - \sigma_t^2)^2 \right]
\]

and

\[
E \left[ L(\hat{\sigma}_t^2, h_{1t}) \right] - E \left[ L(\hat{\sigma}_t^2, h_{2t}) \right] = E \left[ L(\sigma_t^2, h_{1t}) \right] - E \left[ L(\sigma_t^2, h_{2t}) \right]
\]

Thus we’ve shown the equivalence of the “approximate” and the “true” pre-orderings (rankings)
Application of the main result

- Hansen and Lunde verify that the MSE loss function

\[
L \left( \sigma_t^2, h_t \right) = \left( \sigma_t^2 - h_t \right)^2 \Rightarrow \frac{\partial^2 L \left( \sigma^2, h \right)}{\partial (\sigma^2)^2} = 2
\]

satisfies their assumption 2 if used in conjunction with a conditionally unbiased proxy:

\[
E_{t-1} \left[ \hat{\sigma}_t^2 \right] = \sigma_t^2
\]

- The MSE-log loss function does not:

\[
L \left( \sigma_t^2, h_t \right) = \left( \log \sigma_t^2 - \log h_t \right)^2 \Rightarrow \frac{\partial^2 L \left( \sigma^2, h \right)}{\partial (\sigma^2)^2} = 2 \frac{1 - \log \sigma_t^2 + \log h_t}{\sigma_t^4}
\]

- Thus ranking volatility forecasts using MSE and the daily squared returns is equivalent (asymptotically) to ranking them using the true conditional variance. Ranking by MSE on log variances is not.
Regression-based evaluation

- Hansen and Lunde also consider ranking forecasts by the $R^2$ of a Mincer-Zarnowitz regression of a transformation, $\varphi$, of the proxy on the same transformation of the forecast:

$$\varphi \left( \hat{\sigma}^2_t \right) = \beta_0 + \beta_1 \varphi (h_t) + u_t$$

- Common choices for $\varphi$ are $\varphi (x) = x$, $\varphi (x) = \log x$, and $\varphi (x) = \sqrt{x}$.

- The population $R^2$ from this regression is

$$R^2 = \frac{\text{Cov} \left[ \varphi \left( \hat{\sigma}^2_t \right), \varphi (h_t) \right]^2}{V \left[ \varphi (h_t) \right] V \left[ \varphi \left( \hat{\sigma}^2_t \right) \right]}$$

- HL show that rankings using this measure are robust to noise in $\hat{\sigma}^2_t$ if the following condition is satisfied:

$$E_{t-1} \left[ \left( \hat{\sigma}^2_t - \sigma^2_t \right)^j \right] \cdot \varphi(j) \left( \sigma^2_t \right) = c_j \text{ for } j = 1, 2, \ldots \forall t$$
Regression-based evaluation, cont’d

Consider the condition below for a few special cases

\[ E_{t-1} \left[ \eta_t^j \right] \cdot \varphi^{(j)} \left( \sigma_t^2 \right) = c_j \text{ for } j = 1, 2, \ldots \ \forall t \]

1. If \( \varphi \) is **affine**, then only requires that \( E_{t-1} \left[ \hat{\sigma}_t^2 - \sigma_t^2 \right] = c \) (possibly different from zero)

2. If \( \varphi \) is **logarithmic**, then need

\[ E_{t-1} \left[ \eta_t^2 \right] \cdot \frac{1}{\sigma_t^4} = c_1, \text{ OK if } E_{t-1} \left[ \eta_t \right] = 0 \]

\[ E_{t-1} \left[ \eta_t^2 \right] \cdot \frac{1}{\sigma_t^4} = c_2, \text{ so need } V_{t-1} \left[ \hat{\sigma}_t^2 \right] \propto \sigma_t^4, \text{ OK?} \]

In general, need \( E_{t-1} \left[ \eta_t^k \right] \propto \sigma_t^{2k} \ \forall k \)

Thus, the popular regression using logs is not likely to yield a robust ranking.

3. If \( \varphi \) is the **square-root**, then need \( E_{t-1} \left[ \eta_t^k \right] \propto \sigma_t^{2k-1} \ \forall k \) which is also unlikely to hold.
Empirical application

- Hansen and Lunde consider an application to IBM returns over the period Jan 1995 to Feb 2002. They use the first 1250 observations for estimation and the remaining 545 for out-of-sample forecast evaluation.

- They consider 8 variations of ARCH models:
  - ARCH, GARCH, EGARCH, APARCH, T-GARCH, FIGARCH, FIAPARCH.

- They consider 3 volatility proxies:
  - $\tilde{\sigma}^2_{[sc.RV]t} = \left( \frac{1}{T} \sum_{s=1}^{T} \frac{r_s^2}{RV_s} \right) RV_t$
  - $\tilde{\sigma}^2_{[RV+on]t} = RV_t + \left( \log P_t^{open} - \log P_{t-1}^{close} \right)^2$
  - $\tilde{\sigma}^2_{[sq.ret]t} = \left( \log P_t^{close} - \log P_{t-1}^{close} \right)^2$

- HL note that, under some assumptions:
  $$V \left[ \tilde{\sigma}^2_{[sc.RV]t} - \sigma_t^2 \right] \leq V \left[ \tilde{\sigma}^2_{[RV+on]t} - \sigma_t^2 \right] \leq V \left[ \tilde{\sigma}^2_{[sq.ret]t} - \sigma_t^2 \right]$$
# Empirical application - Table 1

Mean squared errors loss for GARCH models of IBM stock returns

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE (level)</th>
<th>MSE (log)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\text{MSE}_{\text{sc.RV}}$</td>
<td>$\text{MSE}_{\text{RV} + \text{on}}$</td>
<td>$\text{MSE}_{\text{sq.ret}}$</td>
</tr>
<tr>
<td>ARCH(1)</td>
<td>39.924 (7.159)</td>
<td>168.30 (126.67)</td>
<td>305.06 (150.21)</td>
</tr>
<tr>
<td>GARCH(1,1)</td>
<td>30.722 (5.820)</td>
<td>159.77 (120.33)</td>
<td>297.01 (142.98)</td>
</tr>
<tr>
<td>EGARCH(1,1)</td>
<td>25.434 (5.389)</td>
<td>155.25 (120.01)</td>
<td>289.37 (142.64)</td>
</tr>
<tr>
<td>A-PARCH(1,1)</td>
<td>23.358 (5.147)</td>
<td><strong>153.66</strong> (120.01)</td>
<td><strong>286.52</strong> (142.54)</td>
</tr>
<tr>
<td>THR-GARCH(1,1)</td>
<td>24.711 (5.111)</td>
<td>155.00 (119.39)</td>
<td>288.49 (141.86)</td>
</tr>
<tr>
<td>FIGARCH(0,0)</td>
<td>31.460 (6.376)</td>
<td>161.89 (123.15)</td>
<td>299.03 (146.01)</td>
</tr>
<tr>
<td>FIGARCH(1,1)</td>
<td>32.057 (6.350)</td>
<td>161.89 (121.08)</td>
<td>299.51 (143.76)</td>
</tr>
<tr>
<td>FIAPARCH(1,1)</td>
<td>24.334 (4.854)</td>
<td>155.53 (119.69)</td>
<td>288.11 (141.96)</td>
</tr>
</tbody>
</table>
### Empirical application - Table 2

<table>
<thead>
<tr>
<th>Model</th>
<th>Level regression</th>
<th>Log-regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$R_{sc.RV}^2$</td>
<td>$R_{RV+on}^2$</td>
</tr>
<tr>
<td>ARCH(1)</td>
<td>0.067</td>
<td>0.010</td>
</tr>
<tr>
<td>GARCH(1,1)</td>
<td>19.907</td>
<td>4.328</td>
</tr>
<tr>
<td>EGARCH(1,1)</td>
<td>30.575</td>
<td>6.365</td>
</tr>
<tr>
<td>A-PARCH(1,1)</td>
<td><strong>36.052</strong></td>
<td><strong>7.329</strong></td>
</tr>
<tr>
<td>THR-GARCH(1,1)</td>
<td>32.425</td>
<td>6.665</td>
</tr>
<tr>
<td>FIGARCH(0,0)</td>
<td>16.867</td>
<td>3.112</td>
</tr>
<tr>
<td>FIGARCH(1,1)</td>
<td>16.338</td>
<td>3.420</td>
</tr>
<tr>
<td>FIAPARCH(1,1)</td>
<td>33.784</td>
<td>6.737</td>
</tr>
</tbody>
</table>
Patton (2011)

- In this paper I extend the work of Hansen and Lunde (2006) in a few useful directions:

1. I derive **analytical results** on the **distortions** in rankings of volatility forecast that occur using some common loss functions and volatility proxies.

2. I provide a necessary and **sufficient** condition on the loss function for it to yield rankings of volatility forecasts that are robust to noise in the volatility proxy.

3. I provide some guidance on the **choice of loss function** for volatility forecast comparison, through:
   - statistical considerations (required number of finite moments)
   - economic considerations (homogeneous loss functions are the only sensible choice in economics)
Framework

- As in earlier forecast studies, I start from the definition of an optimal forecast:

\[ \hat{Y}_{t+h,t} \equiv \arg \min_{\hat{y} \in \mathcal{Y}} E \left[ L \left( Y_{t+h}, \hat{y} \right) | \mathcal{F}_t \right] \]

- If we convert this to volatility forecasting and assume that no proxy is needed we get:

\[ h_t^* \equiv \arg \min_{h \in \mathcal{H}} E \left[ L \left( \sigma_t^2, h \right) | \mathcal{F}_{t-1} \right] = \sigma_t^2, \]

since \( \sigma_t^2 \in \mathcal{F}_{t-1} \) and \( L \left( \sigma_t^2, \sigma_t^2 \right) = 0 \ \forall \ L \)

- But if we are forced to use a volatility proxy we get:

\[ h_t^* \equiv \arg \min_{h \in \mathcal{H}} E \left[ L \left( \hat{\sigma}_t^2, h \right) | \mathcal{F}_{t-1} \right] \]

\[ = \sigma_t^2 \text{ ?? for all/some } L? \]
It is immediate that the optimal volatility forecast is $\sigma_t^2$ when $\sigma_t^2$ is observable (at time $t - 1$).

But if we define an optimal volatility forecast for loss function $L$ as:

$$h_t^* \equiv \arg \min_{h \in \mathcal{H}} E \left[ L \left( \hat{\sigma}_t^2, h \right) | \mathcal{F}_{t-1} \right]$$

then it is not clear that for any choice of $L$ we have $h_t^* = \sigma_t^2$.

Thus, the presence of noise in the proxy makes the choice of loss function very important.

The problem we face here is unusual: we know what the optimal forecast is (we want $h_t^* = \sigma_t^2$), and we want to find loss functions that ensure this.

It will turn out that by making sure that $L$ generates $h_t^* = \sigma_t^2$, we also satisfy Hansen and Lunde’s condition that $\frac{\partial^3 L}{\partial (\hat{\sigma}_t^2)^2 \partial h} = 0$. 
Defining a “robust” loss function

A loss function is “robust” if the ranking of any two (possibly misspecified) forecasts, $h_{1t}$ and $h_{2t}$, by expected loss is the same whether the ranking is done using the true conditional variance, $\sigma^2_t$, or some conditionally unbiased proxy, $\hat{\sigma}^2_t$.

That is, if:

$$E \left[ L \left( \sigma^2_t, h_{1t} \right) \right] \cong E \left[ L \left( \sigma^2_t, h_{2t} \right) \right] \iff E \left[ L \left( \hat{\sigma}^2_t, h_{1t} \right) \right] \cong E \left[ L \left( \hat{\sigma}^2_t, h_{2t} \right) \right]$$

for any $\hat{\sigma}^2_t$ s.t. $E_{t-1} \left[ \hat{\sigma}^2_t \right] = \sigma^2_t$
Characterising the problems with some loss functions

- A necessary condition for a loss function to be robust is that it generates an optimal forecast that is equal to the true conditional variance, i.e. $h^*_t = \sigma^2_t$.

- Thus a measure of the extent of the problem (if any) with a loss function is to derive the optimal forecast under it and check how far it is from $\sigma^2_t$.

- I do this for a collection of 9 loss functions, and three volatility proxies:

  1. $\hat{\sigma}^2_t = r^2_t$

  2. $\hat{\sigma}^2_t = RV_{t+1}^{(m)} \equiv \sum_{j=1}^{m} r_{jt}^2$

  3. $\hat{\sigma}^2_t = RG^*_t \equiv \frac{1}{2 \sqrt{\log 2}} \left( \max_{\tau} \log P_{\tau} - \min_{\tau} \log P_{\tau} \right)$, for $t - 1 < \tau \leq t$
Loss functions

\[
MSE : \quad L(\hat{\sigma}^2, h) = (\hat{\sigma}^2 - h)^2
\]

\[
QLIKE : \quad L(\hat{\sigma}^2, h) = \frac{\hat{\sigma}^2}{h} - \log \frac{\hat{\sigma}^2}{h} - 1
\]

\[
MSE-LOG : \quad L(\hat{\sigma}^2, h) = \left( \log \hat{\sigma}^2 - \log h \right)^2
\]

\[
MSE-SD : \quad L(\hat{\sigma}^2, h) = \left( \hat{\sigma} - \sqrt{h} \right)^2
\]

\[
"HMSE" \text{ aka } MSE-prop : \quad L(\hat{\sigma}^2, h) = \left( \hat{\sigma}^2 / h - 1 \right)^2
\]

\[
MAE : \quad L(\hat{\sigma}^2, h) = \left| \hat{\sigma}^2 - h \right|
\]

\[
MAE-LOG : \quad L(\hat{\sigma}^2, h) = \left| \log \hat{\sigma}^2 - \log h \right|
\]

\[
MAE-SD : \quad L(\hat{\sigma}^2, h) = \left| \hat{\sigma} - \sqrt{h} \right|
\]

\[
MAE-prop : \quad L(\hat{\sigma}^2, h) = \left| \hat{\sigma}^2 / h - 1 \right|
\]
Consider the choice of the MSE loss function with squared returns as the volatility proxy. In that case we know:

\[
\begin{align*}
    h_t^* &\equiv \arg \min_{h \in \mathcal{H}} E_{t-1} \left[ \left( \hat{\sigma}_t^2 - h \right)^2 \right] \\
    &= E_{t-1} \left[ r_t^2 \right] \\
    &= \sigma_t^2 \cdot E_{t-1} \left[ \varepsilon_t^2 \right], \text{ since } r_t = \sigma_t \varepsilon_t \\
    &= \sigma_t^2, \text{ since } \varepsilon_t | \mathcal{F}_{t-1} \sim \text{Student’s } t(0, 1, \nu)
\end{align*}
\]

And so under this combination of loss function and proxy we find \textbf{no} bias.

This is consistent with the result from Hansen and Lunde (2006).
The QLIKE loss function using squared returns as the proxy

- Another popular choice is the so-called QLIKE loss function. In this case:

\[
    h_t^* \equiv \arg \min_{h \in \mathcal{H}} E_{t-1} \left[ \frac{\hat{\sigma}_t^2}{h} - \log \frac{\hat{\sigma}_t^2}{h} - 1 \right]
\]

FOC 0 = \[
E_{t-1} \left[ \frac{\partial}{\partial h} \left( \frac{\hat{\sigma}_t^2}{h^*) - \log \frac{\hat{\sigma}_t^2}{h^*} - 1 \right) \right]
\]

= \[
E_{t-1} \left[ -\frac{\hat{\sigma}_t^2}{h^*^2} + \frac{1}{h^*} \right]
\]

so \[
h_t^* = E_{t-1} \left[ \hat{\sigma}_t^2 \right]
\]

= \[
\sigma_t^2
\]

- So we again find that the optimal forecast is the true conditional variance. What about some other loss functions?
The MAE loss function using squared returns as the proxy

Consider the common choice of the MAE loss function with squared returns as the volatility proxy. In that case we know:

\[ h_t^* \equiv \arg \min_{h \in \mathcal{H}} E_{t-1} \left[ \left| \hat{\sigma}_t^2 - h \right| \right] \]

\[ = \text{Median}_{t-1} \left[ r_t^2 \right] \]

\[ = \sigma_t^2 \cdot \text{Median}_{t-1} \left[ \varepsilon_t^2 \right], \quad \text{since } r_t = \sigma_t \varepsilon_t \]

\[ = \begin{cases} 
\sigma_t^2 \cdot \frac{v-2}{v} \cdot \text{Median} \left[ F_{1,v} \right], & \text{if } r_t | \mathcal{F}_{t-1} \sim \text{Student’s } t \left( 0, \sigma_t^2, v \right) \\
\sigma_t^2 \cdot \text{Median} \left[ \chi^2_1 \right] \approx 0.45 \sigma_t^2, & \text{if } r_t | \mathcal{F}_{t-1} \sim N \left( 0, \sigma_t^2 \right) 
\end{cases} \]

And so under this combination of loss function and proxy we find that the optimal forecast is equal to 0.45 times the true conditional variance.

Thus we will generally be lead to selecting forecasts that are downward biased.
The MSE-SD loss function using sq rets as the proxy

- Another popular choice is the MSE-SD loss function with squared returns as the volatility proxy. In that case we have:

\[
\begin{align*}
    h_t^* & \equiv \arg \min_{h \in \mathcal{H}} E_{t-1} \left[ (\hat{\sigma}_t - \sqrt{h})^2 \right] \\
    &= E_{t-1} [||r_t||^2] \\
    &= \sigma_t^2 \cdot E_{t-1} [||\epsilon_t||^2] \\
    &= \begin{cases} 
        \frac{\nu-2}{\pi} \left( \Gamma \left( \frac{\nu-1}{2} \right) / \Gamma \left( \frac{\nu}{2} \right) \right)^2 \sigma_t^2, & \text{if } r_t | \mathcal{F}_{t-1} \sim \text{Student's t} \left( 0, \sigma_t^2, \nu \right), \\
        \frac{2}{\pi} \sigma_t^2 \approx 0.64 \sigma_t^2, & \text{if } r_t | \mathcal{F}_{t-1} \sim N \left( 0, \sigma_t^2 \right)
    \end{cases}
\end{align*}
\]

- And so under this combination of loss function and proxy we find that the optimal forecast is equal to 0.64 times the true conditional variance

- We will again generally be lead to selecting forecasts that are downward biased
### Summary of results across all loss functions

#### Distribution of daily returns

| Loss function | $r_t | \mathcal{F}_{t-1} \sim F_t (0, \sigma^2_t)$ | $r_t | \mathcal{F}_{t-1} \sim \text{Student's } t (v)$ |
|---------------|-----------------------------------------------|-----------------------------------------------|
| **MSE**       | $\sigma^2_t$                                  | $\sigma^2_t$                                  | $\sigma^2_t$                                  | $\sigma^2_t$                                  | $\sigma^2_t$                                  |
| **QLIKE**     | $\sigma^2_t$                                  | $\sigma^2_t$                                  | $\sigma^2_t$                                  | $\sigma^2_t$                                  | $\sigma^2_t$                                  |
| **MSE-LOG**   | $\exp \{ E_{t-1} [\log \epsilon_t^2] \} \sigma^2_t$ | $0.22\sigma^2_t$                              | $0.25\sigma^2_t$                              | $0.28\sigma^2_t$                              |
| **MSE-SD**    | $(E_{t-1} [||\epsilon_t||])^2 \sigma^2_t$     | $0.56\sigma^2_t$                              | $0.60\sigma^2_t$                              | $0.64\sigma^2_t$                              |
| **MSE-prop**  | Kurtosis$_{t-1} [r_t] \sigma^2_t$             | $6.00\sigma^2_t$                              | $4.00\sigma^2_t$                              | $3.00\sigma^2_t$                              |
| **MAE**       | Median$_{t-1} [r_t^2] \sigma^2_t$             | $0.34\sigma^2_t$                              | $0.39\sigma^2_t$                              | $0.45\sigma^2_t$                              |
| **MAE-LOG**   | Median$_{t-1} [r_t^2] \sigma^2_t$             | $0.34\sigma^2_t$                              | $0.39\sigma^2_t$                              | $0.45\sigma^2_t$                              |
| **MAE-SD**    | Median$_{t-1} [r_t^2] \sigma^2_t$             | $0.34\sigma^2_t$                              | $0.39\sigma^2_t$                              | $0.45\sigma^2_t$                              |
| **MAE-prop**  | n / a                                         | $2.73\sigma^2_t$                              | $2.55\sigma^2_t$                              | $2.36\sigma^2_t$                              |
| **MAE-prop**  | $\dagger$                                     |                                                |                                                |                                                |
Interpreting Patton’s Table 1

Table 1 reveals two important facts:

1. For two loss functions (MSE and QLIKE) the optimal forecast is the true conditional variance. Thus they satisfy at least this necessary condition for robustness (they also satisfy Hansen and Lunde’s sufficient condition).

2. The remaining seven loss functions all generate optimal forecasts that differ from the true conditional variance.

   - The bias is worse for larger kurtosis
   - The bias can be upwards (MSE-prop and MAE-prop) or downwards (the rest)

The fact that the bias can be upwards or downwards explains the conflicting findings various authors have found when looking for the best volatility forecast across a range of (non-robust) loss functions.
Using better volatility proxies

- I next consider the use of **realised volatility** and the (adjusted) **range**, to see how the bias is affected when the noise in the proxy is reduced.

- I do this using a very simple DGP:

\[
\begin{align*}
    r_t &= \sigma_t dW_t \\
    \sigma_\tau &= \sigma_t \forall \tau \in (t-1, t] \\
    r_{i,m,t} &= \int_{(i-1)/m}^{i/m} r_\tau d\tau = \sigma_t \int_{(i-1)/m}^{i/m} dW_\tau \\
    \{r_{i,m,t}\}_{i=1}^m &\sim iid N\left(0, \frac{\sigma_t^2}{m}\right)
\end{align*}
\]

- Although RV theory has been developed for more general DGPs, theory for the range is mostly based on work by Feller (1951) who considered on the above simple case.

- Recent work by Christensen and Podolskij (2006, JoE) provides asymptotic theory for the “realised range” but I do not consider this.
Distributional properties of RV and RG

- This simple DGP allows me to obtain analytically moments and quantiles of RV and RG:

\[ RV_t^{(m)} \equiv \sum_{i=1}^{m} r_{t,i}^2 = \frac{\sigma_t^2}{m} \sum_{i=1}^{m} \varepsilon_{t,i}^2 \]

so \( m\sigma_t^{-2} RV_t^{(m)} \sim \chi_m^2 \)

- Feller (1951) provided the density of the range, and Parkinson (1980) provided a formula for obtaining moments:

\[
\begin{align*}
  f(RG_t; \sigma_t) &= 8 \sum_{k=1}^{\infty} (-1)^{k-1} \frac{k^2}{\sigma_t} \phi \left( \frac{k \cdot RG_t}{\sigma_t} \right) \\
  E[RG_t^p] &= \frac{4}{\sqrt{\pi}} \Gamma \left( \frac{p+1}{2} \right) \left( 2^{p/2} - 2^{2-p/2} \right) \zeta(p-1) \sigma_t^p, \text{ for } p \geq 1
\end{align*}
\]

- where \( \phi \) is the standard normal pdf, \( \text{erfc}(x) \equiv 1 - \text{erf}(x) \), \( \text{erf}(x) \) is the ‘error function’:

\[ \text{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \]

\( \zeta \) is the Riemann zeta function
The MAE loss function using RV or RG as the proxy

- Consider the MAE loss function with RV as the volatility proxy. In that case we know:

\[
    h_t^* = \text{Median}_{t-1}[RV_t] = \frac{\sigma_t^2}{m} \text{Median}\left[\chi_m^2\right]
\]

\[
    \approx \left(1 - \frac{2}{3m} + \frac{1}{9m^2}\right)\sigma_t^2
\]

\[
    = \begin{cases} 
        0.44 \cdot \sigma_t^2 & \text{for } m = 1 \\
        0.95 \cdot \sigma_t^2 & \text{for } m = 13 \\
        0.99 \cdot \sigma_t^2 & \text{for } m = 78 
    \end{cases}
\]

- Using RG* as the proxy we obtain:

\[
    h_t^* = \text{Median}_{t-1}[RG_t^*]
\]

\[
    \approx \frac{2.2938}{4 \log 2} \sigma_t^2 \approx 0.83\sigma_t^2
\]

using numerical methods to invert the cdf of $RG_t^*$. 
- Thus using more accurate volatility proxies does, as expected, reduce the bias in this case.
Summary of results across all loss functions

Volatility proxy

<table>
<thead>
<tr>
<th>Loss function</th>
<th>Range</th>
<th>$m = 1$</th>
<th>$m = 13$</th>
<th>$m = 78$</th>
<th>$m \to \infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$MSE$</td>
<td>$\sigma_t^2$</td>
<td>$\sigma_t^2$</td>
<td>$\sigma_t^2$</td>
<td>$\sigma_t^2$</td>
<td>$\sigma_t^2$</td>
</tr>
<tr>
<td>$QLIKE$</td>
<td>$\sigma_t^2$</td>
<td>$\sigma_t^2$</td>
<td>$\sigma_t^2$</td>
<td>$\sigma_t^2$</td>
<td>$\sigma_t^2$</td>
</tr>
<tr>
<td>$MSE-LOG^+$</td>
<td>$0.85\sigma_t^2$</td>
<td>$0.28\sigma_t^2$</td>
<td>$0.91\sigma_t^2$</td>
<td>$0.98\sigma_t^2$</td>
<td>$\sigma_t^2$</td>
</tr>
<tr>
<td>$MSE-SD$</td>
<td>$0.92\sigma_t^2$</td>
<td>$0.56\sigma_t^2$</td>
<td>$0.96\sigma_t^2$</td>
<td>$0.99\sigma_t^2$</td>
<td>$\sigma_t^2$</td>
</tr>
<tr>
<td>$MSE-prop$</td>
<td>$1.41\sigma_t^2$</td>
<td>$3.00\sigma_t^2$</td>
<td>$1.15\sigma_t^2$</td>
<td>$1.03\sigma_t^2$</td>
<td>$\sigma_t^2$</td>
</tr>
<tr>
<td>$MAE$</td>
<td>$0.83\sigma_t^2$</td>
<td>$0.45\sigma_t^2$</td>
<td>$0.95\sigma_t^2$</td>
<td>$0.99\sigma_t^2$</td>
<td>$\sigma_t^2$</td>
</tr>
<tr>
<td>$MAE-LOG$</td>
<td>$0.83\sigma_t^2$</td>
<td>$0.45\sigma_t^2$</td>
<td>$0.95\sigma_t^2$</td>
<td>$0.99\sigma_t^2$</td>
<td>$\sigma_t^2$</td>
</tr>
<tr>
<td>$MAE-SD$</td>
<td>$0.83\sigma_t^2$</td>
<td>$0.45\sigma_t^2$</td>
<td>$0.95\sigma_t^2$</td>
<td>$0.99\sigma_t^2$</td>
<td>$\sigma_t^2$</td>
</tr>
<tr>
<td>$MAE-prop^+$</td>
<td>$1.19\sigma_t^2$</td>
<td>$2.36\sigma_t^2$</td>
<td>$1.10\sigma_t^2$</td>
<td>$1.02\sigma_t^2$</td>
<td>$\sigma_t^2$</td>
</tr>
</tbody>
</table>
A class of robust loss functions

- Both the MSE and QLIKE loss functions yielded the conditional variance as the optimal forecast.

- This lead me to the question: Is there a general class of such loss functions?

- Proposition 2 suggests a class of loss functions, related to the linear-exponentenional family of densities of Gourieroux, et al. (1984), and to Gourieroux, et al. (1987).
A class of robust loss functions - assumptions

- **A1:** \( E \left[ \hat{\sigma}_t^2 | \mathcal{F}_{t-1} \right] = \sigma_t^2 \)

- **A2:** \( \hat{\sigma}_t^2 | \mathcal{F}_{t-1} \sim F_t \in \tilde{F} \), the set of all absolutely continuous distribution functions (i.e., those that have a pdf) on \( \mathbb{R}_+ \).

- **A3:** \( L \) is twice continuously differentiable with respect to \( h \) and \( \hat{\sigma}^2 \), and has a unique minimum at \( \hat{\sigma}^2 = h \).

- **A4:** There exists some \( h_t^* \in \text{int} (\mathcal{H}) \) such that \( h_t^* = E_{t-1} \left[ \hat{\sigma}_t^2 \right] \), where \( \mathcal{H} \) is a compact subset of \( \mathbb{R}_{++} \).

- **A5:** \( L \) and \( F_t \) are such that: (a) \( E_{t-1} \left[ L \left( \hat{\sigma}_t^2, h \right) \right] < \infty \) for some \( h \in \mathcal{H} \); (b) \( E_{t-1} \left[ \partial L \left( \hat{\sigma}_t^2, \sigma_t^2 \right) / \partial h \right] < \infty \); and (c) \( \left| E_{t-1} \left[ \partial^2 L \left( \hat{\sigma}_t^2, \sigma_t^2 \right) / \partial h^2 \right] \right| < \infty \) for all \( t \).
A class of robust loss functions - Prop 2

Let assumptions A1 to A5 hold. Then a loss function $L$ is "robust" if and only if it takes the following form:

$$L\left(\hat{\sigma}^2, h\right) = \tilde{C}(h) + B\left(\hat{\sigma}^2\right) + C(h)\left(\hat{\sigma}^2 - h\right)$$

where $B$ and $C$ are twice continuously differentiable, $C$ is a strictly decreasing function on $H$, and $\tilde{C}(h) \equiv \int C(h) \, dh$.

Note that if we normalise $L(h, h) = 0$, then the class simplifies to:

$$L\left(\hat{\sigma}^2, h\right) = \tilde{C}(h) - \tilde{C}\left(\hat{\sigma}^2\right) + C(h)\left(\hat{\sigma}^2 - h\right)$$
A class of robust loss functions - Prop 2 proof

- I proved this proposition by showing the equivalence of the following three statements:
  - $S_1$: The loss function takes the form given the statement of the proposition;
  - $S_2$: The loss function is robust in the sense of Definition 1;
  - $S_3$: The optimal forecast under the loss function is the conditional variance.

- I show that $S_1 \Rightarrow S_2 \Rightarrow S_3 \Rightarrow S_1$.
  - $S_1 \Rightarrow S_2$ follows from Hansen and Lunde (2006): their assumption 2 is satisfied given the assumptions for the proposition.
  - $S_2 \Rightarrow S_3$ is simple to show using the fact that rankings using “robust” loss functions are the same whether $\hat{\sigma}_t^2$ or $\sigma_t^2$ is used.
  - Proving $S_3 \Rightarrow S_1$ was harder. It is related to the necessity of the “linear-exponential” family of densities for quasi-maximum likelihood estimation (though my case does not fit directly into that framework). See the paper for details.
Interesting sub-sets of the robust class of loss functions

- The general class of robust loss functions is quite broad: it depends on the univariate function $C$, which is only restricted to be twice continuously differentiable and strictly decreasing function on $H$.

- I was interested in extracting sub-sets of this class that might be easier to handle for applied users. In doing so, I stumbled onto Prop 3:

**Proposition 3:**

- (i) The “MSE” loss function is the **only** robust loss function that depends solely on the **forecast error**, $\hat{\sigma}^2 - h$.

- (ii) The “QLIKE” loss function is the **only** robust loss function that depends solely on the **standardised forecast error**, $\frac{\hat{\sigma}^2}{h}$.
Parametric sub-sets of the robust class of loss functions

- I next tried to extract a **parametric family** of robust loss functions, which nested the MSE and QLIKE cases.

- I did this by using the fact that both these loss functions have a FOC of the form:

\[
0 = E_{t-1} \left[ \frac{\partial L(\hat{\sigma}^2_t, h^*_t)}{\partial h} \right] = (h^*_t)^b \left( E_{t-1} \left[ \hat{\sigma}^2_t \right] - h^*_t \right), \hspace{1cm} b \in \mathbb{R}
\]

- where \( b = 0 \) for MSE loss and \( b = -2 \) for QLIKE loss.

- When we integrate this up, and normalise so that \( L(h, h) = 0 \) we obtain the following:

\[
L(\hat{\sigma}^2, h; b) = \begin{cases} 
\frac{1}{(b+1)(b+2)} (\hat{\sigma}^{2b+4} - h^{b+2}) - \frac{1}{b+1} h^{b+1} (\hat{\sigma}^2 - h), & \text{for } b \notin \{-1, -2\} \\
\frac{\hat{\sigma}^2}{h} \log \frac{\hat{\sigma}^2}{h}, & \text{for } b = -1 \\
\frac{\hat{\sigma}^2}{h} - \log \frac{\hat{\sigma}^2}{h} - 1, & \text{for } b = -2
\end{cases}
\]
Parametric sub-sets of the robust class of loss functions

Robust loss functions for various choices of $b$

Various robust loss functions

- $b=1$
- $b=0.5$
- $b=0$ (MSE)
- $b=-0.5$
- $b=-1$
- $b=-2$ (QLIKE)
- $b=-5$

$hhat (r2=2)$
Parametric sub-sets of the robust class of loss functions

Ratio of losses from neg errors to pos errors, for various choices of b

Ratio of loss from negative forecast errors to positive forecast errors

b = 1
b = 0.5
b = 0 (MSE)
b = -0.5
b = -1
b = -2 (QLIKE)
b = -5

Volatility Forecast Evaluation
The homogeneous sub-set of robust loss functions

- Finally I derived the sub-set of robust loss functions that is **homogeneous** of order $k$:
  \[
  L\left(a\hat{\sigma}^2, ah\right) = a^k L\left(\hat{\sigma}^2, h\right), \forall a > 0
  \]

- It turns out that the sub-set of homogeneous robust loss functions is **exactly** the parametric family I derived previously:
  \[
  L\left(\hat{\sigma}^2, h; b\right) = \begin{cases}
    \frac{1}{(b+1)(b+2)}(\hat{\sigma}^{2b+4} - h^{b+2}) - \frac{1}{b+1} h^{b+1} \left(\hat{\sigma}^2 - h\right), & \text{for } b \notin \{-1, -2\} \\
    h - \hat{\sigma}^2 + \hat{\sigma}^2 \log \frac{\hat{\sigma}^2}{h}, & \text{for } b = -1 \\
    \frac{\hat{\sigma}^2}{h} - \log \frac{\hat{\sigma}^2}{h} - 1, & \text{for } b = -2
  \end{cases}
  \]

- where the degree of homogeneity, $k = b + 2$. 
Why is homogeneity so important? I

- Homogeneity of the loss function is useful in economics because the choice of units is often arbitrary: decimal vs. percent returns, etc

- One would hope that a simple re-scaling of the data does not change any conclusions, but this is not the case. Consider the following example:
Why is homogeneity so important?

- $\sigma_t^2 = 1 \ \forall t, \ (h_{1t}, h_{2t}) = (\gamma_1, \gamma_2) \ \forall t, \ \text{and} \ \hat{\sigma}_t^2 \text{ is s.t. } E_{t-1} \left[ \hat{\sigma}_t^2 \right] = 1 \ \text{a.s.} \ \forall t.$

- One robust but non-homogeneous loss is given by choosing:

  \[ C'(h) = - \log (1 + h) \]

- Given this set-up, we have

  \[
  E \left[ L \left( a\hat{\sigma}_t^2, ah_{it} \right) \right] = \frac{1}{4} \left[ a\gamma_i (3a\gamma_i + 2) - 2 (1 + a\gamma_i)^2 \log (1 + a\gamma_i) \right] \\
  + a \left[ a\gamma_i - (1 + a\gamma_i) \log (1 + a\gamma_i) \right] (1 - \gamma_i) + \text{const}
  \]

Then define

\[
 d_t (\gamma_1, \gamma_2, a) \equiv L \left( a\hat{\sigma}_t^2, a\gamma_1 \right) - L \left( a\hat{\sigma}_t^2, a\gamma_2 \right)
\]

Then note that

\[
 E \left[ d_t (0.33, 1.5, 1) \right] = -0.0087 \Rightarrow h_1 \succ h_2
\]

but

\[
 E \left[ d_t (0.33, 1.5, 2) \right] = +0.0061 \Rightarrow h_1 \prec h_2
\]
Empirical application

- Daily and intra-daily data on IBM from January 1993 to December 2003, 2772 observations

- I consider two simple but widely-used volatility models:

  \[
  \text{Rolling window} \quad h_{1t} = \frac{1}{60} \sum_{j=1}^{60} r_{t-j}^2
  \]

  \[
  \text{RiskMetrics} \quad h_{2t} = \lambda h_{2t-1} + (1 - \lambda) r_{t-1}^2, \quad \lambda = 0.94
  \]

- First 272 observations are used for estimation, last 2500 observations are used for forecast comparison
Empirical application

Daily conditional variance forecasts for IBM, Jan 1994 – Dec 2003

Conditional variance forecasts

60-day rolling window
RiskMetrics
DMW forecast comparison tests

<table>
<thead>
<tr>
<th>Loss function</th>
<th>Daily squared return</th>
<th>65-min realised vol</th>
<th>15-min realised vol</th>
<th>5-min realised vol</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b = 1$</td>
<td>-1.58</td>
<td>-1.66</td>
<td>-1.30</td>
<td>-1.35</td>
</tr>
<tr>
<td>$b = 0$ (MSE)</td>
<td>-0.59</td>
<td>-0.80</td>
<td>-0.03</td>
<td>-0.13</td>
</tr>
<tr>
<td>$b = -1$</td>
<td>1.30</td>
<td>1.04</td>
<td>1.65</td>
<td>-1.55</td>
</tr>
<tr>
<td>$b = -2$ (QLIKE)</td>
<td>1.94</td>
<td>2.21*</td>
<td>2.73*</td>
<td>2.41*</td>
</tr>
<tr>
<td>$b = -5$</td>
<td>-0.17</td>
<td>0.25</td>
<td>1.63</td>
<td>0.65</td>
</tr>
</tbody>
</table>
DMW forecast comparison tests

Figure: *Rolling-window vs. RiskMetrics DMW t-statistics*
Summary

- Hansen and Lunde (2006) and Patton (2011) reveal that the latent nature of volatility (amongst other variables) can cause problems in standard tests for forecast evaluation and comparison.

  - Most tests were developed for the case that $Y_t \in \mathcal{F}_t$, and no such problems arise in that case.

- Hansen and Lunde (2006) provide a sufficient condition on the loss function for it to yield rankings that are robust to (mean zero) noise in the volatility proxy.

- Patton (2011) verifies that many commonly-used loss functions lead to severe biases when used with a noisy proxy.

  - More accurate volatility proxies were shown to alleviate these problems, but they do not completely remove them.

- Patton (2011) provides a necessary and sufficient condition for the loss function to be robust, ruling out some previously-used loss functions.