Forecasting Realized Volatility using Recurrent Neural Networks: A Sequence-Based Approach

Maik Dierkes [1] Florian Sckade [1]

[1]Leibniz University Hannover - Institute of Banking and Finance

Abstract

- We analyze the predictive performance of different recurrent neural networks (RNNs) - simple RNNs, LSTMs, and GRUs - at forecasting realized volatility of the S&P 500.
- We use RNNs to **directly forecast monthly sequences of volatility** as vectors, analyzing the impact of hidden state sizes in single-layer RNNs and the influence of ensembling on results.
- **For short horizons, these sequential RNNs can outperform** traditional econometric methods without requiring very deep architectures.
- However, **their performance breaks down for longer horizons**, offering no overall advantage and performing **worse than the benchmark**.
- Additionally, we analyze the forecasting error distribution over hyperparameters and model ensemble size.

Motivation

- The most common class of time-series models to forecast realized volatility is the heterogeneous autoregressive model (HAR) by [4].
  - ⇒ A large body of literature improves on the standard HAR by incorporating, e.g., jumps or differentiating between positive and negative return volatility.
- With the field of machine learning (ML) advancing rapidly, so do the applications of its methods to financial econometrics. [2] [3]
- RV forecast quality is often evaluated over different horizons h (e.g. $h \in \{1, 5, 22\}$)
  - => Often done with individual models per h [5] [3] or with last forecast carried forward as input [2]
- ML models for sequence data (both NLP or numeric time-series) are often built to forecast a sequence of future values, instead of a single point.

Forecasting Problem Statement

Realized variance (RV) is a daily ex post volatility measurement, estimated from squared intraday returns [1]. The RV on day t is defined as the sum of squared intraday returns:

$$RV_t = \sum_{i=1}^{N} r_{i,t}^2,$$

where we consider N equally spaced intraday returns $r_{i,t}$. In most econometric applications, separate models are estimated for different forecast horizons:

$\{RV_{t-21}, ..., RV_{t-1}, RV_t\}$ → model #1 → $\widehat{RV}_{t+1}$

$\{RV_{t-21}, ..., RV_{t-1}, RV_t\}$ → model #2 → $\widehat{RV}_{t+5}$

$\{RV_{t-21}, ..., RV_{t-1}, RV_t\}$ → model #3 → $\widehat{RV}_{t+22}$

We test "one-size-fits-all" sequence-predicting RNNs for their forecast performance:

$\{RV_{t-21}, ..., RV_{t-1}, RV_t\}$ → RNN → $\{\widehat{RV}_{t+1}, \widehat{RV}_{t+2}, ..., \widehat{RV}_{t+h}\}$

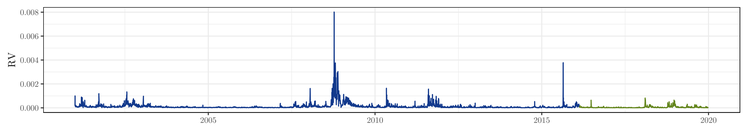
To analyze the out-of-sample performance of all employed methods, we use the MSE. We analyze our models by checking their performance per horizon, calculating the MSE for model i as

$$MSE_i = \frac{1}{T}\sum_{t=1}^{T}(RV_t - \widehat{RV}_{i,t})^2,$$

where $t \in (1, ..., T)$ is the number of forecast observations for the respective horizon h and $\widehat{RV}_i$ is model i's RV forecast.

Data and Benchmark

- We use the **Oxford Realized Library and focus on 5-minute sampled realized variance**. Besides the univariate case, we use S&P excess returns and the VIX as additional explanatory variables in a multivariate setting.
- RNNs are evaluated for a fixed test window and via **ensembled predictions**, while the HAR benchmark is evaluated via rolling estimations. Ensembles are constructed by averaging predictions of the 10 models with the lowest validation error out of 100 total models.
- 70%/10% training/validation data and 20% test data.

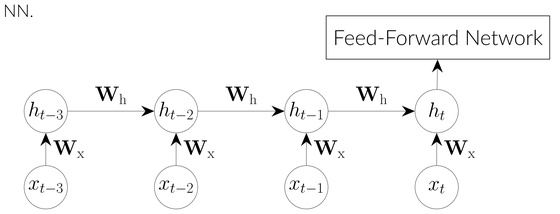


HAR: $RV_{t+h} = \beta_0 + \beta_d RV_t + \beta_w RV_t^{(5)} + \beta_m RV_t^{(22)} + \varepsilon_{t+h}$ with $RV_t^{(n)} = \frac{1}{n}\sum_{i=0}^{n-1} RV_{t-i}$ [4]

RNN Models

Sequential RNN forecasts are produced by stacking a two fully connected feed-forward layer on a single hidden RNN layer. We keep the RV input length fixed at 22, mirroring the standard HAR inputs.

Figure 1. Single-layer ERNN (length four input vector and no activation function) as input into a feed-forward NN.



- For the RNN layer, we use: Elman-RNNs (ERNNs), Long Short-Term-Memory cells (LSTMs), Gated Recurrent Units (GRUs).
- The feed-forward output layer is size 22 to produce the sequence output.
- We vary both the size of the hidden state vector in the recurrent units, the size of the first feed-forward layer, recurrent activation functions, and the inputs (univariate vs. multivariate).

Forecasting Performance

	ERNN			LSTM			GRU		
m_d	$h=1$	$h=5$	$h=22$	$h=1$	$h=5$	$h=22$	$h=1$	$h=5$	$h=22$
$m_h = 8$									
8	0.823*	0.986	1.662	0.841*	0.953	1.243	0.755*	0.945	1.171
16	0.765*	0.988	1.360	0.892*	0.985	1.154	0.809*	1.050	1.260
32	0.827*	0.952*	1.079	0.861*	0.962	1.020	0.870	1.084	1.610
$m_h = 16$									
8	0.832*	0.951	1.088	0.908	0.996	1.291	0.837*	0.984	1.140
16	0.775*	0.988	1.295	0.879*	1.003	1.242	0.794*	0.961	1.177
32	0.763*	0.978	1.190	0.892*	1.027	1.436	0.774*	0.991	1.592
$m_h = 64$									
8	0.877*	0.964	1.090	1.070	1.021	1.220	0.867*	1.097	1.653
16	0.828*	0.981	1.078	1.014	1.145	1.397	0.830*	1.052	1.747
32	0.757*	0.969	1.107	0.873*	0.996	1.350	0.817*	1.145	2.202

Table 1. Relative out-of-sample MSE of multivariate sequential RNN ensembles of size 10. The RNN forecasts are based on the average prediction across 10 estimated models each, with the selected models producing the lowest in-sample validation error out of 100 candidate models. m_h and m_d denote the number of hidden units in the RNN cell and the first feed-forward layer, respectively. *, **, and *** denote if the Diebold-Mariano test's null hypothesis of equal forecast accuracy is rejected at the 10%, 5%, and 1% level, respectively.

Complete Horizon Performance

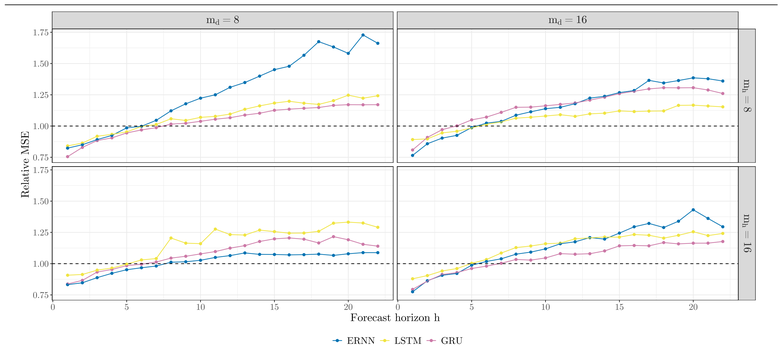


Figure 2. Relative out-of-sample MSE of multivariate sequential RNN ensembles of size 10 over the complete forecasting horizon. The RNN forecasts are based on the average prediction across 10 estimated models each, with the selected models producing the lowest in-sample validation error out of 100 candidate models. m_h and m_d denote the number of hidden units in the RNN cell and the first feed-forward layer, respectively.

Impact of Ensembles

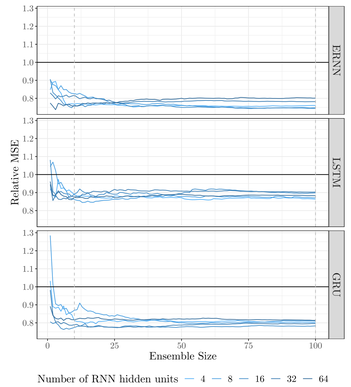


Figure 3. Relative out-of-sample MSE of multivariate sequential RNNs for $h = 1$ over ensemble sizes and number of hidden RNN units with number of feed-forward units kept constant at $m_d = 32$.

- Previously, **each model was estimated 100 times, with final predictions being simple averages out of the 10 model predictions**, sorted by validation error (as is common in the literature ([2] [3]).
- Figure 3 illustrates the relative MSE behavior for $h = 1$ as the ensemble size increases for the different RNNs.

⇒ **For small ensembles**, performance gained by ensembling is substantial, as **individual models do not generalize well** to the test date.

⇒ **For large ensembles**, prediction **error stays relatively consistent** as the ensemble size increases.

References

[1] Torben G. Andersen, Tim Bollerslev, Francis X. Diebold, and Heiko Ebens. The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1):43–76, 2001.

[2] Andrea Bucci. Realized Volatility Forecasting with Neural Networks. *Journal of Financial Econometrics*, 18(3):502–531, 2020.

[3] Kim Christensen, Mathias Siggaard, and Bezirgen Veliyev. A Machine Learning Approach to Volatility Forecasting. *Journal of Financial Econometrics*, 2022.

[4] F Corsi. A Simple Approximate Long-Memory Model of Realized Volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009.

[5] Ines Wilms, Jeroen Rombouts, and Christophe Croux. Multivariate volatility forecasts for stock market indices. *International Journal of Forecasting*, 37(2):484–499, 2021.