Advanced Time Series Econometrics

Week 4: Volatility Modelling with High-Frequency Data

Yi He

January 24, 2023

Plan for Today

1. Introduction
2. Continuous-time Processes
3. Realized Volatility
4. Modeling Realized Volatility
5. Forecasting Realized Volatility
6. Estimating Large-Dimensional Integrated Covariance Matrix

Introduction

High-Frequency Financial Data

- Traditional volatility model focus on daily asset log-returns $r_t = \Delta \log S_t$.
- Here S_t for $t \in \mathbb{N}$ is closing price at end of day t. So r_t is return on an investment buying asset at end of yesterday, and selling at end of today.
- However, assets such as stocks or foreign exchange are traded *during* the day, so price observations are available at times $\{t_i\}_{i=0}^{n_t}$ satisfying

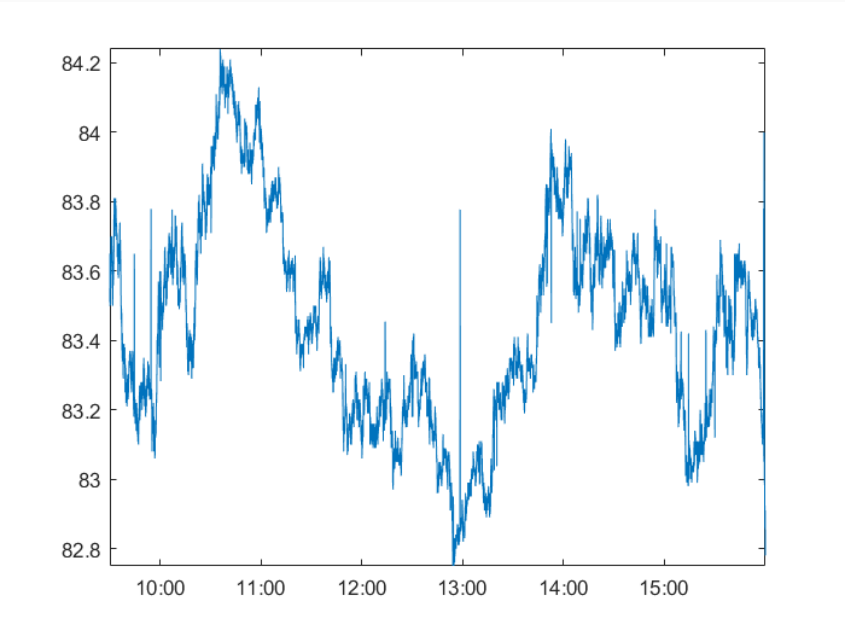
$$t - 1 < t_0 < t_1 < \ldots < t_{n_t} = t.$$

So S_{t_0} is opening price on day t (relevant when market closes, such as stock exchanges).

- The n_t intraday returns on day t:

$$r_{t,i} = \log(S_{t_i}/S_{t_{i-1}}), \qquad i = 1, \ldots, n_t$$

  may be used to estimate variance of daily return $r_t = \sum_{i=1}^{n} r_{t,i}$ (assuming zero overnight return).
- Intraday returns contain more information about volatility than daily returns.



SPY data (fund that follows S&P 500 index) on 2009/02/13

DATE	TIME_M	EX	SYM_ROOT	TR_SCOND	SIZE	PRICE	TR_CORR	TR_SEQNUM
20090213	"9:30:00.031"	T	SPY	T	400	83.5500	0	55492
20090213	"9:30:00.031"	T	SPY	T	500	83.5500	0	55493
20090213	"9:30:00.031"	T	SPY	T	300	83.5500	0	55494
20090213	"9:30:00.031"	T	SPY	T	300	83.5500	0	55495
20090213	"9:30:00.031"	T	SPY	T	300	83.5500	0	55496
20090213	"9:30:00.031"	T	SPY	T	500	83.5500	0	55497
20090213	"9:30:00.031"	T	SPY	T	300	83.5500	0	55498
20090213	"9:30:00.031"	T	SPY	T	200	83.5500	0	55499
20090213	"9:30:00.031"	T	SPY	T	300	83.5500	0	55514
20090213	"9:30:00.031"	T	SPY	T	300	83.5500	0	55515
20090213	"9:30:00.031"	T	SPY	T	200	83.5500	0	55516
20090213	"9:30:00.031"	T	SPY	FT	100	83.5500	0	55517
20090213	"9:30:00.031"	T	SPY	FT	300	83.5500	0	55518
20090213	"9:30:00.061"	T	SPY	FT	200	83.5500	0	55519
20090213	"9:30:00.061"	T	SPY	FT	400	83.5500	0	55520
20090213	"9:30:00.061"	T	SPY	T	100	83.5500	0	55521
20090213	"9:30:00.061"	T	SPY	FT	500	83.5600	0	55522
20090213	"9:30:00.061"	T	SPY	FT	500	83.5600	0	55523
20090213	"9:30:00.061"	T	SPY	FT	200	83.5600	0	55524
20090213	"9:30:00.061"	T	SPY	FT	160	83.5600	0	55525
20090213	"9:30:00.061"	T	SPY	FT	300	83.5600	0	55526
20090213	"9:30:00.061"	T	SPY	FT	238	83.5600	0	55527
20090213	"9:30:00.061"	T	SPY	FT	200	83.5600	0	55528
20090213	"9:30:00.061"	T	SPY	FT	300	83.5600	0	55529
20090213	"9:30:00.061"	T	SPY	FT	300	83.5600	0	55530
20090213	"9:30:00.061"	T	SPY	FT	300	83.5600	0	55531
20090213	"9:30:00.061"	T	SPY	FT	300	83.5600	0	55532
20090213	"9:30:00.061"	T	SPY	FT	300	83.5600	0	55533

Millisecond SPY Data

- Data on all trades and quotes for a particular stock within a day are collected and sold by commercial parties. An important example is the TAQ database, which covers stocks traded on the New York Stock Exchange (NYSE).
- Multiple trades in the same fund against different prices, on different markets, within the same one-second period.
- Leads to large data files; data processing cannot be done using e.g. MATLAB with text or Excel files. Substantial data cleaning is needed before analysing the data.

Continuous-time Processes

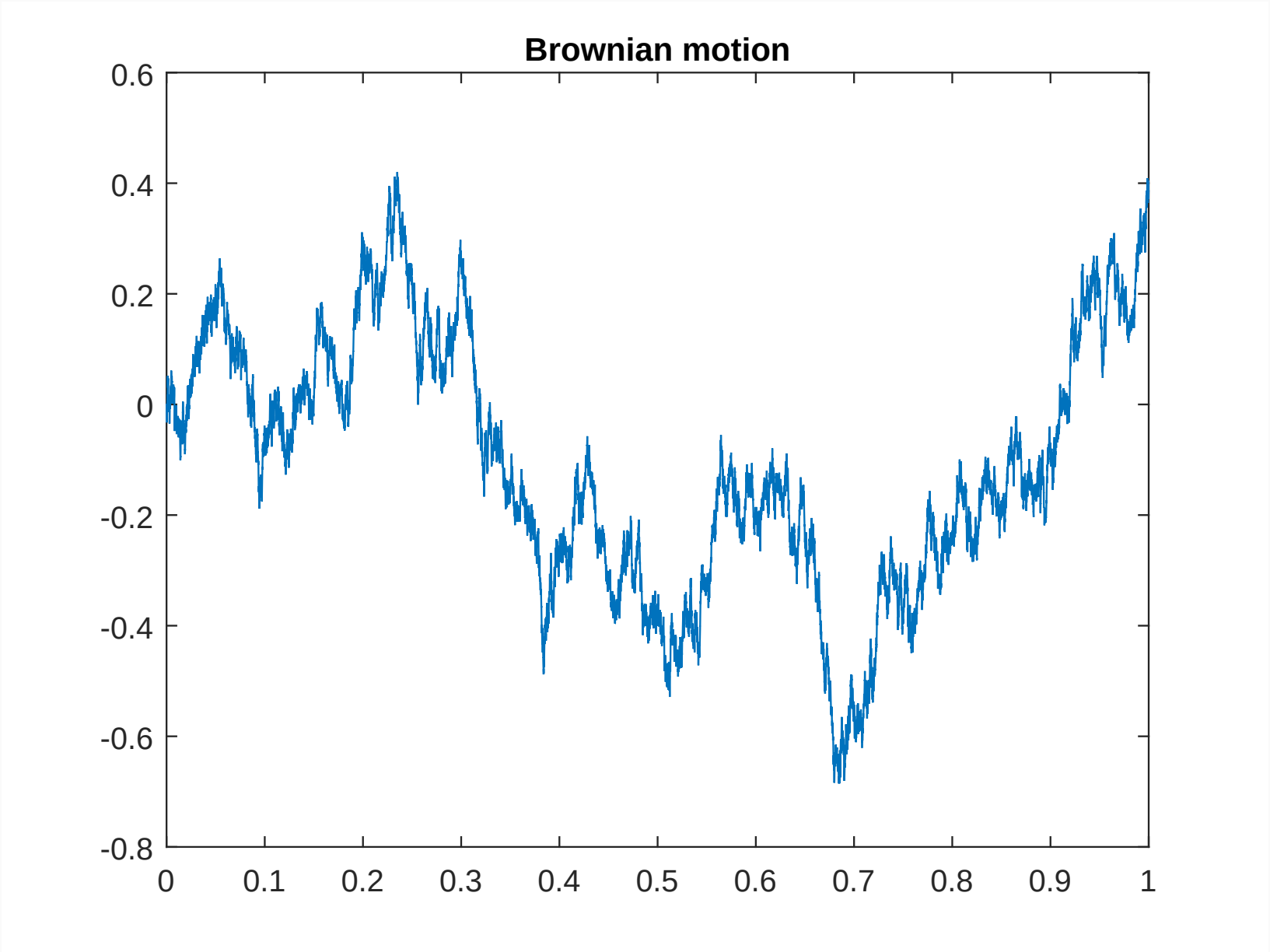
Continuous-time Processes

- **Brownian motion** $\{W_t\}_{t\geq 0}$ is stochastic process satisfying:
  1. $W_0 = 0$;
  2. $W_{t+s} - W_t \sim N(0, s)$ for $t \geq 0$ and $s \geq 0$;
  3. independent increments;
  4. continuous sample paths.
- Main building block for continuous-time processes.
- Continuous but non-differentiable sample paths, of unbounded variation: with probability one

$$\sup_{0\leq t_0\leq \ldots \leq t_n \leq t} \sum_{i=1}^{n} \left| W_{t_i} - W_{t_{i-1}} \right| = \infty.$$

- **Martingale** w.r.t. its own **filtration** $\{\mathcal{F}_t^W\}_{t\geq 0}$, with $\mathcal{F}_t^W = \sigma(\{W_s\}_{0\leq s\leq t})$:

$$E(W_t|\mathcal{F}_s^W) = W_s, \qquad 0 \leq s \leq t.$$


Brownian motion
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Itô Processes

- **Itô integral** is defined as

$$\int_0^t \sigma_s dW_s = \lim_{n\to\infty} \sum_{i=0}^{n-1} \sigma_{t_i}(W_{t_{i+1}} - W_{t_i}),$$

  where:
  - **volatility** $\{\sigma_t\}_{t\geq 0}$ is adapted to $\{\mathcal{F}_t\}_{t\geq 0}$, and $\{W_t\}_{t\geq 0}$ is martingale w.r.t. $\{\mathcal{F}_t\}_{t\geq 0}$;
  - $P\left[\int_0^T \sigma_t^2 dt < \infty\right] = 1$;
  - the limit is in L^2, with $0 = t_0 \leq \ldots \leq t_n = t$ and $\max_{1\leq i\leq n} |t_i - t_{i-1}| \to 0$.
- Leads to class of **Itô processes** (Brownian semimartingales):

$$\begin{aligned} X_t &= X_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s, \\ \Leftrightarrow \quad dX_t &= \mu_t dt + \sigma_t dW_t, \end{aligned}$$

  where **drift** $\{\mu_t\}_{t\geq 0}$ is adapted to $\{\mathcal{F}_t\}_{t\geq 0}$ and integrable.

Example 1

Brownian motion with drift:

$$dX_t = \mu dt + \sigma dW_t.$$

- Black-Scholes model for log-asset prices ($X_t = \log S_t$).
- Discrete-time observations $\{x_i = X_{i\Delta t}\}_{i=0}^{n}$ follow Gaussian random walk with drift:

$$x_i = x_{i-1} + \mu\Delta t + \varepsilon_i,$$

  with $\varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2\Delta t)$.
- Hence normal log-returns $r_i = \Delta x_i$.

Example 2

Ornstein-Uhlenbeck (OU) process:

$$dX_t = \alpha(\mu - X_t)dt + \sigma dW_t, \qquad \alpha > 0.$$

- Displays mean-reversion to μ.
- Discrete-time observations $\{x_i = X_{i\Delta t}\}_{i=0}^{n}$ follow Gaussian AR(1) process:

$$x_i = \mu + \rho(x_{i-1} - \mu) + \varepsilon_i,$$

with $\rho = e^{-\alpha\Delta t}$ and $\varepsilon_i \sim$ i.i.d. $N(0, \sigma^2(1 - e^{-2\alpha\Delta t})/(2\alpha))$.

Itô's formula

- **Itô's formula**: if $dX_t = \mu_t dt + \sigma_t dW_t$ and $Y_t = f(t, X_t)$, then

$$\begin{aligned} dY_t &= \left[\dot{f}(t, X_t) + \tfrac{1}{2} f''(t, X_t)\sigma_t^2\right] dt + f'(t, X_t) dX_t \\ &= \left[\dot{f}(t, X_t) + f'(t, X_t)\mu_t + \tfrac{1}{2} f''(t, X_t)\sigma_t^2\right] dt \\ &\quad + f'(t, X_t)\sigma_t dW_t, \end{aligned}$$

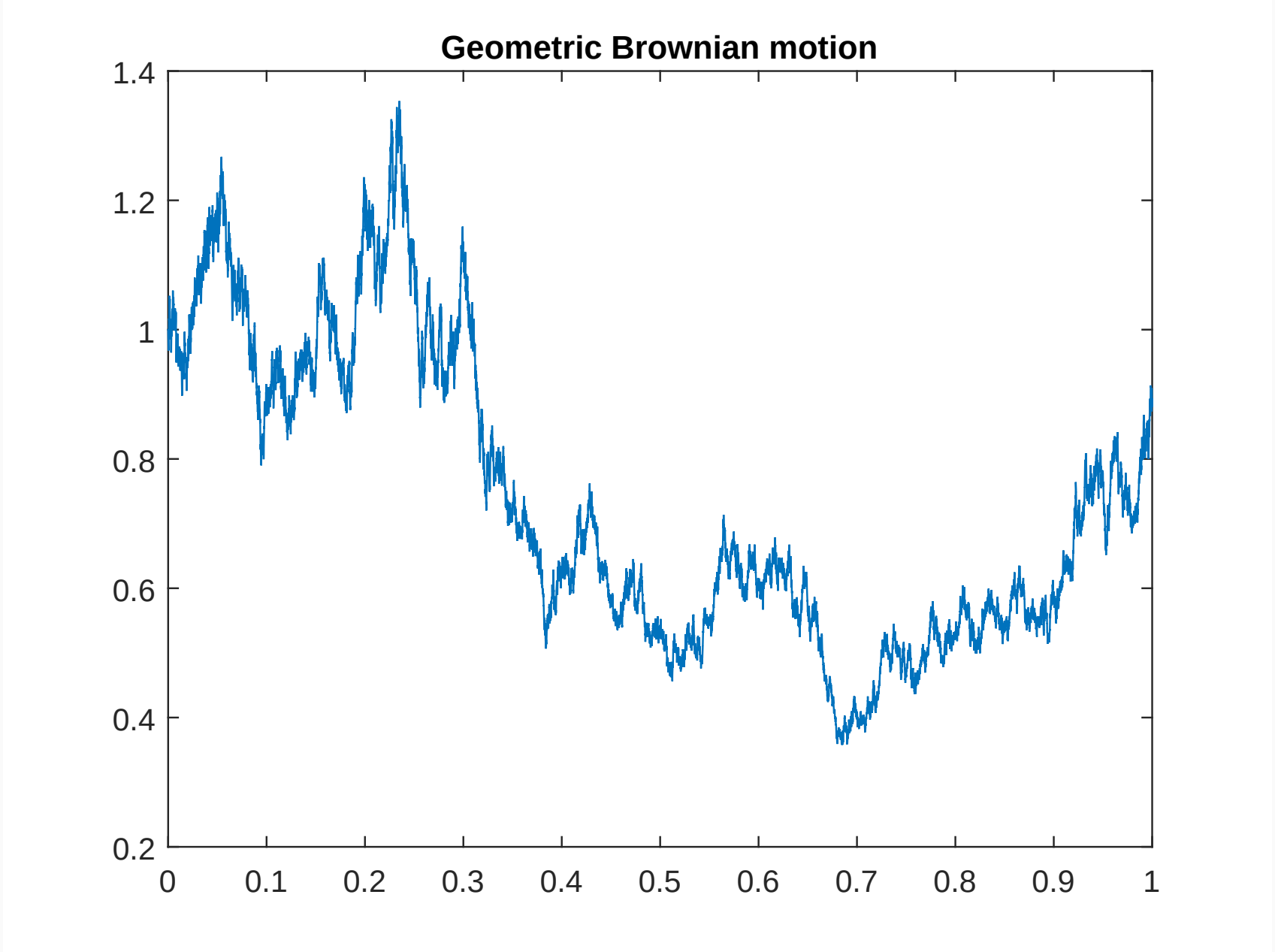
  where

$$\dot{f}(t,x) = \frac{\partial f(t,x)}{\partial t}, \; f'(t,x) = \frac{\partial f(t,x)}{\partial x}, \; f''(t,x) = \frac{\partial^2 f(t,x)}{\partial x^2}.$$

- Example: if $dX_t = \mu dt + \sigma dW_t$ and $S_t = \exp(X_t)$, then

$$dS_t = \left(\mu + \tfrac{1}{2}\sigma^2\right) S_t dt + \sigma S_t dW_t.$$

  Process $\{S_t\}_{t \geq 0}$ is called **Geometric Brownian motion**.



Geometric BM with $\sigma = 1$ and $\mu = -0.5$

Jump-Diffusion Process

- **Poisson process** $\{N_t\}_{t\geq 0}$ is integer-valued stochastic process with:
  1. $N_0 = 0$;
  2. $N_{t+s} - N_t \sim \text{POISSON}(\lambda s)$ for $t \geq 0$ and $s \geq 0$;
  3. independent increments.
- Jump times $\{\tau_i\}_{i=1}^{N_t}$ defined by $N_{\tau_i} = N_{\tau_i -} + 1$, where $N_{\tau_i -} = \lim_{t \uparrow \tau_i} N_t$.
- Integral with respect to N_t is defined as

$$\int_0^t \kappa_s dN_s = \sum_{i=1}^{N_t} \kappa_{\tau_i} = \sum_{0<s\leq t} J_s, \qquad J_t = \kappa_t dN_t.$$

- Leads to Itô process with jumps, or **jump-diffusion**:

$$dX_t = \mu_t dt + \sigma_t dW_t + \kappa_t dN_t,$$

where $\{\mu_t, \sigma_t, \kappa_t\}_{t\geq 0}$ are predictable ($\mathcal{F}_{t-}$-measurable).

Quadratic Variation

- For any semimartingale $\{X_t\}_{t\geq 0}$ (sum of local martingale and bounded variation process), the **quadratic variation** process is

$$QV_t = [X]_t = \int_0^t (dX_s)^2 = \text{plim}_{n\to\infty} \sum_{i=1}^{n} (X_{t_i} - X_{t_{i-1}})^2,$$

  where again $0 = t_0 \leq \ldots \leq t_n = t$ and $\max_{1\leq i\leq n} |t_i - t_{i-1}| \to 0$.
- Examples:
  - Brownian motion: $[W]_t = t$.
  - Poisson proces: $[N]_t = N_t$.
  - Jump-diffusion $dX_t = \mu_t dt + \sigma_t dW_t + \kappa_t dN_t$:

$$[X]_t = \int_0^t \sigma_s^2 ds + \sum_{0<s\leq t} J_s^2,$$

    or $QV_t = IV_t + \sum_{0\leq s\leq t} J_s^2$, where $IV_t = \int_0^t \sigma_s^2 ds$ (**integrated variance**).

Realized Volatility

Realized Volatility

- Assume we have $n+1$ intra-day observations $\{X_{t,i}\}_{i=0}^{n}$ on the log-stock price for day t, where

$$X_{t,i} = X_{t-1+i/n},$$

  so $X_{t,n} = X_t$ is closing price for day t.
- Corresponding intra-day log-returns are $r_{t,i} = X_{t,i} - X_{t,i-1}, i = 1, \ldots, n$.
- Time interval is $\delta = 1/n$; could be generalized to varying δ_i.
- The **realized variance** is the sample variance of $\{r_{t,i}\}_{i=1}^{n}$, assuming zero mean and expressed as variance of daily return $r_t = \sum_{i=1}^{n} r_{t,i}$:

$$RV_{n,t} = \sum_{i=1}^{n} r_{t,i}^2.$$

- The **realized volatility** is simply the square root of $RV_{n,t}$.

Realized Volatility – Brownian Motion with Drift

- Brownian motion with drift $dX_t = \mu dt + \sigma dW_t$ implies

  $r_{t,i} = \mu\delta + \sigma(W_{t_i} - W_{t_{i-1}}) = \mu\delta + \sigma\sqrt{\delta}Z_i,\ Z_i \overset{iid}{\sim} N(0,1),\ i = 1, \ldots, n.$

- Law of large numbers for Z_i^2:

$$\begin{aligned} RV_{n,t} = \sum_{i=1}^{n} r_{t,i}^2 &= n\mu^2\delta^2 + 2\mu\sigma\delta^{3/2}\sum_{i=1}^{n} Z_i + \sigma^2\delta\sum_{i=1}^{n} Z_i^2 \\ &= \frac{\mu^2}{n} + \frac{2\mu\sigma}{n}(W_t - W_{t-1}) + \sigma^2\frac{1}{n}\sum_{i=1}^{n} Z_i^2 \\ &\xrightarrow{p} \sigma^2. \end{aligned}$$

  Note that mean return μ has negligible effect

- Central limit theorem for $Z_i^2 - 1$:

$$\sqrt{n}\left(RV_{n,t} - \sigma^2\right) = \sigma^2\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i^2 - 1) + o_p(1) \xrightarrow{d} N(0, 2\sigma^4).$$

Realized Volatility – Jumps

- If log-price is jump diffusion $dX_t = \mu_t dt + \sigma_t dW_t + \kappa_t dN_t$, then

$$\text{plim}_{n\to\infty} RV_{n,t} = QV_{t-1}^t = IV_{t-1}^t + \sum_{t-1<s\leq t} J_s^2.$$

- Barndorff-Nielsen and Shephard (2004) propose to estimate integrated variance using **realized bipower variation**

$$BV_{n,t} = \frac{\pi}{2} \sum_{i=2}^{n} |r_{t,i}| \, |r_{t,i-1}| \, .$$

Factor $\pi/2$ is needed because $E[|Z_i|] = \sqrt{2/\pi}$. Consistent estimator of IV_{t-1}^t because probability of two subsequent jumps is negligible.

- Alternative approach to estimate integrated variance is **truncation** (Aït-Sahalia and Jacod, 2014):

$$TRV_n = \sum_{i=1}^{n} r_{t,i}^2 \mathbf{1}_{\{|r_{t,i}|<a_n\}},$$

where a_n is threshold converging to 0 at a rate slower than $n^{-1/2}$.

Market microstructure

- In practice log-asset prices are not exact semimartingales, because of **market microstructure** (trading mechanisms and their effect on asset prices).
- Market makers are traders that provide liquidity (willingness to act as counterparty for others). They charge higher price when selling an asset (ask price S_a) than when buying same asset at same time (bid price S_b).
- The bid-ask spread $\Delta = S_a - S_b$ implies that traded price S_t may differ from fundamental price S_t^*, e.g. as

$$S_t = \begin{cases} S_t^* + \frac{1}{2}\Delta, & \text{if trade is buyer-initiated,} \\ S_t^* - \frac{1}{2}\Delta, & \text{if trade is seller-initiated.} \end{cases}$$

- Simple model for microstructure noise is

$$\begin{aligned} Y_t &= \log S_t = X_t + U_t, \\ dX_t &= \mu_t dt + \sigma_t dW_t, \end{aligned}$$

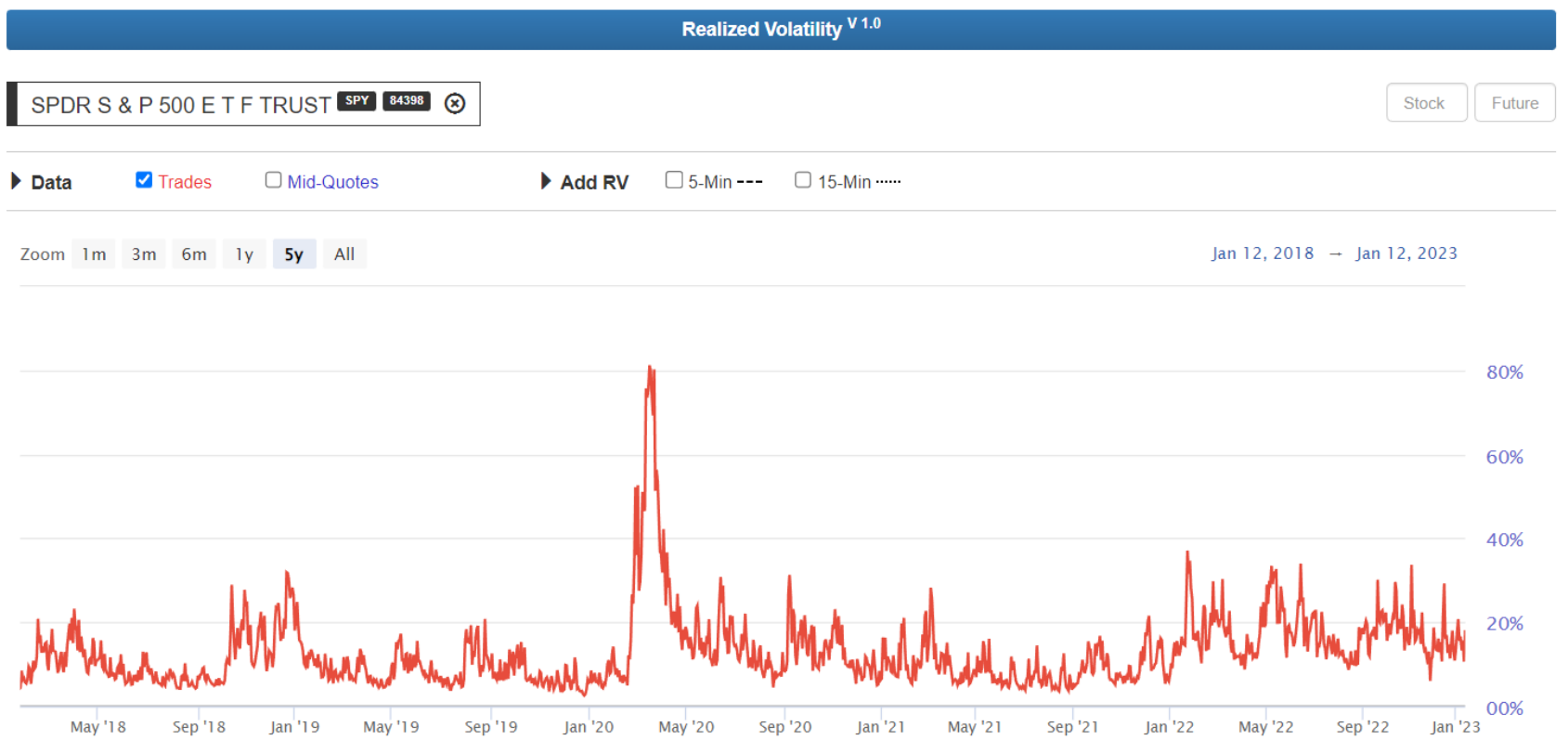
  where U_t is white noise (or moving-averaging process) with variance ω^2, independent of X_t.
- Implies that $RV_{n,t}$ is not consistent: using

$$RV_{n,t} = \sum_{i=1}^{n} \Delta Y_{t,i}^2 = \sum_{i=1}^{n} \Delta X_{t,i}^2 + \sum_{i=1}^{n} \Delta U_{t,i}^2 + 2 \sum_{i=1}^{n} \Delta X_{t,i} \Delta U_{t,i},$$

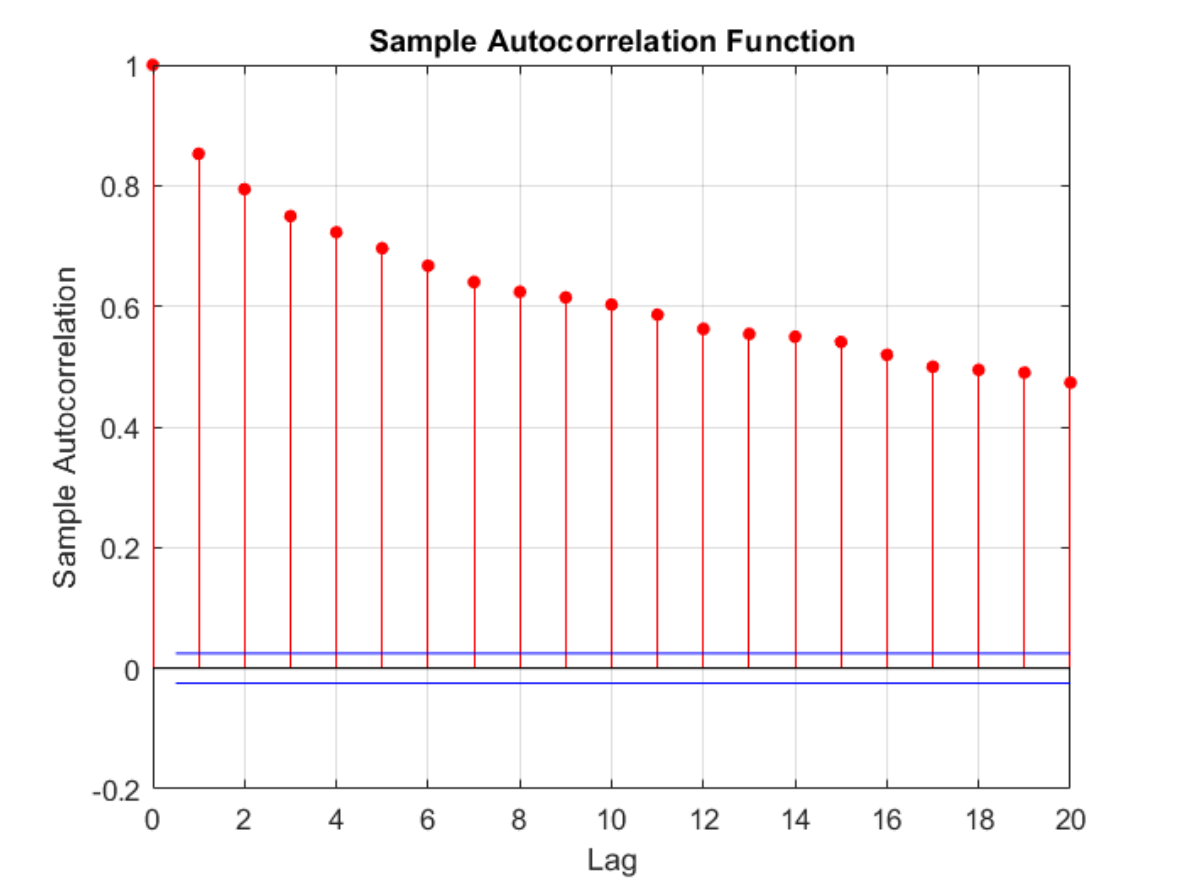
  it follows that $\text{plim}_{n\to\infty} n^{-1} RV_{n,t} = 2\omega^2 = \text{var}(\Delta U_t)$. So in the limit, the noise dominates the signal.
- Reason for using, e.g., 20-minute returns in construction of RV, instead of all available high-frequency data.
- Da and Xiu, 2021: Estimate QV_t allowing for moving-average microstructure noise

Modeling Realized Volatility

I extract the realized volatility from Dacheng Xiu's website:



Daily realized volatility of S&P500 index



ACF of log-realized variance of S&P500 index, 2010/01/01–2023/01/20.

Unit root can be easily rejected.

Modeling Realized Variance

- Daily (log-)realized variance time series display long memory: autocorrelations decay slowly, hyperbolically. Yet the time series does not appear to be $I(1)$.
- Suggests the use of fractionally integrated models:

$$\phi_p(L)(1-L)^d \log RV_t = \theta_q(L)\varepsilon_t$$

  with $d \approx 0.5$ in our example.
- Alternative long-memory model is so-called heterogeneous autoregressive, HAR model (Corsi, 2009):

$$RV_{t+1} = \beta_0 + \beta_D RV_t + \beta_W RV_{t-5,t} + \beta_M RV_{t-22,t} + \varepsilon_{t+1},$$

  where $RV_{t-h,t} = h^{-1}(RV_{t-h+1} + \ldots + RV_t)$
- ... or the logHAR model replacing RV with $\log RV$

Realized Variance In GARCH Models

- Realized variance can also be used as explanatory variable in GARCH model for daily returns:

$$\begin{aligned} r_t &= \mu_t + \varepsilon_t = \mu_t + h_t^{1/2} z_t, \\ h_t &= \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1} + \gamma RV_{t-1}. \end{aligned}$$

- Estimation typically leads to $\widehat{\alpha} \approx 0$, $\widehat{\beta} + \widehat{\gamma} \approx 1$, and $\widehat{\beta} < 0.8$. So RV_{t-1} is much more informative than ε_{t-1}^2 for forecasting daily volatility, and h_t reacts relatively quickly to news.

- Multi-step volatility forecasting requires an additional equation for RV_t. Example is the realized GARCH model (Hansen, Huang and Shek, 2012):

$$\begin{aligned} h_t &= \omega + \beta h_{t-1} + \gamma RV_{t-1}, \\ RV_t &= \delta_0 + \delta_1 h_t + \delta_2 z_t + \delta_3 (z_t^2 - 1) + u_t, \end{aligned}$$

  where u_t is an error term and $\delta_2 z_t + \delta_3 (z_t^2 - 1)$ reflects the leverage effect.

Forecasting Realized Volatility

Paye (2012)

$$\log RV_t = \alpha + \rho_1 \log RV_{t-1} + \rho_2 \log RV_{t-2} + \beta' X_{t-1} + \varepsilon_t$$

- Daily return on S&P 500 index rather than intraday data
- X_t includes a set of macroeconomic variables
- In-sample: some variables are useful
- Out-of-sample: "*It is more difficult to find evidence that forecasts exploiting macroeconomic variables outperform a univariate benchmark out-of-sample.*"
- Asymmetry: "*Predictive power associated with macroeconomic variables appears to concentrate around the onset of recessions.*"

Bucci (2020)

- Also use daily data rather than intraday data
- The NARX (Nonlinear AutoRegressive with eXogenous inputs) model could outperform traditional time series model

$$\log RV_t = f\left(\log RV_{t-1}, \ldots, \log RV_{t-q}, X_t\right) + \varepsilon_t$$

- Fit f by using neural networks: to be discussed later
- In particular, including X_t improves prediction MSE substantially
- LSTM neural networks works similarly: not to be discussed

Christensen K., Siggaard M., Veliyev B. (2022)

- RV based on intraday data
- In general, consider the regression model

$$RV_t = f(Z_{t-1}; \beta) + \varepsilon_t$$

where Z_{t-1} is a set of lagged economic variables, including the variables used by (log)HAR

- Parametric Models: Regularized linear regression using Ridge, LASSO or elastic-net regression

$$f(Z_{t-1}) = \beta_0 + Z'_{t-1}\beta$$

- Non-parametric Models: bagging (random forest), boosting, neural network, assuming f is a (nonlinear) smooth function
- Low-dimensional setting: a few variables with small p/n

Elastic Net Regression

$$\frac{1}{2n}\sum_{t=1}^{n}(Y_t - \beta_0 - X_t'\beta)^2 + \lambda\left[(1-\alpha)\frac{1}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1\right]$$

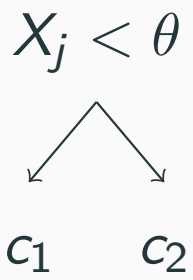
where

$$\|\beta\|_2^2 = \sum_{j=1}^{p}\beta_j^2, \quad \|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$$

- Take $Y_t = RV_t$ and $X_t = Z_{t-1}$.
- $\alpha = 1$ gives LASSO, $\alpha = 0$ gives ridge, in general for $0 < \alpha \leq 1$ there is also variable selection
- Ridge regression often outperforms LASSO in economic forecasting: same in CSV2022
- Elastic net works similarly as Ridge in CSV2022
- Variable selection does not help forecasting here, perhaps because the population model is dense rather than sparse.

Regression Stump

Target $Y \in \mathbb{R}$ and features $X = (X_1, \ldots, X_p)'$



- Make predictions at the terminal nodes (leafs)
- root node = a condition for a **single** feature
- Condition TRUE $\rightarrow$ left child
- Condition FALSE $\rightarrow$ right child
- Given j and θ, minimizing the mean squared error gives the optimal predictions

$$c_1 = \mathbb{E}[Y|X_j < \theta], \quad c_2 = \mathbb{E}[Y|X_j > \theta]$$

In practice, use the sample averages.

Least-Squares Boosting

- Initialize $F^{(0)}(x)$ with zero or a constant $F^{(0)}(x) = \bar{Y}$ the sample average of the target values.
- For $m = 1$ to M do:
  1. Calculate the residuals $\widetilde{Y}_i = Y_i - F^{(m-1)}(X_i)$
  2. Fit a base to the residuals:

$$f_m^* = \underset{f \in \mathcal{G}}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} \left(f(X_i) - \widetilde{Y}_i^{(m-1)} \right)^2 .$$

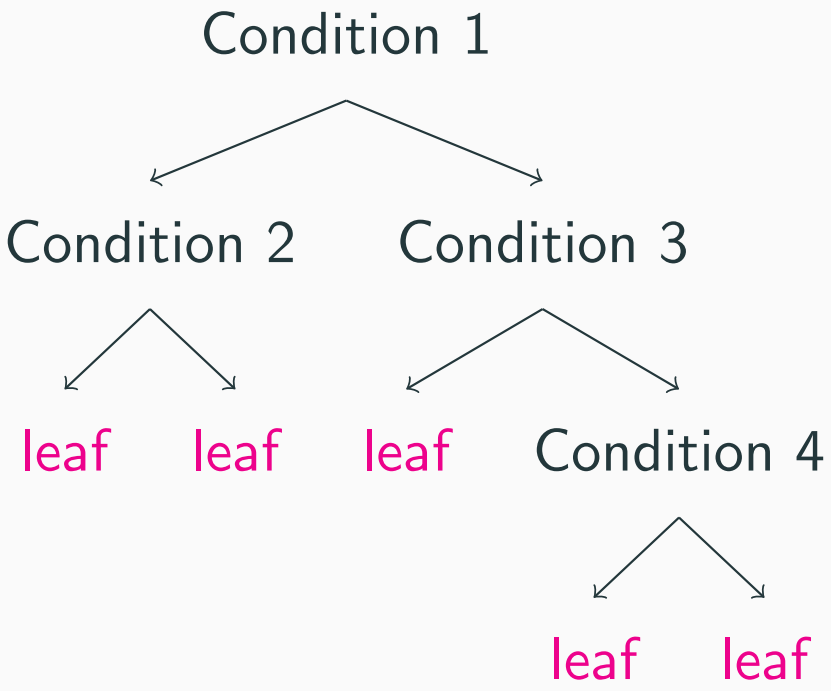
  3. Set $f_m = \eta \cdot f_m^* \in \mathcal{G}$. Accumulate the base learners:

$$F^{(m)} = F^{(m-1)} + f_m = F^{(m-1)} + \eta \cdot f_m^*$$

- Output $\widehat{f}(x) = F^{(M)}(x)$.

The hyperparameter $\eta \in (0, 1)$ is called 'learning rate' or 'shrinkage parameter' to balance the bias and variance. By default we take the base $\mathcal{G}$ as the set of stumps (4 parameters to be fitted).

Unstable Model: Decision Tree



- Make predictions at the terminal nodes (leafs)
- Interior node = a condition for a **single** feature $X_j < \theta$
- Condition TRUE $\rightarrow$ left child
- Condition FALSE $\rightarrow$ right child
- Decision stump is a one-split decision tree

Random Forest

- Resample B bootstrap datasets of the same size n from the raw dataset $\{Y_t, X_t : t = 1, \ldots, n\}$
- Generate individual trees using CART algorithm:

$$f^{(b)}(x; \{\widehat{R}_\tau^{(b)}, \widehat{c}_\tau^{(b)}\}) = \sum_{\tau=1}^{\mathrm{T}} \widehat{c}_\tau^{(b)} \mathbb{1}[x \in \widehat{R}_\tau^{(b)}]$$

  For each split, we compare a large pool of candidate features and thresholds, and then selects the one minimizes the sample variance allowing making a prediction on the left child and another on the right child.
- Average over trees to get the final estimator

$$\widehat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} f^{(b)}(x; \{\widehat{R}_\tau^{(b)}, \widehat{c}_\tau^{(b)}\})$$

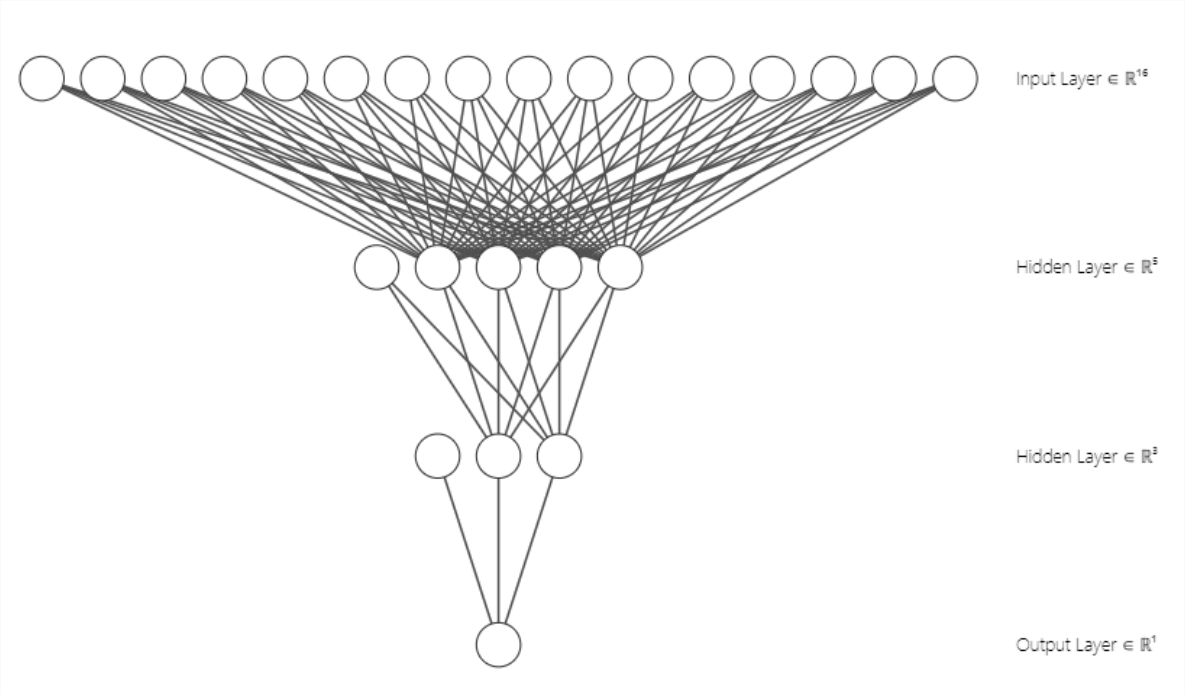
- RF uses a random subset of candidate features for each split

ReLU Neural Network

Starting from $z^{(0)} = x$ and for $\ell = 1, \ldots, D$

$$a^{(\ell)} = W^{(\ell)} z^{(\ell)} + b^{(\ell)} \in \mathbb{R}^{M_\ell}, \quad z^{(\ell)} = \sigma_\ell(a^{(\ell)}),$$

where we apply $\sigma_\ell : \mathbb{R} \to \mathbb{R}$ element-wisely with $\sigma_\ell(a) = \max\{0, a\}$ for $\ell < D$ but $\sigma_D(x) = x$. Fit parameters by least-squares method.



Example: $M_0 = 15$, $M_1 = 4$, $M_2 = 2$, $M_3 = 1$, $D = 3$

CSV2022: One-Day-Ahead Prediction

- logHAR outperforms HAR
- With Only HAR terms: no (substantial) improvement with ML relative to logHAR.
- With all variables: no (substantial) improvement with ML relative to logHAR.

CSV2002: One-Month-Ahead Prediction

With all variables:

Table 7 One-month-ahead relative MSE and Diebold–Mariano test for dataset $\mathcal{M}_{\text{ALL}}$

	HAR	HAR-X	LogHAR	LevHAR	SHAR	HARQ	RR	LA	EN	A-LA	P-LA	BG	RF	GB	NN_1^1	NN_1^{10}	NN_2^1	NN_2^{10}	NN_3^1	NN_3^{10}	NN_4^1	NN_4^{10}
LogHAR	1.169	1.402	–	1.438	1.430	1.672	1.212	1.470	1.203	1.476	1.459	***0.703***	***0.690***	***0.938***	***0.913***	***0.905***	***0.941***	***0.906***	***0.894***	***0.903***	1.013	1.072
BG	1.693	2.149	1.512	2.211	2.179	2.618	1.855	2.296	1.842	2.311	2.276	–	0.988	1.337	1.326	1.305	1.340	1.302	1.305	1.315	1.495	1.601

- All non-parametric estimators (bagging, RF, Boosting, NN) tend to outperform the linear estimators (including Ridge/LASSO/ElasticNet)
- The non-parametric estimators are all comparable: not surprising as they are estimating the same population regression function
- Make sense as the linear models are likely to be mis-specified, and the ML estimators benefit from the non-linearity

Estimating Large-Dimensional Integrated Covariance Matrix

Inconsistency of RCV Estimator

Recall the univariate RV estimator

$$RV_{n,t} = \sum_{i=1}^{n} r_{t,i}^2$$

The high-dimensional analogy for p-asset is

$$\Sigma_p^{\text{RCV}} = \sum_{i=1}^{n} r_{t,i} r_{t,i}', \quad r_{t,i} \in \mathbb{R}^p$$

Zheng and Li (2011): Even in very favorable cases $dX_t = \gamma_t dW_t$, $X_t \in \mathbb{R}^p$, $\gamma_t \in \mathbb{R}$ and W_t is a standard p-dimensional Brownian motion with independent components, Σ_p^{RCV} is not consistent when $p \asymp n$.

- Lam, Feng and Hu (2017) provides a non-linear shrinkage estimator; see also Lam and Qian (Working paper)

	$p=73$
NER-TSRVM	4.0
TSRVM	141.7
NER-MSRVM	4.0
MSRVM	137.5
NER-mMSRVM	4.2
mMSRVM	157.8
NER-KRPVM	4.0
KRPVM	137.5
NER-KRVM	3.9
KRVM	140.6
NER-PRPVM	3.9
PRPVM	125.2
NER-PRVM	3.8
PRVM	128.1

Lam and Qian (Working paper) proposed a non-parametric eigenvalue regularization (NER) method for different RCV estimators. Here shows the annualized out-of-sample standard deviation of the minimum variance portfolio constructed using 15-minute data on 73 US stocks.

Exercises

1. Consider the mean-zero OU process $dX_t = -\alpha X_t dt + \sigma dW_t$. Use Itô's formula to show that $Y_t = e^{\alpha t} X_t$ satisfies $dY_t = e^{\alpha t} \sigma dW_t$, and use this to obtain the solution $X_t = e^{-\alpha t} X_0 + \int_0^t e^{-\alpha(t-s)} \sigma dW_s$.
2. Suppose that we calculate $RV_{n,t}$ for a deterministic, continuous and bounded volatility process σ_t in $X_t = X_0 + \int_0^t \sigma_s dW_s$. Prove, for this case, consistency and asymptotic normality of $RV_{n,t}$.
3. Let $Y_t = X_t + U_t$, where $X_t = X_0 + \sigma W_t$ (constant volatility), and $U_t \sim \text{i.i.d. } N(0, \omega^2)$, independent of X_t. Derive the autocovariance function $\{\gamma_k\}_{k \geq 0}$ of the intra-day returns $r_{t,i} = Y_{t-1+i/n} - Y_{t-1+(i-1)/n}$, and show that it implies that $\{r_{t,i}\}$ is an MA(1) process with negative MA parameter. Next, show that $\gamma_0 + 2\gamma_1 = \sigma^2 \delta$.