Using Neural Networks and Other Machine Learning Algorithms with Non-Standard Loss Functions in Realized Volatility Forecasting

Iakov Grigoryev

New Economic School

January 29, 2021

Outline

- Why Is It Important?

Outline

- Why Is It Important?
- Theory
    - realized volatility and integrated variance

Outline

- Why Is It Important?
- Theory
    - realized volatility and integrated variance
    - models: HAR, MEM, Random Forest, LSTM

Outline

- Why Is It Important?
- Theory
  - realized volatility and integrated variance
  - models: HAR, MEM, Random Forest, LSTM
  - loss functions: MSE, LinEx, ALS

Outline

- Why Is It Important?
- Theory
    - realized volatility and integrated variance
    - models: HAR, MEM, Random Forest, LSTM
    - loss functions: MSE, LinEx, ALS
- Empirical Research
    - data

Outline

- Why Is It Important?
- Theory
  - realized volatility and integrated variance
  - models: HAR, MEM, Random Forest, LSTM
  - loss functions: MSE, LinEx, ALS
- Empirical Research
  - data
  - results

Outline

- Why Is It Important?
- Theory
  - realized volatility and integrated variance
  - models: HAR, MEM, Random Forest, LSTM
  - loss functions: MSE, LinEx, ALS
- Empirical Research
  - data
  - results
  - plots

Outline

- Why Is It Important?
- Theory
  - realized volatility and integrated variance
  - models: HAR, MEM, Random Forest, LSTM
  - loss functions: MSE, LinEx, ALS
- Empirical Research
  - data
  - results
  - plots
- Conclusion

Why Is It Important?

Volatility Forecasting:

- asset allocation

Why Is It Important?

Volatility Forecasting:

- asset allocation
- risk management

Why Is It Important?

Volatility Forecasting:

- asset allocation
- risk management
- pricing derivatives

Why Is It Important?

Volatility Forecasting:

- asset allocation
- risk management
- pricing derivatives

Example:

- Black–Scholes model[1]

$$C(S_t, t) = N(d_1)S_t - N(d_2)PV(K),$$

where $d_1 = \frac{1}{\sigma\sqrt{T-1}}\left[\ln\left(\frac{S_t}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)(T-t)\right]$
and $d_2 = d_1 - \sigma\sqrt{T-t}$

[1]Black, F. & Scholes, M. (1973, Journal of Political Economy)

Theory: Realized Volatility and Integrated Variance

Notation:

- P_t is the asset price at time $t \in [0, T]$
- $p_t = \log P_t$
- $r_{t,\Delta t} = p_{t+\Delta t} - p_t$ is the asset return

Theory: Realized Volatility and Integrated Variance

Notation:

- P_t is the asset price at time $t \in [0, T]$
- $p_t = \log P_t$
- $r_{t,\Delta t} = p_{t+\Delta t} - p_t$ is the asset return

Diffusion equation:

$$dp_t = \mu_t dt + \sigma_t dW_t, \ 0 \leq t \leq T,$$

where

- W_t is Brownian motion
- μ_t and σ_t are predictable processes
- σ_t is independent of W_t

Theory: Realized Volatility and Integrated Variance

Asset return: $r_{t,\Delta t} = p_{t+\Delta t} - p_t$
Diffusion equation: $dp_t = \mu_t dt + \sigma_t dW_t, \ 0 \leq t \leq T$

Integrated variance:

$$IV_t = \int_0^{\Delta t} \sigma_{t+\tau}^2 d\tau$$

Theory: Realized Volatility and Integrated Variance

Asset return: $r_{t,\Delta t} = p_{t+\Delta t} - p_t$
Diffusion equation: $dp_t = \mu_t dt + \sigma_t dW_t,\ 0 \leq t \leq T$

Integrated variance:

$$IV_t = \int_0^{\Delta t} \sigma^2_{t+\tau} d\tau$$

Realized variance:

$$RV_t = \sum_{i=0}^{m-1} r^2_{t+\frac{\Delta t}{m}\cdot i, \frac{\Delta t}{m}}$$

Theory: Realized Volatility and Integrated Variance

Asset return: $r_{t,\Delta t} = p_{t+\Delta t} - p_t$
Diffusion equation: $dp_t = \mu_t dt + \sigma_t dW_t,\ 0 \leq t \leq T$

Integrated variance:

$$IV_t = \int_0^{\Delta t} \sigma_{t+\tau}^2 d\tau$$

Realized variance:

$$RV_t = \sum_{i=0}^{m-1} r_{t+\frac{\Delta t}{m}\cdot i, \frac{\Delta t}{m}}^2$$

Consistency: $m \to +\infty \Longrightarrow RV_t \xrightarrow{p} IV_t$

Theory: Realized Volatility and Integrated Variance

Asset return: $r_{t,\Delta t} = p_{t+\Delta t} - p_t$
Diffusion equation: $dp_t = \mu_t dt + \sigma_t dW_t,\ 0 \leq t \leq T$

Integrated variance:

$$IV_t = \int_0^{\Delta t} \sigma^2_{t+\tau} d\tau$$

Realized variance:

$$RV_t = \sum_{i=0}^{m-1} r^2_{t+\frac{\Delta t}{m}\cdot i, \frac{\Delta t}{m}}$$

Consistency: $m \to +\infty \Longrightarrow RV_t \xrightarrow{p} IV_t$

Realized volatility: $\sqrt{RV_t} = \sqrt{\sum_{i=0}^{m-1} r^2_{t+\frac{\Delta t}{m}\cdot i, \frac{\Delta t}{m}}} := RV_t$

Theory: The HAR Model

Denote: $RV_t^n = \frac{\sum_{i=0}^{n-1} RV_{t-i}}{n}$

Theory: The HAR Model

Denote: $RV_t^n = \frac{\sum_{i=0}^{n-1} RV_{t-i}}{n}$

Daily HAR model:

$$RV_t = \beta_0 + \beta_d RV_{t-1}^1 + \beta_w RV_{t-1}^5 + \beta_m RV_{t-1}^{22} + \epsilon_t$$

Theory: The HAR Model

Denote: $RV_t^n = \frac{\sum_{i=0}^{n-1} RV_{t-i}}{n}$
Daily HAR model:

$$RV_t = \beta_0 + \beta_d RV_{t-1}^1 + \beta_w RV_{t-1}^5 + \beta_m RV_{t-1}^{22} + \epsilon_t$$

- 'd' is 'day', 'w' is 'week', 'm' is 'month'

Theory: The HAR Model

Denote: $RV_t^n = \frac{\sum_{i=0}^{n-1} RV_{t-i}}{n}$

Daily HAR model:

$$RV_t = \beta_0 + \beta_d RV_{t-1}^1 + \beta_w RV_{t-1}^5 + \beta_m RV_{t-1}^{22} + \epsilon_t$$

- 'd' is 'day', 'w' is 'week', 'm' is 'month'
- we use $\log RV_t$ instead of RV_t

Theory: The MEM Model

Used only for time series with non-negative elements!

Theory: The MEM Model

Used only for time series with non-negative elements!

Model:

- $RV_t = \psi_t \eta_t$

Theory: The MEM Model

Used only for time series with non-negative elements!

Model:

- $RV_t = \psi_t \eta_t$
- $\eta_t \sim \exp(\lambda = 1)$

Theory: The MEM Model

Used only for time series with non-negative elements!

Model:

- $RV_t = \psi_t \eta_t$
- $\eta_t \sim \exp(\lambda = 1)$
- $\psi_t = \beta_0 + \beta_1 RV_{t-1} + \beta_2 \psi_{t-1}$

Theory: The MEM Model

Used only for time series with non-negative elements!

Model:

- $RV_t = \psi_t \eta_t$
- $\eta_t \sim \exp(\lambda = 1)$
- $\psi_t = \beta_0 + \beta_1 RV_{t-1} + \beta_2 \psi_{t-1}$
- $\beta_0 > 0$ ($\Rightarrow \psi_t > 0$)

Theory: The MEM Model

Used only for time series with non-negative elements!

Model:

- $RV_t = \psi_t \eta_t$
- $\eta_t \sim \exp(\lambda = 1)$
- $\psi_t = \beta_0 + \beta_1 RV_{t-1} + \beta_2 \psi_{t-1}$
- $\beta_0 > 0$ $(\Rightarrow \psi_t > 0)$

Log-likelihood function:

$$L = -\sum_{t=1}^{T} \left(\frac{RV_t}{\psi_t} + \log \psi_t \right) \longrightarrow \max_{\beta_0, \beta_1, \beta_2},$$

where $\psi_t = \beta_0 + \beta_1 RV_{t-1} + \beta_2 \psi_{t-1}$.

Theory: The Random Forest Model

Regressors: $RV_{t-1}, \ldots, RV_{t-k},\ t \in [k, T]$
Target: $RV_t,\ t \in [k, T]$

Theory: The Random Forest Model

Regressors: $RV_{t-1}, \ldots, RV_{t-k}, \;\; t \in [k, T]$
Target: $RV_t, \;\; t \in [k, T]$

Decision tree:

- Node N, data X

Theory: The Random Forest Model

Regressors: $RV_{t-1}, \ldots, RV_{t-k}, \; t \in [k, T]$
Target: $RV_t, \; t \in [k, T]$

Decision tree:

- Node N, data X
- $Q(X, F, T) > 0$?

Theory: The Random Forest Model

Regressors: $RV_{t-1}, \ldots, RV_{t-k},\ t \in [k, T]$
Target: $RV_t,\ t \in [k, T]$

Decision tree:

- Node N, data X
- $Q(X, F, T) > 0$?
- Yes $\Rightarrow$ feature F and threshold T (optimizing $Q(X, F, T)$):
  - $F < T \Rightarrow$ subtree with node N_{TRUE}, data X_{TRUE}
  - $F \geq T \Rightarrow$ subtree with node N_{FALSE}, data X_{FALSE}

Theory: The Random Forest Model

Regressors: $RV_{t-1}, \ldots, RV_{t-k},\ t \in [k, T]$
Target: $RV_t,\ t \in [k, T]$

Decision tree:

- Node N, data X
- $Q(X, F, T) > 0$?
- Yes $\Rightarrow$ feature F and threshold T (optimizing $Q(X, F, T)$):
  - $F < T \Rightarrow$ subtree with node N_{TRUE}, data X_{TRUE}
  - $F \geq T \Rightarrow$ subtree with node N_{FALSE}, data X_{FALSE}
- No $\Rightarrow$ N is a leaf node

Theory: The Random Forest Model

Regressors: $RV_{t-1}, \ldots, RV_{t-k},\ t \in [k, T]$
Target: $RV_t,\ t \in [k, T]$

Decision tree:

- Node N, data X
- $Q(X, F, T) > 0$?
- Yes $\Rightarrow$ feature F and threshold T (optimizing $Q(X, F, T)$):
  - $F < T \Rightarrow$ subtree with node N_{TRUE}, data X_{TRUE}
  - $F \geq T \Rightarrow$ subtree with node N_{FALSE}, data X_{FALSE}
- No $\Rightarrow$ N is a leaf node

$$Q(X, F, T) = H(X) - \frac{|X_{TRUE}|}{|X|} H(X_{TRUE}) - \frac{|X_{FALSE}|}{|X|} H(X_{FALSE}),$$

where $H(X) = \min\limits_{c \in \mathbb{R}} \frac{1}{|X|} \sum\limits_{t=k}^{T} (y^t - c)^2$

Theory: The Random Forest Model

Regressors: $RV_{t-1}, \ldots, RV_{t-k}, \;\; t \in [k, T]$
Target: $RV_t, \;\; t \in [k, T]$

Random Forest:

- Bootstrap n_{tree} samples

Theory: The Random Forest Model

Regressors: $RV_{t-1}, \ldots, RV_{t-k}, \ t \in [k, T]$
Target: $RV_t, \ t \in [k, T]$

Random Forest:

- Bootstrap n_{tree} samples
- On each sample: decision tree with a reduced number of regressors

Theory: The Random Forest Model

Regressors: $RV_{t-1}, \ldots, RV_{t-k}, \ t \in [k, T]$
Target: $RV_t, \ t \in [k, T]$

Random Forest:

- Bootstrap n_{tree} samples
- On each sample: decision tree with a reduced number of regressors
- Return the mean of n_{tree} predictions

Theory: Neural Network

General neural network:

- neurons, grouped in layers, connected to each other

Theory: Neural Network

General neural network:

- neurons, grouped in layers, connected to each other
- weights of neurons

Theory: Neural Network

General neural network:

- neurons, grouped in layers, connected to each other
- weights of neurons
- (non-linear) activation functions for each layer

Theory: Neural Network

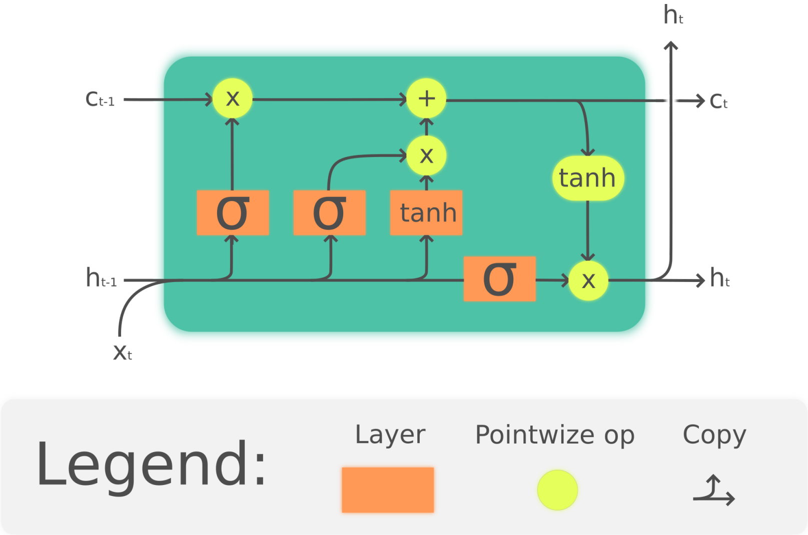
General neural network:

- neurons, grouped in layers, connected to each other
- weights of neurons
- (non-linear) activation functions for each layer

A single hidden layer neural network can approximate any non-linear function given enough number of neurons in this hidden layer.[2]

[2] Donaldson, G. & Kamstra, M. (1996, Journal of Forecasting)

Theory: The LSTM Model



Theory: Cost Function

Cost function:

$$C(y, \hat{y}) = \frac{1}{m} \sum_{i=1}^{m} L(y^i - \hat{y}^i),$$

where m is a number of examples, $L(\cdot)$ is a loss function.

Theory: The MSE Loss Function

MSE loss: $L(h) = h^2$

- widely used in regression problems

Theory: The MSE Loss Function

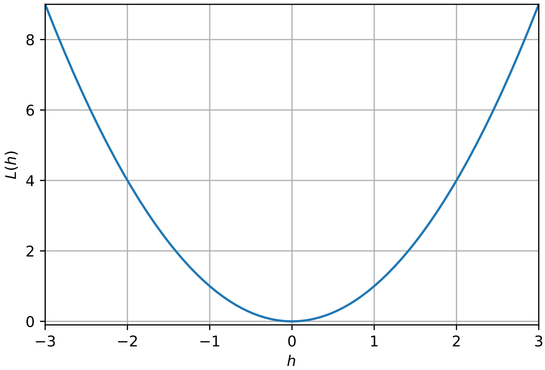
MSE loss: $L(h) = h^2$

- widely used in regression problems
- symmetric

Theory: The MSE Loss Function

MSE loss: $L(h) = h^2$

- widely used in regression problems
- symmetric



MSE loss

Theory: The LinEx Loss Function

LinEx loss: $L(h) = \exp(\alpha h) - \alpha h - 1$

- $\alpha > 0 \Rightarrow$ it is smaller for overprediction than for underprediction

Theory: The LinEx Loss Function

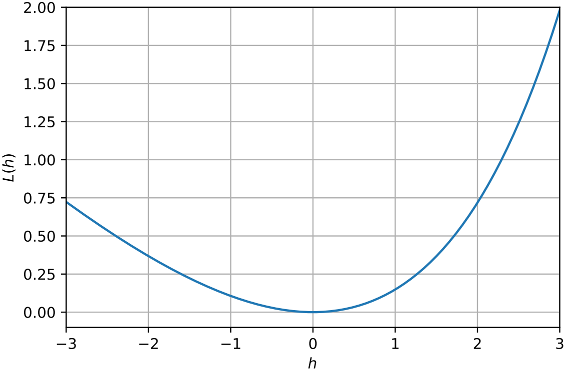
LinEx loss: $L(h) = \exp(\alpha h) - \alpha h - 1$

- $\alpha > 0 \Rightarrow$ it is smaller for overprediction than for underprediction
- $\alpha < 0 \Rightarrow$ vice versa

Theory: The LinEx Loss Function

LinEx loss: $L(h) = \exp(\alpha h) - \alpha h - 1$

- $\alpha > 0 \Rightarrow$ it is smaller for overprediction than for underprediction
- $\alpha < 0 \Rightarrow$ vice versa



LinEx loss, $\alpha = 0.5$

Theory: The ALS Loss Function

ALS loss: $L(h) = |\alpha - \mathbb{1}_{(h<0)}| \times h^2$

- $\alpha > 0.5 \Rightarrow$ it is smaller for overprediction than for underprediction

Theory: The ALS Loss Function

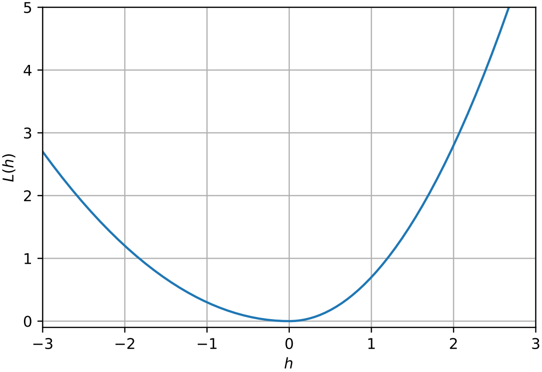
ALS loss: $L(h) = |\alpha - \mathbb{1}_{(h<0)}| \times h^2$

- $\alpha > 0.5 \Rightarrow$ it is smaller for overprediction than for underprediction
- $\alpha < 0.5 \Rightarrow$ vice versa

Theory: The ALS Loss Function

ALS loss: $L(h) = |\alpha - \mathbb{1}_{(h<0)}| \times h^2$

- $\alpha > 0.5 \Rightarrow$ it is smaller for overprediction than for underprediction
- $\alpha < 0.5 \Rightarrow$ vice versa



ALS loss, $\alpha = 0.7$

Empirical Research: Data

- from Oxford-Man Institute of Quantitative Finance library

Empirical Research: Data

- from Oxford-Man Institute of Quantitative Finance library
- daily realized volatility of S&P 500 Index (ticks of 5 minutes)

Empirical Research: Data

- from Oxford-Man Institute of Quantitative Finance library
- daily realized volatility of S&P 500 Index (ticks of 5 minutes)
- period from 1/3/2000 to 1/14/2021

Empirical Research: Data

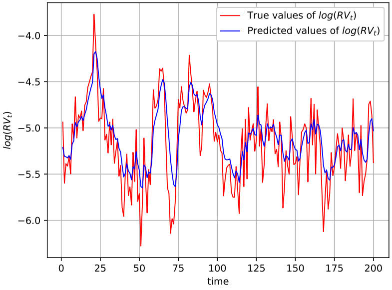
- from Oxford-Man Institute of Quantitative Finance library
- daily realized volatility of S&P 500 Index (ticks of 5 minutes)
- period from 1/3/2000 to 1/14/2021
- in-sample to out-of-sample proportion: 7 : 3

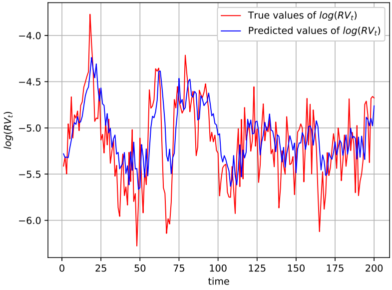
Empirical Research: Results

Model	MSE for RV_t	MSE for $\log(RV_t)$
MEM	4.67×10^{-6}	0.057
Random Forest	1.11×10^{-5}	0.136
HAR	1.21×10^{-5}	0.110
LSTM	3.96×10^{-5}	0.453

Model	LinEx for $\log(RV_t)$	ALS for $\log(RV_t)$
HAR	0.014	0.052
LSTM	0.067	0.246

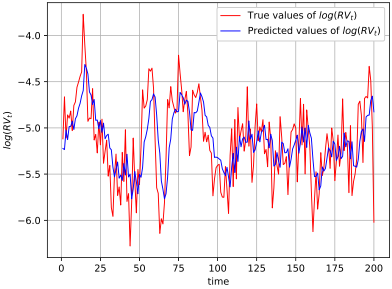
Empirical Research: Plots

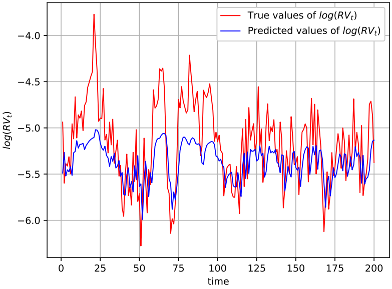




The MEM model with QMLE (left) and the Random Forest model with MSE loss (right), one-step-ahead predictions

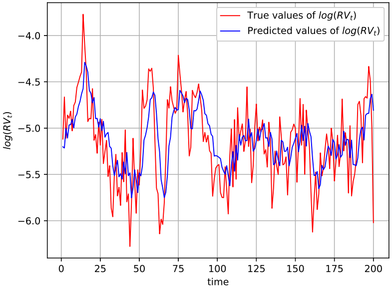
Empirical Research: Plots

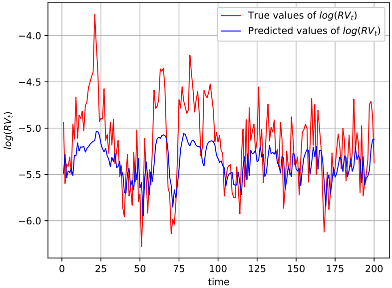




The HAR model (left) and the LSTM model (right) with MSE loss, one-step-ahead predictions

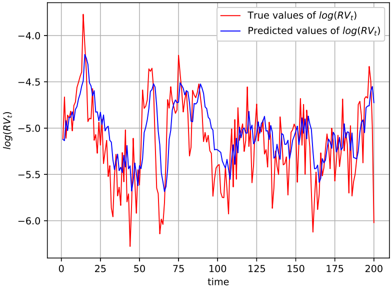
Empirical Research: Plots

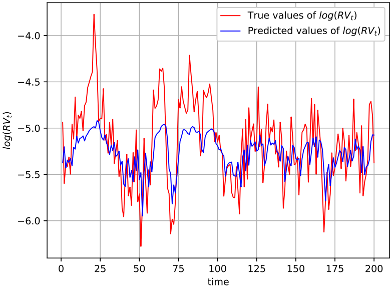




The HAR model (left) and the LSTM model (right) with LinEx loss, one-step-ahead predictions

Empirical Research: Plots





The HAR model (left) and the LSTM model (right) with ALS loss, one-step-ahead predictions

Conclusion

- MSE criterion
  - The MEM model is the best by margin

Conclusion

- MSE criterion
  - The MEM model is the best by margin
  - The Random Forest and the HAR models are similar in terms of predictive power

Conclusion

- MSE criterion
  - The MEM model is the best by margin
  - The Random Forest and the HAR models are similar in terms of predictive power
  - The LSTM model is the worst

Conclusion

- MSE criterion
  - The MEM model is the best by margin
  - The Random Forest and the HAR models are similar in terms of predictive power
  - The LSTM model is the worst
- LinEx/ALS loss & criterion
  - The HAR model is better in both cases than the LSTM model

Conclusion

- MSE criterion
  - The MEM model is the best by margin
  - The Random Forest and the HAR models are similar in terms of predictive power
  - The LSTM model is the worst
- LinEx/ALS loss & criterion
  - The HAR model is better in both cases than the LSTM model
- Further research
  - New criteria and loss functions

Conclusion

- MSE criterion
  - The MEM model is the best by margin
  - The Random Forest and the HAR models are similar in terms of predictive power
  - The LSTM model is the worst
- LinEx/ALS loss & criterion
  - The HAR model is better in both cases than the LSTM model
- Further research
  - New criteria and loss functions
  - New models

Conclusion

- MSE criterion
  - The MEM model is the best by margin
  - The Random Forest and the HAR models are similar in terms of predictive power
  - The LSTM model is the worst
- LinEx/ALS loss & criterion
  - The HAR model is better in both cases than the LSTM model
- Further research
  - New criteria and loss functions
  - New models
  - Addition of jumps