Are combination forecasts of S&P 500 volatility statistically superior?

Paper replication
Applied Time Series with R
Department of Economics
University of Zurich

Jan Wälty
Hügelistrasse 5
5040 Schöftland
Student Nr 20-704-227

1 Paper overview

The paper by Becker & Clements (2008) assesses models that are used to predict asset return volatility. Lopez (2001) argues that volatility predictions are crucial for various economic applications such as option pricing or exchange rate forecasting for national banks. The primary question they want to answer is whether implied volatility via volatility index (VIX) is superior to other model-based approaches in terms of predictive power. The former forecasts expected volatility based on option prices in the market, while the latter solely rely on statistical anaylsis. For this purpose, they employ various GARCH model specification and other stochastic models. These models are heavily in use as they are able to capture changing return variance over time (heteroscedasticity), which is a common feature of financial market returns (see Figure 1). It shows that volatility seems to be serially correlated, meaning that there is a certain short-term persistence that might be modelled by lags of the volatility itself. The authors find that VIX yields better results than classical volatility models such as basic GARCH specifications. However, it is inferior to models that incorporate realized volatility or form a combination of various models.

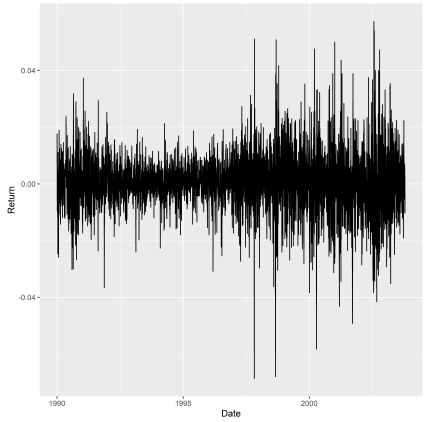

Figure 1: Daily returns S&P 500 (1990-2003)

2 Narrow replication

2.1 Methodology

The following part introduces the replication procedure as well as the implemented volatility models. In order to assess the performance of model-based predictions, the same rolling-window procedure as in Becker & Clements (2008) is applied. The data consists of daily returns of S&P 500 from 2 January 1990 to 17 October 2003 (corresponding to 3481 observations).
The procedure works as follows: 1000 data points are used to fit the models and predict the conditional volatility for the subsequent 22 trading days, which corresponds to the forecast horizon of the VIX index. These values are averaged and compared to the average of a proxy of realized volatility for the same horizon. We need a substitute for the true conditional volatility as the procedure in the paper relies on intraday trading data that is not available. Then, the procedure is repeated by shifting the sample one day ahead resulting in 2460 prediction intervals.

The corresponding MSE is computed as follows:

$$\text{MSE} = \frac{\Sigma_{t=1}^{2460}(\overline{RV}_{t+22} - f_t)^2}{N}, \tag{1}$$

where $\overline{RV}_{t+22}$ corresponds to the average realized volatility as measured by one of the proxies and f_t to the average of predicted volatilities.

2.2 Models

The replication comprises the modeling of S&P 500 conditional volatility. The subsequent models assume the following return specification:

$$Y_t = \sigma_t \epsilon_t \tag{2}$$

Y_t denotes the de-meaned asset return, σ_t the conditional standard deviation and ϵ_t a mean zero random process that is assumed to be N(0,1) in the implementation by Becker & Clements (2008). However, this assumption is usually not realistic.
The ARCH model describes the conditional volatility as a function of lagged returns. The standard ARCH(1) model can be formalized as follows:

$$\sigma_t^2 = \gamma + \alpha \cdot y_{t-1}^2 \tag{3}$$

This model was not considered in the original paper by Becker & Clements (2008) but added in the replication because it is one of the standard models in use. GARCH models are natural extensions of ARCH models. In addition to using lags of returns, they incorporate lags of the conditional variance itself. It is therefore able to generate volatility clustering (i.e, variability in conditional variance that is persistent over time). The standard GARCH(1,1) model can be written as follows:

$$\sigma_t^2 = \gamma + \alpha \cdot y_{t-1}^2 + \beta \cdot \sigma_{t-1}^2 \tag{4}$$

It is well known that negative returns tend to induce larger shocks in financial markets than positive returns. However, standard GARCH models are unable to incorporate such effects as only squared returns are considered. GJR-GARCH models are capable of capturing such effects by adding a specific term that serves as dummie for negative returns (Franses & Van Dijk, 1996). Formally, it can be modelled as

$$\sigma_t^2 = \gamma + (\alpha + \delta 1_{(y_{t-1}<0)}) \cdot y_{t-1}^2 + \beta \cdot \sigma_{t-1}^2 \tag{5}$$

Poon & Granger (2003) state that these models should generally perform better than classical Garch-models.
As already mentioned, Becker & Clements (2008) also incorporate combination forecast to test whether they perform better than single models. In their implementation, they obtain the corresponding model weights by the following multiple regression:

$$\overline{RV_{t+22}} = \alpha_0 + \alpha_1 f_t^1 + ... + f_t^n, \tag{6}$$

where f_t^i corresponds to the forecast of one of the models employed for the given forecast horizon. In contrast to the implementation in the original paper, only one combination forecast is considered (the one that contains all single forecasts as opposed to all possible subsets).

In order to compare the performance of the models in describing the underlying volatility dynamics, a benchmark (the realized or true conditional volatility) is needed. As established before, a proxy is needed as the construction by the authors cannot directly be implemented. Andersen & Bollerslev (1998) state that squared returns are often used as proxy for realized volatility. This estimator is unbiased and is linked to the definition in Equation 2:

$$E_{t-1}[Y_t^2] = E_{t-1}[\sigma_t^2 \cdot \epsilon_t^2] = \sigma_t^2 \tag{7}$$

However, the authors argue that this may yield noisy estimates due to the innovation term ϵ_t. Another straightforward proxy for realized volatility can be found by calculating the percentage deviation of daily high and low prices:

$$\hat{\sigma_t}^2 = \frac{P_{t,high} - P_{t,low}}{P_{t,low}} \tag{8}$$

It is reasonable to assume that larger deviations are positively correlated with conditional volatility.

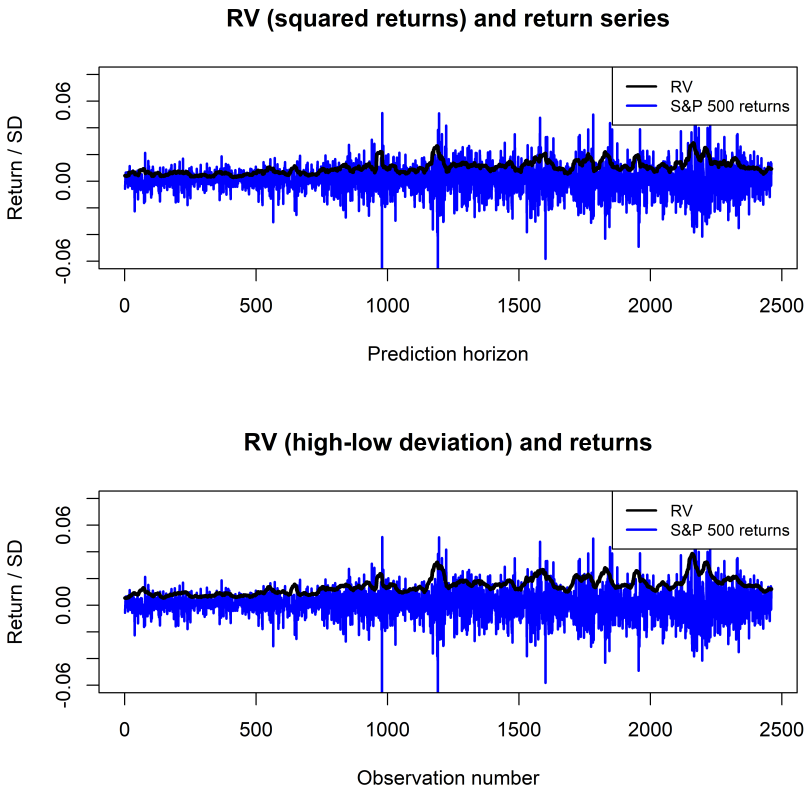

Figure 2: Volatility predictions for various models

Given that a proxy for realized volatility is used as benchmark, it is important to check the relationsship between returns and the proxy itself. Figure 2 shows the return time series and the corresponding RV series. By construction, the squared return volatility proxy reacts to hikes in returns. This is even more pronounced when using the high-low proxy. However, this is only a partially valid observation as the true conditional volatility remains unknown and cannot directly be observed by looking at the daily return plot. Intraday data is necessary for that purpose.

2.3 Results

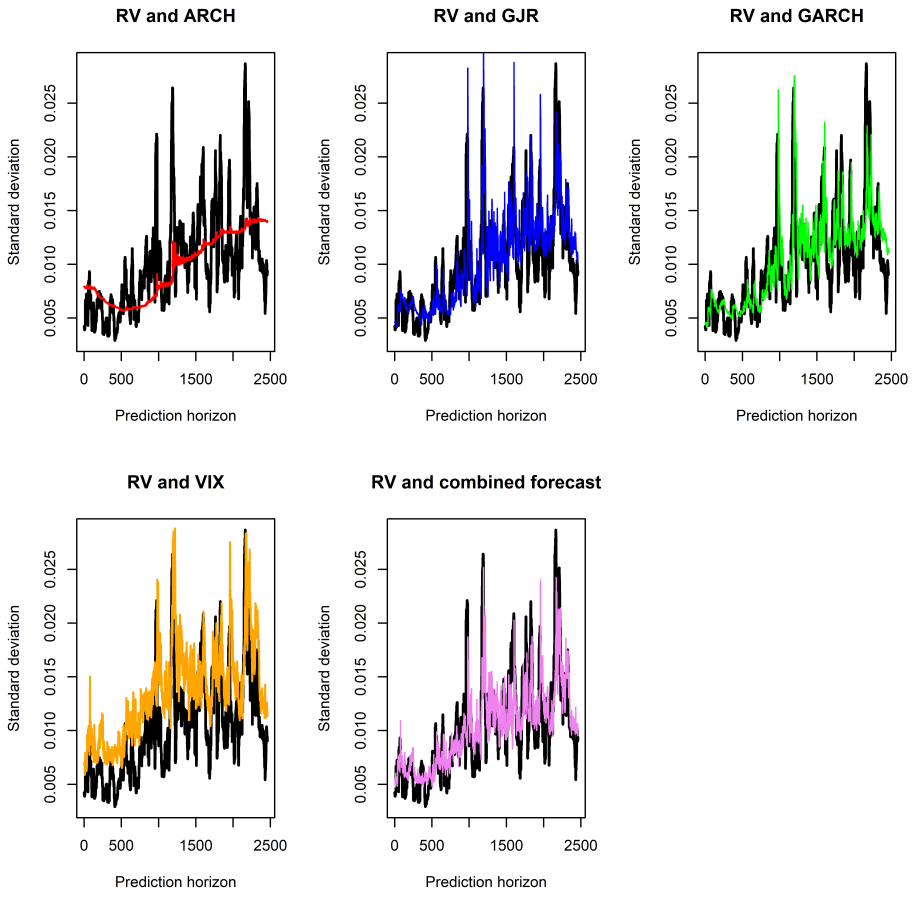

Figure 3: Comparison of models (squared return proxy)

Figure 3 displays the volatility model predictions and the realized volatility based on the squared return proxy. The upper-left plot underlines the characteristics of the basic ARCH model. In particular, it is not able to quickly adapt to volatility changes. Instead, it averages the values around a particular observation. On the other hand, models from the GARCH family quickly adapt to volatility changes and are able to model serial correlation between subsequent conditional volatilities (volatility clustering). Lastly, the VIX model tends to overestimate the effective volatility the most. This is a well-documented feature and is related to arbitrage considerations (Bandi & Perron, 2006).

Model	MSE
Combination	0.023
Garch(1,1)	0.029
GJR-Garch(1,1)	0.032
Arch(1,1)	0.039
VIX	0.052

Table 1: Squared returns as proxy

Model	MSE
VIX	0.037
Combination	0.078
Garch(1,1)	0.079
GJR-Garch(1,1)	0.080
Arch(1,1)	0.11

Table 2: Intraday low/high as proxy

Tables 1 and 2 show the MSE results for the rolling-window procedure when using the proposed proxies as benchmark. Apparently, adding volatility lags is beneficial given that both GARCH models yield a smaller MSE in both versions than the ARCH model. However, accounting for leverage effects in the GJR-GARCH model did not provide better results. The tables also underline that the result are prone to errors made by benchmark choice. This particularly applies to the VIX predictions that yield the best results when using the intraday low/high ratio as proxy and the worst result in the case of squared returns. It can also be seen that combination model yields better results than single (model-based) model specifications. This is in line with the findings of Becker & Clements (2008).

3 Reflection and further steps

The authors did not provide any replication material. Various models could not be implemented due to the lack of intraday data. Therefore, different assumptions had to be made such as the introduction of a proxy for realized volatility. Usually, it is not optimal to use squared returns due to their stochastic nature. We have seen that the model ranking based on the mean squared error is sensitive to benchmark selection. Naturally, using intraday data would be the preferred option as a larger sample size is able to capture more of the underlying dynamics than a samewhat artificial proxy. Therefore, the results obtained in this replication cannot directly be compared to the original paper, even though the ranking based on the predictive power of the single models is similar as in Becker & Clements (2008).
Further possible steps comprise the use of benchmarks other than squared daily returns. Additionally, the implemented models can be extended by adding more lags, which might be helpful to capture more of the underlying volatility dynamics. The GARCH model family offers a variety of other specifications that are able to capture other stylized characteristics of asset returns. Of course, testing the validity of the results using data from other markets might be helpful as market returns usually reflect the behavior of a specific regional economy. Therefore, we might not be able to generalize these results to other indices. Lastly, assuming a normal distribution for ϵ_t is usually assumed to be unrealistic. This is due to the fact that the distribution of high-frequency returns usually shows heavy-tails. Therefore, a distribution with such properties might be more adequate (e.g log-normal).

4 References

Andersen, T. G., & Bollerslev, T. (1998). *Answering the skeptics: Yes, standard volatility models do provide accurate forecasts.* International Economic Review, 885–905.

Bandi, F. M., & Perron, B. (2006). *Long memory and the relation between implied and realized volatility.* Journal of Financial Econometrics, 4(4), 636–670.

Becker, R., & Clements, A. E. (2008). *Are combination forecasts of S&P 500 volatility statistically superior?.* International Journal of Forecasting, 24(1), 122–133.

Franses, P. H., & Van Dijk, D. (1996). *Forecasting stock market volatility using (non-linear) Garch models.* Journal of Forecasting, 15(3), 229–235.

Lopez, J. A. (2001). *Evaluating the predictive accuracy of volatility models.* Journal of Forecasting, 20(2), 87–109.

Poon, S. H., & Granger, C. W. J. (2003). *Forecasting volatility in financial markets: A review.* Journal of Economic Literature, 41(2), 478–539.