Outline 1

1. High-Frequency Data
   1.1 Introduction
      1.2 Handling High-Frequency Data

2. Quadratic Variation & Realized Variance
   2.1 Quadratic Variation
   2.2 Realized Variance

3. Modelling Realized Variances
   3.1 Empirical Properties of RV
   3.2 Modelling RV
Once upon a time...
...and today

Source: NANEX website
High-Speed on Markets

- Dramatic changes of the trading landscape through the last decade
  - Algorithmic Trading
  - High-Frequency Trading
  - Dark Trading
  - New Types of Trading Platforms
  - Increased Competition between Exchanges & Market Fragmentation

- Consequences unclear ...
Order Driven Markets

- Traders directly enter orders in an (electronic) order book.
- Type of orders:
  - Market orders: Immediate execution.
  - Limit orders: Only execution when a certain price limit is achieved. Otherwise queueing in the book according to price/time priority rules.
- Order attributes: "all or nothing", "fill and kill", "day only", "iceberg" etc.
- A market order can be matched against several limit orders.
Financial Transaction Data

- Financial transaction data provide details on each individual transaction on a market:
  - Exact time stamp
  - Transaction price
  - Transaction volume
  - Best available bid and ask quote
  - Volumes associated with bid and ask quotes
  - Order attributes

- The informational limiting case: Limit order book data
  - Data on the complete order arrival process
  ⇒ Allows for the reconstruction of the complete limit order book
Bid/ask queues in a limit order book
Outline I

1. High-Frequency Data
   1.1 Introduction
   1.2 Handling High-Frequency Data

2. Quadratic Variation & Realized Variance
   2.1 Quadratic Variation
   2.2 Realized Variance

3. Modelling Realized Variances
   3.1 Empirical Properties of RV
   3.2 Modelling RV
**TAQ trade record for MSFT, June 1st, 2009**

<table>
<thead>
<tr>
<th>SYMBOL</th>
<th>DATE</th>
<th>TIME</th>
<th>EX</th>
<th>PRICE</th>
<th>SIZE</th>
<th>COND</th>
<th>CORR</th>
<th>G127</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>36601</td>
<td>Z</td>
<td>21.1900</td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>36601</td>
<td>Z</td>
<td>21.1900</td>
<td>1000</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>36601</td>
<td>Z</td>
<td>21.1900</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>36601</td>
<td>B</td>
<td>21.1900</td>
<td>400</td>
<td>@F</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>36601</td>
<td>B</td>
<td>21.1900</td>
<td>400</td>
<td>@F</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>36602</td>
<td>D</td>
<td>21.1912</td>
<td>470</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>36602</td>
<td>Z</td>
<td>21.1900</td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>36602</td>
<td>Q</td>
<td>21.1900</td>
<td>900</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>36602</td>
<td>Q</td>
<td>21.1900</td>
<td>100</td>
<td>@F</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>36602</td>
<td>Q</td>
<td>21.1900</td>
<td>100</td>
<td>@F</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>36602</td>
<td>Q</td>
<td>21.1900</td>
<td>300</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>36602</td>
<td>Q</td>
<td>21.1900</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>36602</td>
<td>D</td>
<td>21.1900</td>
<td>100</td>
<td>@F</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
## TAQ quote record for MSFT, June 1st, 2009

<table>
<thead>
<tr>
<th>SYMBOL</th>
<th>DATE</th>
<th>EX</th>
<th>TIME</th>
<th>BID</th>
<th>BID SZ</th>
<th>OFFER</th>
<th>OFF SZ</th>
<th>MODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>Z</td>
<td>36001</td>
<td>21.1100</td>
<td>43</td>
<td>21.1300</td>
<td>38</td>
<td>12</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>T</td>
<td>36001</td>
<td>21.1100</td>
<td>97</td>
<td>21.1200</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>T</td>
<td>36001</td>
<td>21.1100</td>
<td>92</td>
<td>21.1200</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>T</td>
<td>36001</td>
<td>21.1100</td>
<td>82</td>
<td>21.1200</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>I</td>
<td>36001</td>
<td>21.1100</td>
<td>9</td>
<td>21.1200</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>T</td>
<td>36001</td>
<td>21.1100</td>
<td>72</td>
<td>21.1200</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>B</td>
<td>36001</td>
<td>21.1100</td>
<td>30</td>
<td>21.1300</td>
<td>22</td>
<td>12</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>D</td>
<td>36001</td>
<td>21.1000</td>
<td>8</td>
<td>21.2100</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>B</td>
<td>36001</td>
<td>21.1100</td>
<td>31</td>
<td>21.1300</td>
<td>22</td>
<td>12</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>B</td>
<td>36002</td>
<td>21.1100</td>
<td>30</td>
<td>21.1300</td>
<td>22</td>
<td>12</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>T</td>
<td>36002</td>
<td>21.1100</td>
<td>72</td>
<td>21.1200</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>MSFT</td>
<td>2009-06-01</td>
<td>I</td>
<td>36002</td>
<td>21.1100</td>
<td>9</td>
<td>21.1300</td>
<td>33</td>
<td>12</td>
</tr>
</tbody>
</table>
NASDAQ ITCH data ...

<table>
<thead>
<tr>
<th>TIME</th>
<th>ORDER ID</th>
<th>EVENT</th>
<th>PRICE</th>
<th>SIZE</th>
<th>SIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>40900995</td>
<td>135132726</td>
<td>E</td>
<td></td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>40900995</td>
<td>135133117</td>
<td>E</td>
<td></td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>40900996</td>
<td>135126512</td>
<td>D</td>
<td></td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>40900996</td>
<td>135135501</td>
<td>A</td>
<td>2428</td>
<td>100</td>
<td>-1</td>
</tr>
<tr>
<td>40900996</td>
<td>135125636</td>
<td>D</td>
<td></td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>40900996</td>
<td>132601833</td>
<td>P</td>
<td>2427</td>
<td>500</td>
<td>-1</td>
</tr>
<tr>
<td>40900996</td>
<td>132601833</td>
<td>P</td>
<td>2427</td>
<td>250</td>
<td>-1</td>
</tr>
<tr>
<td>40900996</td>
<td>132601833</td>
<td>P</td>
<td>2427</td>
<td>144</td>
<td>-1</td>
</tr>
<tr>
<td>40900997</td>
<td>135135542</td>
<td>A</td>
<td>2427</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>40900997</td>
<td>135135544</td>
<td>A</td>
<td>2428</td>
<td>200</td>
<td>-1</td>
</tr>
<tr>
<td>40900997</td>
<td>135135580</td>
<td>A</td>
<td>2426</td>
<td>200</td>
<td>1</td>
</tr>
<tr>
<td>40900998</td>
<td>135135501</td>
<td>D</td>
<td></td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>40900998</td>
<td>135135591</td>
<td>A</td>
<td>2432</td>
<td>100</td>
<td>-1</td>
</tr>
<tr>
<td>40900999</td>
<td>135135631</td>
<td>A</td>
<td>2428</td>
<td>4000</td>
<td>-1</td>
</tr>
</tbody>
</table>
Matching Trades and Quotes

- Many exchanges record trades and quotes in different files ⇒ requires data matching

- **Problem**: Time stamps often not sufficiently exact!

- Researchers have to apply matching rules:
  - search for corresponding order book updates in windows around trade time stamp ("perfect" matching)
  - account for the possibility of hidden volume ("imperfect" matching)
  - "round time matching": match with order book record closest to trade’s time stamp
Data Errors

- Typical data errors are due to
  (i) wrong recording,
  (ii) delayed recording of trade or quote information

- Delayed records are due to ex post corrections of mis-recorded trades

- Can cause severe price jumps
Recording Errors ...
Identification of Buys and Sells

- When trades and quotes come from different files, buyer- and seller-initiated trades cannot be identified directly.

- Quote method: Comparing trade price and prevailing mid-quote
  - Prices above (below) the mid-quote are classified as buys (sells).

- Tick test: Comparing to recent trade history
  - If current trade occurs at higher (lower) price than previous trade, it is classified as buy (sell).
Data Sampling

- Transaction time sampling (TRTS): sampling whenever a trade occurs
- Tick time sampling (TTS): sampling whenever price or midquote changes by at least one tick
  - TRTS and TTS create irregularly spaced observations
  - In case of multivariate processes: Creates asynchronicity of observations
- Calendar time sampling (CTS): sampling in equi-distant time intervals
  - Induces loss of information but eases modeling due to equi-distant data
Outline I

1. High-Frequency Data
   1.1 Introduction
   1.2 Handling High-Frequency Data

2. Quadratic Variation & Realized Variance
   2.1 Quadratic Variation
   2.2 Realized Variance

3. Modelling Realized Variances
   3.1 Empirical Properties of RV
   3.2 Modelling RV
Diffusion Processes

An Itô drift-diffusion process $X_t$ is given by the stochastic differential equation (SDE)

$$dX_t = a_t \, dt + \sigma_t \, dW_t,$$

where $a_t$ denotes the drift, $\sigma_t$ is the volatility and $W_t$ is a Brownian motion.

Equivalent representation:

$$X_T - X_0 = \int_0^T a_t \, dt + \int_0^T \sigma_t \, dW_t.$$

Approximation for small increments $\Delta$ by assuming the drift and volatility to be constant:

$$X_{t+\Delta} - X_t | X_t \sim N(a_t \Delta, \sigma_t^2 \Delta).$$
Special Case

- Bachelier (1900).
- SDE with constant drift and volatility:

  \[ dX_t = adt + \sigma dW_t \]

  and thus

  \[ X_t - X_0 = at + \sigma W_t. \]

- Then:

  \[ X_t - X_s \sim N(a(t - s), \sigma^2(t - s)), \quad t > s. \]
Consider a single risky asset whose price may be observed at \( n \) equally-spaced discrete points, namely 
\[ t = 0, 1/n, 2/n, ..., 1 - (1/n), 1. \]

We denote the logarithmic asset price at time \( t \) by \( X_t \) and the continuously compounded returns over \( [t - k, t] \) by 
\[ r(t, k) := X_t - X_{t-k}, \]
where \( 0 \leq t - k < t \leq 1 \) and \( k = j/n \) for some positive integer \( j \).

Assume the continuously compounded returns are driven by a simple time-invariant Brownian motion,

\[ dX_t = adt + \sigma dW_t, \]
\[ 0 \leq t \leq 1, \]

where \( a \) and \( \sigma \) denote the constant drift and diffusion coefficients.
For a given measurement period, $[0, T]$, we have $nT$ intraday returns $r(t, 1/n) = X_t - X_{t-1/n}$, $t = 1/n, \ldots, (n - 1)T/n$, $T$, that are i.i.d. normally distributed with mean $a/n$ and variance $\sigma^2/n$.

The drift $a$ is estimated by ML by

$$\hat{a}_n = \frac{1}{T} \sum_{j=1}^{nT} r(j/n, 1/n) = \frac{r(T, T)}{T} = \frac{X_T - X_0}{T}$$

with

$$\mathbb{V}[\hat{a}_n] = \frac{\sigma^2}{T}.$$

The estimator of the drift is independent of the sampling frequency. Hence, the mean drift cannot be estimated consistently over any fixed interval.
2. Quadratic Variation & Realized Variance  |  2.1 Quadratic Variation  |  24  |  53

**Estimating $\sigma^2$**

- We have

$$\mathbb{E}[r(j/n, 1/n)^2] = \frac{a^2}{n^2} + \frac{\sigma^2}{n},$$

and

$$\mathbb{E}[r(j/n, 1/n)^4] = \frac{a^4}{n^4} + 6 \frac{a^2\sigma^2}{n^3} + 3 \frac{\sigma^4}{n^2}.$$

⇒ Terms involving drift coefficient are an order of magnitude smaller, for $n$ large!

- Allows to estimate return variation with high degree of precision even without specifying the mean drift, i.e.,

$$\widehat{\sigma}_n^2 = \frac{1}{T} \sum_{j=1}^{nT} r^2(j/n, 1/n).$$
We can show that

\[ \mathbb{E}[\hat{\sigma}^2_n] = \frac{a^2}{n} + \sigma^2 \]

and

\[ \mathbb{V}[\sigma^2_n] = 4 \frac{a^2 \sigma^2}{n^2 T} + 2 \frac{\sigma^4}{nT} \]

and thus \( \hat{\sigma}^2_n \to \sigma^2 \) for \( n \to \infty \).

Hence, the realized variation measure is a biased but consistent estimator of the underlying (squared) volatility coefficient.

For \( n \to \infty \), the bias is negligible and we have

\[ \sqrt{nT}(\hat{\sigma}^2_n - \sigma^2) \to^d \mathcal{N}(0, 2\sigma^4). \]
2. Quadratic Variation & Realized Variance

2.1 Quadratic Variation

Integrated Variation

Assume now the diffusion

\[ dX_t = a_t \, dt + \sigma_t \, dW_t, \quad 0 \leq t \leq T, \]

where \( a_t \) and \( \sigma_t \) are predictable processes, \( a_t \) is of finite variation, and \( \sigma_t \) is strictly positive and square integrable, i.e.,

\[ \mathbb{E} \left[ \int_0^t \sigma_s^2 \, ds \right] < \infty. \]

The continuously compounded return over the time interval from \( t - k \) to \( t \), \( 0 < k \leq t \) is therefore

\[ r(t, k) = X_t - X_{t-k} = \int_{t-k}^t a_\tau \, d\tau + \int_{t-k}^t \sigma_\tau \, dW_\tau. \]
The quadratic variation $QV(t, t-k)$ is given by

$$QV(t-k, t) := \lim_{n \to \infty} \sum_{i=1}^{n} r \left( t - k + \frac{i}{n}, \frac{1}{n} \right)^2.$$ 

The diffusive sample path variation over $[t-k, t]$ is called the integrated variance $IV(t-k, k)$,

$$IV(t-k, t) := \int_{t-k}^{t} \sigma_{\tau}^2 d\tau.$$ 

As long as there is only a diffusive component but no jumps:

$$QV(t-k, t) = IV(t-k, t).$$ 

Recall that innovations to the mean component $a_t$ do not affect the sample path variation of the return!
Outline 1

1. High-Frequency Data
   1.1 Introduction
   1.2 Handling High-Frequency Data

2. Quadratic Variation & Realized Variance
   2.1 Quadratic Variation
   2.2 Realized Variance

3. Modelling Realized Variances
   3.1 Empirical Properties of RV
   3.2 Modelling RV
Realized Variance

- Interval of interest: $[0, T]$. Here, consider $T = 1$.
- How to measure the integrated variance

$$IV(0, 1) := IV_1 := IV = \int_0^1 \sigma^2 d\tau$$

- Intuitive: Sum of $n$ squared returns of length $\Delta_n = [T/n] = 1/n$.

$$RV^{(n)} := \sum_{j=1}^{n} (X_{j\Delta} - X_{(j-1)\Delta})^2 =: \sum_{j=1}^{n} r_{j,n}^2.$$ 

- $RV$ is called the *realized variance*. 
RV Estimation in a Discrete-Time Setting

Assume that intraday (log) returns follow a stochastic volatility process

\[ r_{j,n} = \sigma_{j,n} u_{j,n}, \quad u_{j,n} \sim \mathcal{N}(0,1), \quad j = 1, \ldots, n, \]

where \( \sigma_{j,n}^2 \) denotes the stochastic intraday variance and \( u_{j,n} \) is independent of \( \sigma_{j,n} \).

Then, we have:

\[
\nabla \left[ \sum_{j=1}^{n} r_{j,n} \right] = \mathbb{E} \left[ \sum_{j=1}^{n} \sigma_{j,n}^2 \right] = \mathbb{E}[RV^{(n)}].
\]
Conditional variance of RV, given $\sigma^2_{j,n}$, $j = 1, \ldots, n$:

$$\nabla[RV^{(n)}|\sigma^2_{j,n}, j = 1, \ldots, n] = 2 \sum_{j=1}^{n} \sigma^4_{j,n}.$$ 

If $\sigma_{j,n}$, $j = 1, \ldots, n$, are independent, then,

$$\nabla[RV^{(n)}] = 2n \mathbb{E}[\sigma^4_{j,n}].$$

Moreover, if $\sigma^2_{j,n} = \sigma^2 / n$, with $\sigma^2$ denoting a random variable, then,

$$\nabla[RV^{(n)}] = \frac{2}{n} \mathbb{E}[\sigma^4].$$
2. Quadratic Variation & Realized Variance

2.2 Realized Variance

ML Estimation of $\sigma^2$

- Assume diffusion without drift and constant volatility:

$$dX_t = \sigma \, dW_t.$$  

- The log returns $r_{i,n} = X_{i\Delta_n} - X_{(i-1)\Delta_n} = \sigma (W_{i\Delta_n} - W_{(i-1)\Delta_n})$, $i = 1 \ldots, n$, $n = T/\Delta_n$ are then i.i.d. $N(0, \sigma^2 \Delta_n)$, so the likelihood function is

$$\ell(\sigma^2) = -n \ln(2\pi \sigma^2 \Delta_n)/2 - (2\sigma^2 \Delta_n)^{-1} \mathbf{r}_n' \mathbf{r}_n,$$

with $\mathbf{r}_n := (r_{1,n}, \ldots, r_{n,n})'$. 
The MLE of $\sigma^2$ is given by

$$\hat{\sigma}^2 = \frac{1}{n\Delta_n} \sum_{i=1}^{n} r_{i,n}^2 = \frac{1}{T} \sum_{i=1}^{n} r_{i,n}^2$$

with

$$\mathbb{E}[\hat{\sigma}^2] = \frac{1}{T} \sum_{i=1}^{n} \mathbb{E}[r_{i,n}^2] = \frac{n\sigma^2 \Delta_n}{T} = \sigma^2$$

$$\mathbb{V}[\hat{\sigma}^2] = \frac{1}{T^2} \mathbb{V} \left[ \sum_{i=1}^{n} r_{i,n}^2 \right] = \frac{n2\sigma^4 \Delta_n^2}{T^2} = \frac{2\sigma^4 \Delta_n}{T}$$
We have the following asymptotic distribution

\[ T^{1/2}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{T \to \infty} N(0, \varsigma), \]

where

\[ \varsigma = \mathbb{A} \mathbb{V}[\hat{\sigma}^2] = 2\sigma^4 \Delta_n. \]

Asymptotic variance declines with decreasing \( \Delta_n \)!

'In-fill' asymptotics: Sampling on highest possible frequencies crucial!
Theory of Quadratic Variation

Assume now

\[ dX_t = \sigma_t dW_t \]

It can be shown that (e.g., Barndorff-Nielsen/Shepard, 2002)

\[ \sqrt{n}(RV^{(n)} - IV) | IQ \xrightarrow{d} N(0, 2IQ) \]

The integrated quarticity \( IQ \) can be consistently estimated using the realized quarticity:

\[ RQ^{(n)} := \frac{n}{3} \sum_{j=1}^{n} r_{j,n}^4 \xrightarrow{p} IQ := \int_0^1 \sigma^4(\tau)d\tau \]

Then:

\[ \sqrt{n} \frac{RV^{(n)} - IV}{\sqrt{2RQ^{(n)}}} \approx N(0, 1). \]
In-Fill Asymptotics at Work ...
Outline I

1. High-Frequency Data
   1.1 Introduction
   1.2 Handling High-Frequency Data

2. Quadratic Variation & Realized Variance
   2.1 Quadratic Variation
   2.2 Realized Variance

3. Modelling Realized Variances
   3.1 Empirical Properties of RV
   3.2 Modelling RV
Note: RV is a (non-parametric) ex-post estimator of the integrated variance $\int_{t-1}^{t} \sigma^2(s)ds$.

In order to predict $\int_{t}^{t+1} \sigma^2(s)ds$, we need a model for $RV_t$, $t = 1, \ldots$.

⇒ Need to understand the statistical and empirical properties of $RV_t$, $t = 1, \ldots$.
  ▶ Distributional properties
  ▶ Dynamic properties
Unconditional Distribution

- The unconditional distribution of realized volatility is approximately log-normal.
The Distribution of $r/RV^{1/2}$

Empirical ACFs of Realized Volatility

Source: Andersen et al (2001, JASA)
Implications

- The distribution of realized volatility is approximately log-normal.
- Realized volatility is fractionally integrated.
- RV-standardized returns are approximately normal.
- Distribution of returns is approximately lognormal-normal as advocated in Taylor’s (1986) SV model!
Outline I

1. High-Frequency Data
   1.1 Introduction
   1.2 Handling High-Frequency Data

2. Quadratic Variation & Realized Variance
   2.1 Quadratic Variation
   2.2 Realized Variance

3. Modelling Realized Variances
   3.1 Empirical Properties of RV
   3.2 Modelling RV
The Heterogeneous AR (HAR) model

- Motivation: Heterogeneous Market Hypothesis (Müller et al. 1993): Main heterogeneity due to differences in agents’ time horizons ⇒ agents perceive, react and cause different volatility components.

- Consider $RV$, aggregated as follows:

$$RV_{t-1|t-k} = \frac{1}{k} \sum_{j=1}^{k} RV_{t-j}.$$ 

- Choose 3 different time horizons: daily $k = 1$, weekly $k = 5$, monthly $k = 22$. 

Then, the HAR model (Corsi, 2009) is given by

\[ r_t = \tilde{\sigma}_t^d \varepsilon_t, \quad \varepsilon_t \sim (0, 1) \]

\[ \tilde{\sigma}_{t+1}^m = c^m + \phi^m RV_{t|t-21} + \tilde{u}_{t+1}^m \]

\[ \tilde{\sigma}_{t+1}^w = c^w + \phi^w RV_{t|t-4} + \gamma^w \mathbb{E}_t[\tilde{\sigma}_{t+1}^m] + \tilde{u}_{t+1}^w \]

\[ \tilde{\sigma}_{t+1}^d = c^d + \phi^d RV_t + \gamma^d \mathbb{E}_t[\tilde{\sigma}_{t+1}^w] + \tilde{u}_{t+1}^d \]

Interpretation: Expectation for the next period volatility is based on

- current RV experienced at same time scale,
- expectation for next longer horizon partial volatility.
By recursive substitution we obtain

$$\sigma_{t+1}^{d} = c + \beta^{d} RV_{t} + \beta^{w} RV_{t|t-4} + \beta^{m} RV_{t|t-21} + \tilde{u}_{t+1}^{d}$$

With

$$\sigma_{t+1}^{d} = RV_{t+1}^{d} + u_{t+1}^{d},$$

where $u_{t+1}^{d}$ subsumes estimation errors, we get

$$RV_{t+1} = c + \beta^{d} RV_{t} + \beta^{w} RV_{t|t-4} + \beta^{m} RV_{t-1|t-21} + u_{t+1},$$

with $u_{t+1} = \tilde{u}_{t+1}^{d} - u_{t+1}^{d}$.

⇒ HAR(3) model
HARQ model (Bollerslev et al, 2016)

- $RV_t$ is a noisy estimate of $IV_t = \int_{t-1}^{t} \sigma_s^2 ds$:

  $$RV_t = IV_t + \eta_t, \quad \eta_t \sim MN(0, \sigma_\eta^2),$$

  where $\sigma_\eta^2 = 2\Delta_n IQ_t$, and $IQ_t = \int_{t-1}^{t} \sigma_s^4 ds$, which can be consistently estimated by $RQ = \frac{n}{3} \sum_{i=1}^{n} r_{i,n}^4$.

- Assume for illustration an AR(1) model for $IV_t$:

  $$IV_t = \phi_0 + \phi_1 IV_{t-1} + u_t, \quad u_t \sim (0, \sigma_u^2)$$

- If $u_t$ and $\eta_t$ are both i.i.d., then $RV_t$ follows an ARMA(1,1) process with AR parameter $\phi_1$ and MA parameter

  $$\theta_t = \frac{-\sigma_u^2 - (1 + \phi_1^2)\sigma_\eta^2 \sqrt{\sigma_u^4 + 2(1 + \phi_1^2)\sigma_u^2 \sigma_\eta^2 + (1 - \phi_1^2)^2 \sigma_\eta^4}}{2\phi_2 \sigma_\eta^2},$$

  which increases in $\sigma_\eta^2$ and $\theta_1 \to 0$ if $\sigma_\eta^2 \to 0$ or $\phi_1 \to 0$.  

Suppose researchers specify an (incorrect) AR(1) model for $RV_t$, then

$$IV_t + \eta_t = \beta_0 + \beta_1 (IV_{t-1} + \eta_{t-1}) + u_t.$$ 

If $u_t$ and $\eta_t$ are both i.i.d., then IQ is constant (homoscedasticity), $\text{Cov}[RV_t, RV_{t-1}] = \phi_1 \text{V}[IV_t]$ and $\text{V}[RV_t] = \text{V}[IV_t] + 2\Delta_nIQ$, and the population value for $\beta_1$ is

$$\beta_1 = \phi_1 \left(1 + \frac{2\Delta_nIQ}{\text{V}[IV_t]}\right)^{-1} < \phi_1.$$ 

Discrepancy between $\beta_1$ and $\phi_1$ is the so-called attenuation bias resulting from measurement errors.
Bollerslev, Patton and Quadvelig (2016, JoE) propose a more flexible specification:

\[ RV_t = \beta_0 + (\beta_1 + \beta_1 Q RQ_{t-1}^{1/2}) R V_{t-1} + u_t \]

The so-called HARQ model is then given by

\[ RV_t = \beta_0 + (\beta_1 + \beta_1 Q RQ_{t-1}^{1/2}) R V_{t-1} + (\beta_2 + \beta_2 Q RQ_{t-1|t-5}^{1/2}) R V_{t-1|t-5} \]

\[ + (\beta_3 + \beta_3 Q RQ_{t-1|t-22}^{1/2}) R V_{t-1|t-22} + u_t, \]

where \( RQ_{t-1|t-k} = \frac{1}{k} \sum_{j=1}^{k} RQ_{t-j}. \)

As the measurement error variance for the weekly and monthly RV are much smaller, a simplified version of the HARQ model is

\[ RV_t = \beta_0 + (\beta_1 + \beta_1 Q RQ_{t-1}^{1/2}) R V_{t-1} + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + u_t, \]
Realized GARCH

Hansen et al (2012) propose a Realized GARCH model of the form

\[ r_t = \sqrt{h_t} z_t \]
\[ h_t = \omega + \beta h_{t-1} + \gamma x_{t-1}, \]
\[ x_t = \xi_t + \phi h_t + \tau(z_t) + u_t, \]

where \( x_t \) is a realized variance measure, \( z_t \sim iid(0, 1) \), \( u_t \sim iid(0, \sigma_u^2) \), \( h_t = \mathbb{V}[r_t|\mathcal{F}_{t-1}] \), and \( \tau(z_t) \) is a (leverage) function of \( z_t \).

As a consequence, \( h_t \) follows AR(1) with

\[ h_t = \mu + \pi h_{t-1} + w_{t-1}, \]

where \( \mu = \omega + \gamma \xi \), \( \pi = \beta + \phi \gamma \), and \( w_t = \gamma \tau(z_t) + \gamma u_t \).
Second equation is referred to as *measurement* equation: Integrated variance is seen as conditional variance plus random innovation $\tau(z_t) + u_t$.

Simple specification for the leverage function:

$$\tau(z) = \tau_1 z + \tau_2(z^2 - 1).$$

If realized measure is computed from a shorter period (trading hours) than the interval that the conditional variance refers to (e.g., 24 hours), we have $\phi < 1$. 
Log Realized GARCH

A Realized GARCH with log-linear specification is given by

\[
\ln h_t = \omega + \sum_{i=1}^{p} \beta_i \ln h_{t-i} + \sum_{j=1}^{q} \gamma_j \ln x_{t-j} + \sum_{j=1}^{q} \alpha_j \ln r_{t-j}^2 \\
\ln x_t = \xi + \phi \ln h_t + \tau(z_t) + u_t,
\]

where \( q = \max_i \{ (\alpha_i, \gamma_i) \neq (0, 0) \} \), \( z_t = r_t / \sqrt{h_t} \sim iid(0, 1) \), \( u_t \sim iid(0, \sigma^2_u) \) and \( E[\tau(z_t) = 0] \).
Other or extended models

- HEAVY models (Shephard & Sheppard, 2010)
- Time-varying coefficient realized GARCH with measurement errors: Gerlach, Naimoli & Storti (2018)
- Multiplicative error models (Caporin et al, 2017, Engle/Gallo, 2006)
- ...