Using Machine Learning Algorithms in Realized Volatility Forecasting in Presence of Jumps

Iakov Grigoryev

New Economic School

April 16, 2021

Theory: Realized Volatility and Jumps

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 Jumps

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:

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

where

- W_t is Brownian motion
- μ_t and σ_t are predictable processes
- σ_t is independent of W_t
- q_t is the number of jumps with time-varying intensity κ_t

Theory: Realized Volatility and Jumps

<u>Asset return</u>: $r_{t,\Delta t} = p_{t+\Delta t} - p_t$
<u>Diffusion equation</u>: $dp_t = \mu_t dt + \sigma_t dW_t + \kappa_t dq_t, \; 0 \leq t \leq T$

<u>Realized variance</u>:

$$\mathrm{RV}_t^2 = \sum_{i=0}^{m-1} r^2_{t+\frac{\Delta t}{m}\cdot i, \frac{\Delta t}{m}} \xrightarrow[m\to\infty]{p} \underbrace{\int_0^{\Delta t} \sigma^2_{t+\tau} d\tau}_{\mathrm{IV}_t} + \underbrace{\sum_{t<\tau\leq t+\Delta t} \kappa_\tau^2}_{K_t}$$

Theory: Realized Volatility and Jumps

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

<u>Realized variance</u>:

$$\mathrm{RV}_t^2 = \sum_{i=0}^{m-1} r^2_{t+\frac{\Delta t}{m}\cdot i, \frac{\Delta t}{m}} \xrightarrow[m\to\infty]{p} \underbrace{\int_0^{\Delta t} \sigma^2_{t+\tau} d\tau}_{\mathrm{IV}_t} + \underbrace{\sum_{t<\tau\leq t+\Delta t} \kappa^2_\tau}_{K_t}$$

<u>Realized volatility</u>: $\mathrm{RV}_t = \sqrt{\sum_{i=0}^{m-1} r^2_{t+\frac{\Delta t}{m}\cdot i, \frac{\Delta t}{m}}}$

Theory: Decomposition into Continuous and Jump Parts

Consistency: $\mathrm{RV}_t^2 \xrightarrow[m\to\infty]{p} \mathrm{IV}_t + K_t$

Goal: find observable C_t and J_t such that $\mathrm{RV}_t^2 = C_t + J_t$,

$$C_t \xrightarrow[m\to\infty]{p} \mathrm{IV}_t, \text{ and } J_t \xrightarrow[m\to\infty]{p} K_t$$

Theory: Decomposition into Continuous and Jump Parts

Consistency: $\mathrm{RV}_t^2 \xrightarrow[m\to\infty]{p} \mathrm{IV}_t + K_t$

Goal: find observable C_t and J_t such that $\mathrm{RV}_t^2 = C_t + J_t$,

$$C_t \xrightarrow[m\to\infty]{p} \mathrm{IV}_t, \text{ and } J_t \xrightarrow[m\to\infty]{p} K_t$$

Approach: use median realized variance estimator[1]:

$$C_t \equiv \mathrm{MedRV}_t = \frac{\pi}{6 - 4\sqrt{3} + \pi}\left(\frac{m}{m-2}\right) \times$$

$$\times \sum_{i=1}^{m-2} \mathrm{Med}\left(\left|r_{t+\frac{\Delta t}{m}\cdot(i-1),\frac{\Delta t}{m}}\right|, \left|r_{t+\frac{\Delta t}{m}\cdot i,\frac{\Delta t}{m}}\right|, \left|r_{t+\frac{\Delta t}{m}\cdot(i+1),\frac{\Delta t}{m}}\right|\right)^2,$$

$$J_t = \mathrm{RV}_t^2 - C_t$$

[1] Andersen et al. (2012, JoE)

Theory: HAR-CJ Model (modified)

Denote:

$$c_t = \log C_t, \quad j_t = \log(J_t + 1)$$

$$c_t^n = \frac{\sum_{i=0}^{n-1} c_{t-i}}{n}, \quad j_t^n = \frac{\sum_{i=0}^{n-1} j_{t-i}}{n}$$

Theory: HAR-CJ Model (modified)

Denote:

$$c_t = \log C_t, \quad j_t = \log(J_t + 1)$$

$$c_t^n = \frac{\sum_{i=0}^{n-1} c_{t-i}}{n}, \quad j_t^n = \frac{\sum_{i=0}^{n-1} j_{t-i}}{n}$$

Daily HAR-CJ model (modified to estimate both parts of RV_t^2):

$$c_t = \beta_0^c + \beta_{cd}^c c_{t-1}^1 + \beta_{cw}^c c_{t-1}^5 + \beta_{cm}^c c_{t-1}^{22} + \beta_{jd}^c j_{t-1}^1 + \beta_{jw}^c j_{t-1}^5 + \beta_{jm}^c j_{t-1}^{22} + \epsilon_t^c$$

$$j_t = \beta_0^j + \beta_{cd}^j c_{t-1}^1 + \beta_{cw}^j c_{t-1}^5 + \beta_{cm}^j c_{t-1}^{22} + \beta_{jd}^j j_{t-1}^1 + \beta_{jw}^j j_{t-1}^5 + \beta_{jm}^j j_{t-1}^{22} + \epsilon_t^j$$

Theory: HAR-CJ Model (modified)

Denote:

$$c_t = \log C_t, \quad j_t = \log(J_t + 1)$$

$$c_t^n = \frac{\sum_{i=0}^{n-1} c_{t-i}}{n}, \quad j_t^n = \frac{\sum_{i=0}^{n-1} j_{t-i}}{n}$$

Daily HAR-CJ model (modified to estimate both parts of RV_t^2):

$$c_t = \beta_0^c + \beta_{cd}^c c_{t-1}^1 + \beta_{cw}^c c_{t-1}^5 + \beta_{cm}^c c_{t-1}^{22} + \beta_{jd}^c j_{t-1}^1 + \beta_{jw}^c j_{t-1}^5 + \beta_{jm}^c j_{t-1}^{22} + \epsilon_t^c$$

$$j_t = \beta_0^j + \beta_{cd}^j c_{t-1}^1 + \beta_{cw}^j c_{t-1}^5 + \beta_{cm}^j c_{t-1}^{22} + \beta_{jd}^j j_{t-1}^1 + \beta_{jw}^j j_{t-1}^5 + \beta_{jm}^j j_{t-1}^{22} + \epsilon_t^j$$

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

Theory: Random Forest Model

Regressors: $c_{t-1}, \dots, c_{t-k}, j_{t-1}, \dots, j_{t-k},\ t \in [k, T]$
Target: $(c_t, j_t),\ t \in [k, T]$

Theory: Random Forest Model

Regressors: $c_{t-1}, \ldots, c_{t-k}, j_{t-1}, \ldots, j_{t-k}, \;\; t \in [k, T]$
Target: $(c_t, j_t), \;\; t \in [k, T]$

Decision tree:

- Node N, data X

Theory: Random Forest Model

Regressors: $c_{t-1}, \ldots, c_{t-k}, j_{t-1}, \ldots, j_{t-k},\ t \in [k, T]$
Target: $(c_t, j_t),\ t \in [k, T]$

Decision tree:

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

Theory: Random Forest Model

Regressors: $c_{t-1}, \ldots, c_{t-k}, j_{t-1}, \ldots, j_{t-k},\ t \in [k, T]$
Target: $(c_t, j_t),\ t \in [k, T]$

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: Random Forest Model

Regressors: $c_{t-1}, \ldots, c_{t-k}, j_{t-1}, \ldots, j_{t-k}, \; t \in [k, T]$
Target: $(c_t, j_t), \; t \in [k, T]$

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: Random Forest Model

Regressors: $c_{t-1}, \ldots, c_{t-k}, j_{t-1}, \ldots, j_{t-k},\ t \in [k, T]$
Target: $(c_t, j_t),\ t \in [k, T]$

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 $H(X) = \min_{c \in \mathbb{R}} \frac{1}{|X|} \sum_{t=k}^{T} (y^t - c)^2$

Theory: Random Forest Model

Regressors: $c_{t-1}, \dots, c_{t-k}, j_{t-1}, \dots, j_{t-k}, \; t \in [k, T]$
Target: $(c_t, j_t), \; t \in [k, T]$

Random Forest:

- Bootstrap n_{tree} samples

Theory: Random Forest Model

Regressors: $c_{t-1}, \dots, c_{t-k}, j_{t-1}, \dots, j_{t-k}, \; t \in [k, T]$
Target: $(c_t, j_t), \; t \in [k, T]$

Random Forest:

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

Theory: Random Forest Model

Regressors: $c_{t-1}, \ldots, c_{t-k}, j_{t-1}, \ldots, j_{t-k},\ t \in [k, T]$
Target: $(c_t, j_t),\ t \in [k, T]$

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

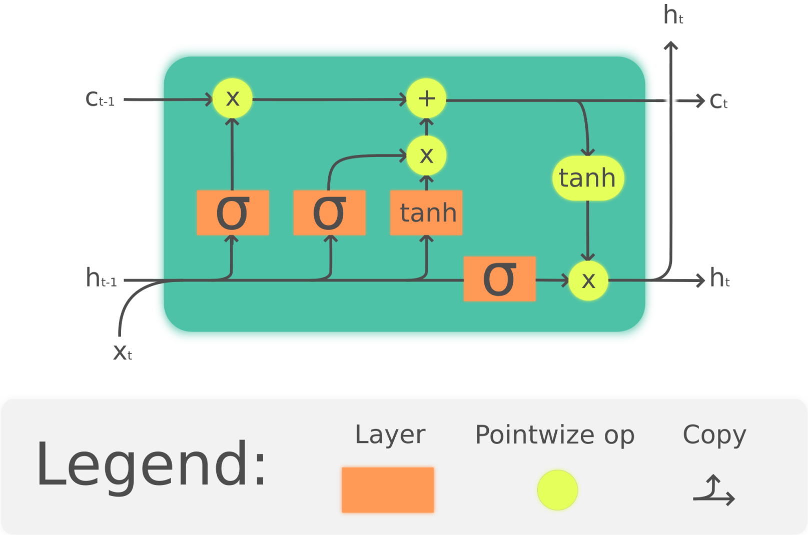
General neural network:

- 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 & Kamstra (1996, J. Forecast.)

Theory: LSTM Model



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 and median realized variance estimator

Empirical Research: Data

- from Oxford-Man Institute of Quantitative Finance library
- daily realized volatility and median realized variance estimator
- S&P 500 Index

Empirical Research: Data

- from Oxford-Man Institute of Quantitative Finance library
- daily realized volatility and median realized variance estimator
- S&P 500 Index
- ticks every 5 minutes

Empirical Research: Data

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

Empirical Research: Data

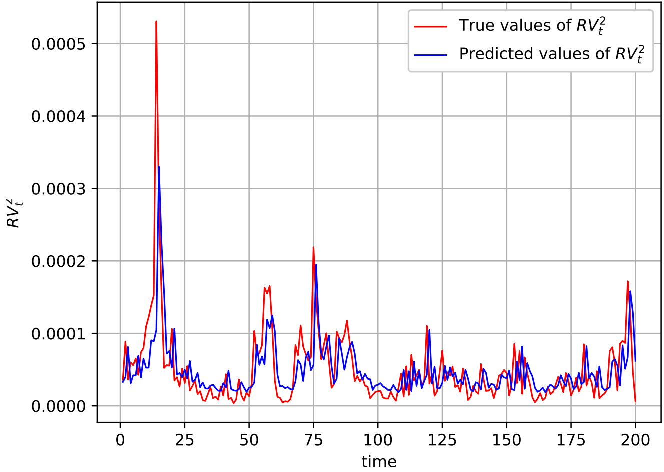
- from Oxford-Man Institute of Quantitative Finance library
- daily realized volatility and median realized variance estimator
- S&P 500 Index
- ticks every 5 minutes
- period from 1/3/2000 to 1/14/2021
- in-sample to out-of-sample proportion: 7 : 3

Empirical Research: Results

- predictions are one-step-ahead
- errors are calculated on out-of-sample data

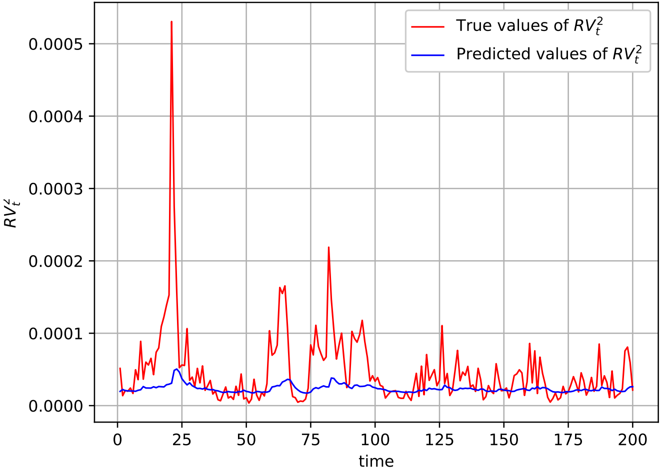
Error/Model	HAR-CJ	LSTM	Random Forest
MSE for c_t	0.518	1.697	0.520
MSE for j_t	2.52×10^{-8}	2.82×10^{-8}	2.07×10^{-8}
MSE for C_t	5.29×10^{-9}	1.37×10^{-8}	4.87×10^{-9}
MSE for J_t	2.52×10^{-8}	2.83×10^{-8}	2.08×10^{-8}
MSE for RV_t^2	3.61×10^{-8}	6.99×10^{-8}	3.36×10^{-8}

Empirical Research: Plots



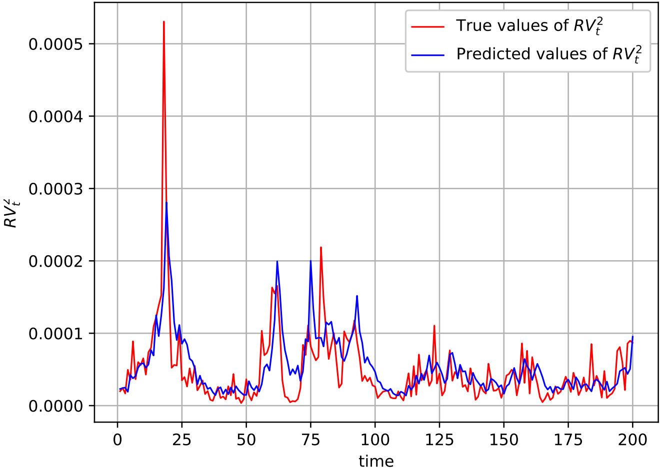
HAR-CJ model for RV_t^2, one-step-ahead predictions

Empirical Research: Plots



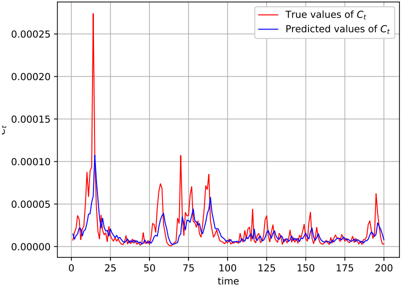
LSTM model for RV_t^2, one-step-ahead predictions

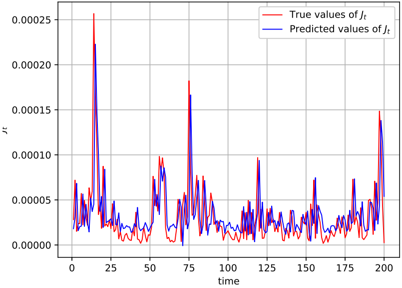
Empirical Research: Plots



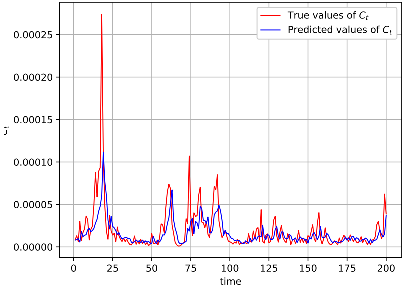
Random Forest model for RV_t^2, one-step-ahead predictions

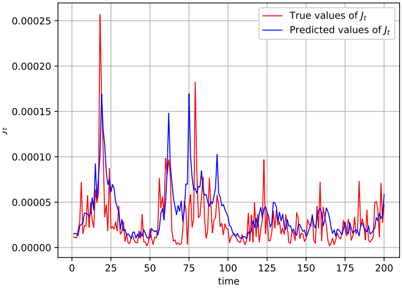
Empirical Research: Plots





HAR-CJ model for C_t (left) and J_t (right), one-step-ahead predictions





Random Forest model for C_t (left) and J_t (right), one-step-ahead predictions

Conclusion

- Random Forest predicts RV_t^2 (as a sum of continuous part and jumps) slightly better than HAR-CJ model.
- Random Forest predicts jumps of RV_t^2 better than HAR-CJ model.
- Random Forest model predicts continuous part of RV_t^2 approximately with the same accuracy as HAR-CJ model.
- LSTM model is the worst (due to small in-sample training set) in predicting continuous part, jumps and RV_t^2, even with regularization.