# Time Series Modeling of Financial Data

Prof. Daniel P. Palomar

The Hong Kong University of Science and Technology (HKUST)

MAFS6010R- Portfolio Optimization with R
MSc in Financial Mathematics
Fall 2019-20, HKUST, Hong Kong

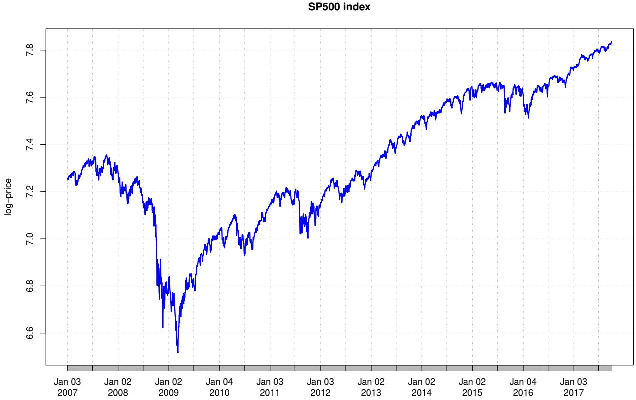
# Outline

# Outline

# Asset log-prices

- Let $p_t$ be the price of an asset at (discrete) time index $t$.
- The fundamental model is based on modeling the log-prices $y_t \triangleq \log p_t$ as a random walk:

$$y_t = \mu + y_{t-1} + \epsilon_t$$



# Asset returns

- Simple return (a.k.a. linear or net return) is

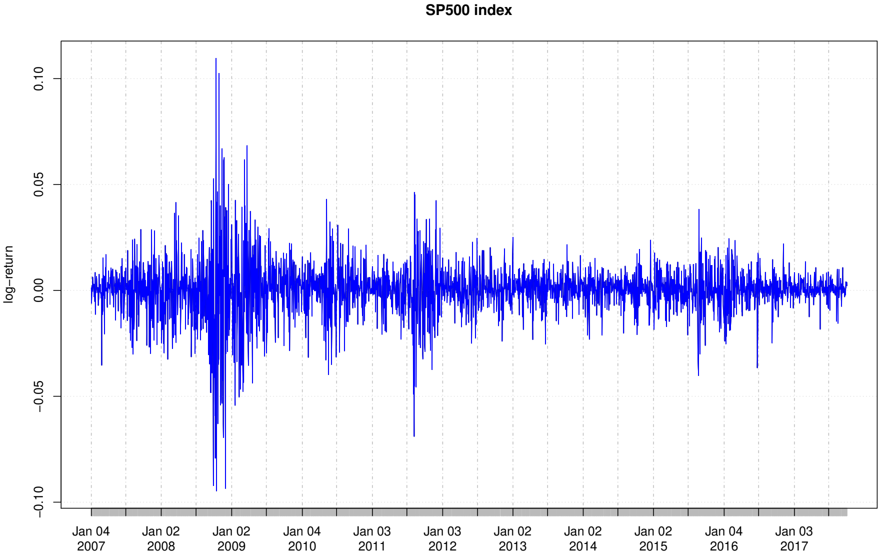
$$R_t \triangleq \frac{p_t - p_{t-1}}{p_{t-1}} = \frac{p_t}{p_{t-1}} - 1.$$

- Log-return (a.k.a. continuously compounded return) is

$$r_t \triangleq y_t - y_{t-1} = \log \frac{p_t}{p_{t-1}} = \log \left(1 + R_t\right).$$

- Note $r_t = \log\left(1 + R_t\right) \approx R_t$ when $R_t$ is small.

# S&P 500 index - log-returns



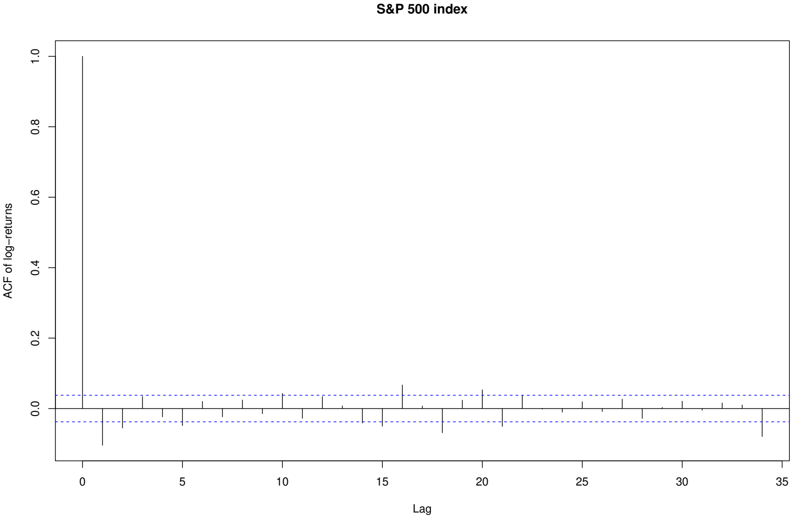
# Stylized facts of financial data

- A set of properties, common across many instruments, markets, and time periods, has been observed by independent studies and classified as "stylized facts."[1]
- **Lack of stationarity**: past returns do not necessarily reflect future performance (watch out fund's brochures)
- **Absence of autocorrelations**: autocorrelations of returns are often insignificant (efficient market hypothesis)
- **Heavy tails**: Gaussian distributions generally do not hold in financial data (even after correcting for volatility clustering)
- **Gain/loss asymmetry**: basically asymmetry of the pdf
- **Aggregational Gaussianity**: for lower frequencies, the distribution tends to become more Gaussian.
- **Volatility clustering**: high-volatility evens tend to cluster in time

[1] R. Cont, "Empirical properties of assets returns: Stylised facts and statistical issues", *Quantitative Finance*, vol. 1, pp. 223–236, 2001.

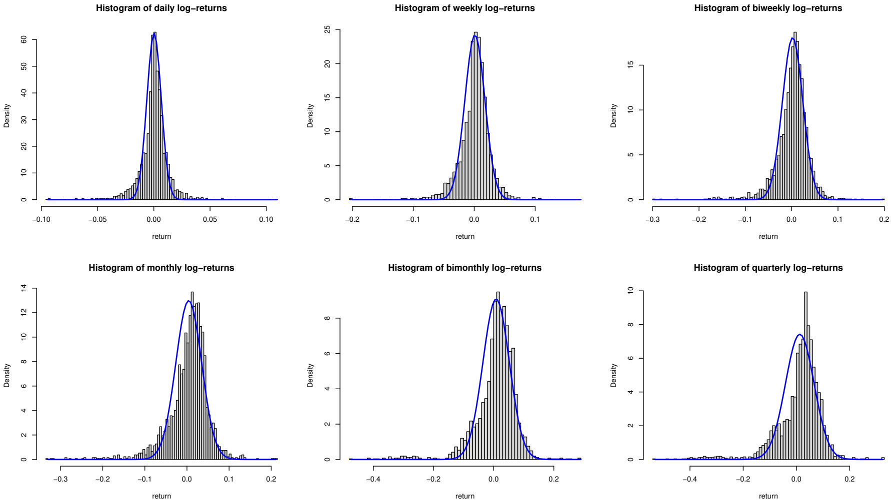
# Autocorrelation

- ACF of S&P 500 log-returns:



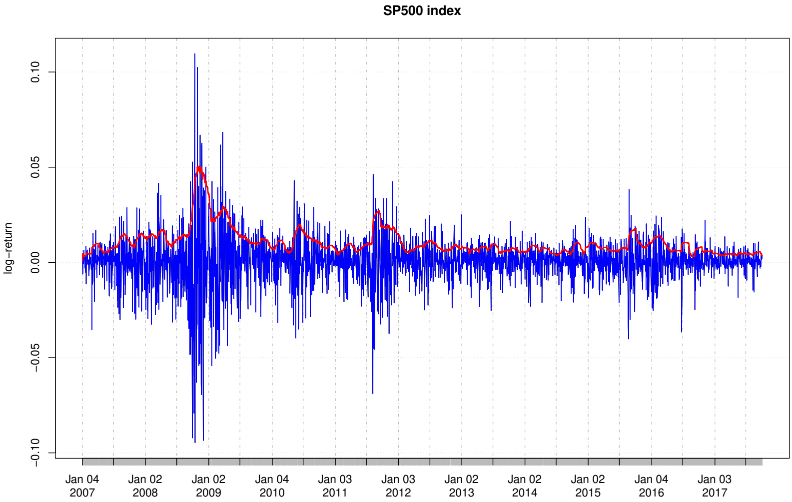
# Non-Gaussianity and asymmetry

- Histograms of S&P 500 log-returns:



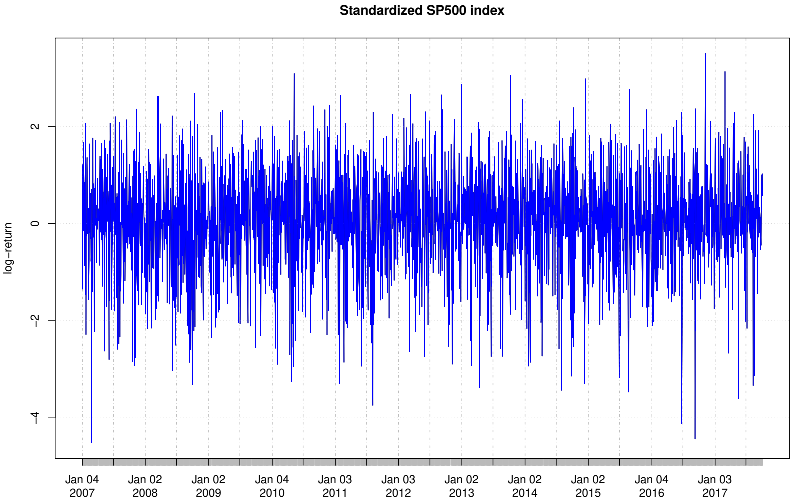
# Volatility clustering

- S&P 500 log-returns:



# Volatility clustering removed

- Standardized S&P 500 log-returns:



# Outline

# General structure of a model

- Denote log-return of $N$ assets as $\mathbf{r}_t \in \mathbb{R}^N$.
- Denote $\mathcal{F}_{t-1}$ as the previous historical data.
- Financial modeling aims at modeling $\mathbf{r}_t$ conditional on $\mathcal{F}_{t-1}$.

- Conditional on $\mathcal{F}_{t-1}$, we can decompose $\mathbf{r}_t \in \mathbb{R}^N$ as follows:

$$\mathbf{r}_t = \boldsymbol{\mu}_t + \mathbf{w}_t$$

  where

  - $\boldsymbol{\mu}_t$ is the *conditional mean*

$$\boldsymbol{\mu}_t = \mathsf{E}[\mathbf{r}_t | \mathcal{F}_{t-1}]$$

  - $\mathbf{w}_t$ is a white noise with zero mean and *conditional covariance*

$$\boldsymbol{\Sigma}_t = \mathsf{E}[(\mathbf{r}_t - \boldsymbol{\mu}_t)(\mathbf{r}_t - \boldsymbol{\mu}_t)^T | \mathcal{F}_{t-1}].$$

# i.i.d. model

- It assumes $\mathbf{r}_t$ follows an i.i.d. distribution.
- That is, both the conditional mean and conditional covariance are constant

$$\boldsymbol{\mu}_t = \boldsymbol{\mu},$$
$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_w.$$

- Very simple model, however, it is one of the most fundamental assumptions for many important works, e.g., the Nobel prize-winning Markowitz portfolio theory[2].

[2] H. Markowitz, "Portfolio selection", *J. Financ.*, vol. 7, no. 1, pp. 77–91, 1952.

# Factor model

- The factor model is
$$\mathbf{r}_t = \boldsymbol{\alpha} + \mathbf{B}\mathbf{f}_t + \mathbf{w}_t,$$
where
  - $\boldsymbol{\alpha}$ denotes a constant vector
  - $\mathbf{f}_t \in \mathbb{R}^K$ with $K \ll N$ is a vector of a few factors that are responsible for most of the randomness in the market,
  - $\mathbf{B} \in \mathbb{R}^{N \times K}$ denotes how the low dimensional factors affect the higher dimensional market;
  - $\mathbf{w}_t$ is a white noise residual vector that has only a marginal effect.
- The factors can be explicit or implicit.
- Widely used by practitioners (they buy factors at a high premium).
- Connections with Principal Component Analysis (PCA)[3].

[3] I. Jolliffe, *Principal Component Analysis*. Springer-Verlag, 2002.

# i.i.d. vs factor models

- Factor models are special cases of the i.i.d. model with the variation being decomposed into two parts: **low dimensional factors** and **marginal noise**.
- The explicit factor model
  - explains the log-returns with a smaller number of fundamental or macroeconomic variables,
  - however, in general there is no systematic method to choose the right factors.
- The hidden factor model
  - explores the structure of the covariance matrix,
  - is a more systematical approach and thus it may provide a better explanatory power,
  - however, does not have explicit econometric interpretations.

# Outline

# Outline

# Time series decomposition

- Time series data can exhibit a variety of patterns, and it is often helpful to split a time series into several components, each representing an underlying pattern category.[4]
- Additive decomposition:

$$y_t = S_t + T_t + R_t$$

where $y_t$is the data, $S_t$ is the seasonal component, $T_t$ is the trend-cycle component, and $R_t$ is the remainder (noisy) component.

- Multiplicative decomposition:

$$y_t = S_t \times T_t \times R_t.$$

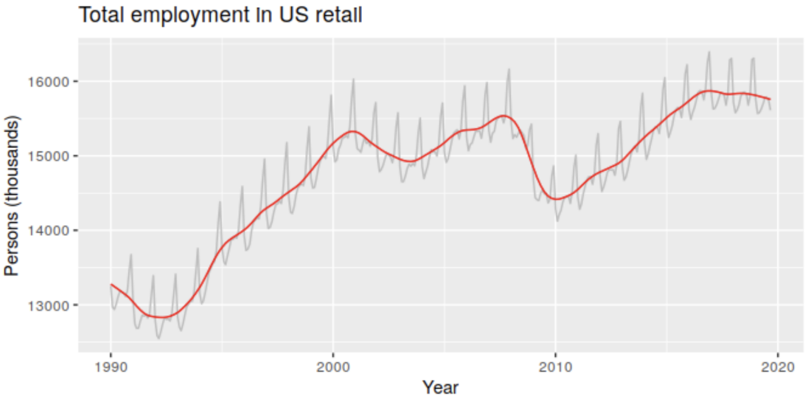
- Multiplicative decompositions are common with economic time series.
- An alternative to using a multiplicative decomposition is to first use a log transformation and then use an additive decomposition:

$$\log y_t = \log S_t + \log T_t + \log R_t.$$

[4] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2014.

# Time series decomposition: Example

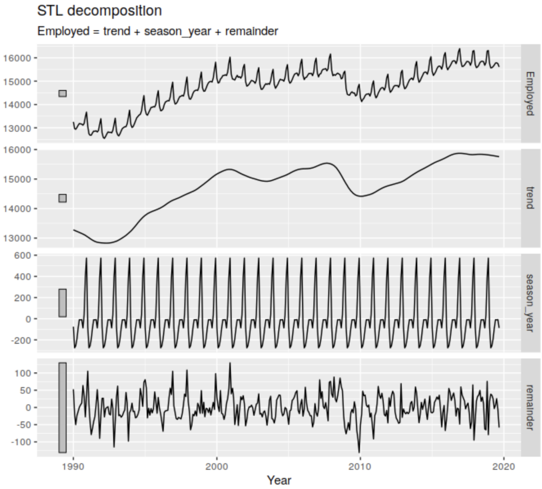
Time series (grey) with trend-cycle component (red):[5]




[5] Credit: Hyndman and Athanasopoulos at https://otexts.com/fpp3

# Time series decomposition: Example

Time series decomposition into trend-cycle component, seasonal component, and residual component:[6]




[6]Credit: Hyndman and Athanasopoulos at https://otexts.com/fpp3

# Moving average (MA) smoothing

- One classical way to obtain the trend-cycle component of a time series is with the moving average.
- A moving average of order $m$ is

$$\hat{y}_t = \frac{1}{m}\sum_{i=1}^{m} y_{t-i}.$$

- We can also use a centered moving average for smoothing (not forecasting purposes):

$$\hat{y}_t = \frac{1}{m}\sum_{i=-k}^{k} y_{t-i}$$

  where $m = 2k + 1$.
- It is also called rolling means since it is computing the mean on a rolling-window basis.

# Classical decomposition

- The classical decomposition method for $y_t = S_t + T_t + R_t$ originated in the 1920s.[7]
- It is a relatively simple procedure, and forms the starting point for most other methods of time series decomposition.
- Steps:
  1. Compute the trend-cycle component $\hat{T}_t$ using an MA.
  2. Detrend series: $y_t - \hat{T}_t$
  3. To estimate the seasonal component $\hat{S}_t$ for each season, simply average the detrended time series for that season; for example, with monthly data, the seasonal component for March is the average of all the detrended March values.
  4. Compute the remainder as $\hat{R}_t = y_t - \hat{T}_t - \hat{S}_t$.
- While classical decomposition is still widely used, it is not recommended, as there are now several much better methods.

[7] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2014.

# Exponential smoothing

- Exponential smoothing was proposed in the late 1950s, and has motivated some of the most successful forecasting methods.[8]
- Forecasts produced using exponential smoothing methods are weighted averages of past observations, with the weights decaying exponentially as the observations get older.
- The simplest of the exponentially smoothing methods is naturally called simple exponential smoothing:

$$\hat{y}_t = \alpha y_{t-1} + (1-\alpha)\hat{y}_{t-1}$$

  with $0 \leq \alpha \leq 1$.
- Recall the expression for the MA:

$$\hat{y}_t = \frac{1}{m}\sum_{i=1}^{m} y_{t-i}.$$

[8] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2014.

# Exponential smoothing: Component form

- The simple exponential smoothing

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t$$

can be rewritten in a different form called **component form**.

- Component form of the simple exponential smoothing:

$$\begin{array}{rrcl} \text{Forecast equation} & \hat{y}_{t+1} & = & \ell_t \\ \text{Smoothing equation} & \ell_t & = & \alpha y_t + (1-\alpha)\ell_{t-1} \end{array}$$

where $\ell_t$ is the level (or smoothed value) of the series at time $t$.

# Holt's linear trend method

- Holt extended in 1957 the simple exponential to allow a trend (slope of the level):

$$
\begin{aligned}
\text{Forecast equation} \qquad \hat{y}_{t+1} &= \ell_t + b_t \\
\text{Level equation} \qquad \ell_t &= \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
\text{Trend equation} \qquad b_t &= \beta(\ell_t - \ell_{t-1}) + (1-\beta)\, b_{t-1}
\end{aligned}
$$

  where $\ell_t$ is the level (or smoothed value) and $b_t$ denotes the trend (slope) of the series at time $t$.

# Holt-Winters' method with seasonality

- Holt and Winters extended Holt's method to capture seasonability in 1960:

$$
\begin{array}{rrcl}
\text{Forecast equation} & \hat{y}_{t+1} & = & \ell_t + b_t + s_{t+1-m} \\
\text{Level equation} & \ell_t & = & \alpha\left(y_t - s_{t-m}\right) + (1-\alpha)\left(\ell_{t-1} + b_{t-1}\right) \\
\text{Trend equation} & b_t & = & \beta\left(\ell_t - \ell_{t-1}\right) + (1-\beta)\, b_{t-1} \\
\text{Seasonal equation} & s_t & = & \gamma\left(y_t - \ell_{t-1} - b_{t-1}\right) + (1-\gamma)\, s_{t-m}
\end{array}
$$

  where $m$ denotes the period of the seasonality (so for monthly data $m = 12$, i.e., one year).

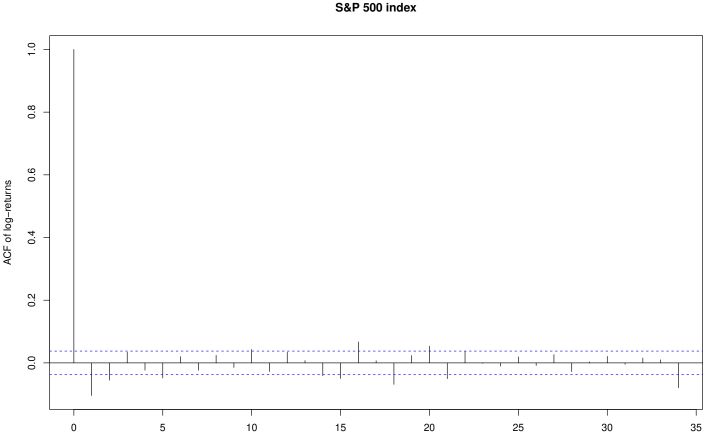
# Taxonomy of exponential smoothing methods

- More generally, while the level equation is always there, one can choose whether to have trend and seasonal terms and also one can choose whether they are additive or multiplicative.
- This can be expressed compactly with three letters (E,T,S) defining the error, trend, and seasonality type (to choose from None, Additive, and Multiplicative):
  - ETS(A,N,N) corresponds to the simple exponential smoothing;
  - ETS(A,A,N) corresponds to Holt's method;
  - ETS(A,A,A) corresponds to Holt-Winters' method.
- In R, the package **forecast** allows to compute all these variations conveniently.

# Outline

# ARIMA models to capture time-correlation

- ARIMA models provide another approach to time series modeling and forecasting.
- While exponential smoothing models are based on a description of the trend and seasonality in the data, ARIMA models aim to describe the autocorrelations in the data.



# VAR(1) model

- Recall that we model the log-returns $\mathbf{r}_t = \Delta \mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$, where $\mathbf{y}_t$ denotes the log-prices.
- The VAR (Vector Auto-Regressive) model or order 1 is

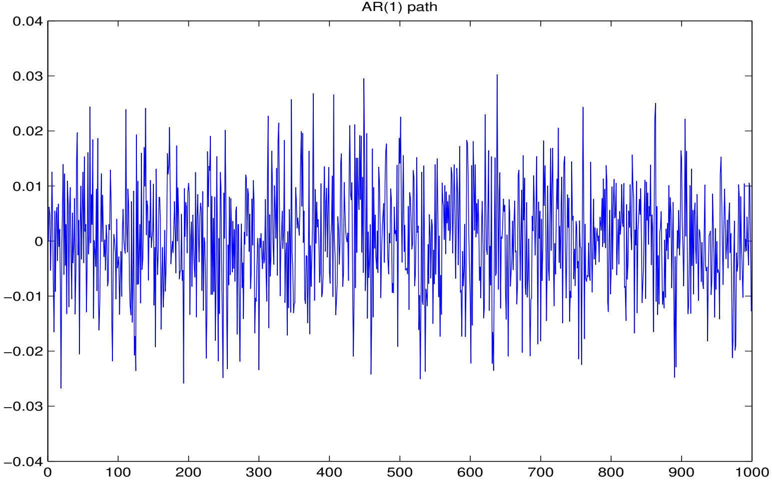
$$\mathbf{r}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Phi}_1 \mathbf{r}_{t-1} + \mathbf{w}_t,$$

  where
  - the vector $\boldsymbol{\phi}_0 \in \mathbb{R}^N$ and the matrix $\boldsymbol{\Phi}_1 \in \mathbb{R}^{N \times N}$ are parameters,
  - $\mathbf{w}_t$ is a white noise series with zero mean and constant covariance matrix $\boldsymbol{\Sigma}_w$.
- The conditional mean and covariance matrix are

$$\begin{aligned} \boldsymbol{\mu}_t &= \boldsymbol{\phi}_0 + \boldsymbol{\Phi}_1 \mathbf{r}_{t-1}, \\ \boldsymbol{\Sigma}_t &= \boldsymbol{\Sigma}_w. \end{aligned}$$

# AR(1) example

- A univariate AR(1) path looks like



# VAR($p$) model

- The VAR (Vector Auto-Regressive) model of order $p$ is

$$\mathbf{r}_t = \phi_0 + \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{r}_{t-i} + \mathbf{w}_t,$$

  where $p$ is a nonnegative integer and
  - the vector $\phi_0 \in \mathbb{R}^N$ and the matrices $\mathbf{\Phi}_i \in \mathbb{R}^{N \times N}$ are parameters,
  - $\mathbf{w}_t$ is a white noise series with zero mean and constant covariance matrix $\mathbf{\Sigma}_w$.
- The conditional mean and covariance matrix are

$$\begin{aligned} \boldsymbol{\mu}_t &= \phi_0 + \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{r}_{t-i}, \\ \mathbf{\Sigma}_t &= \mathbf{\Sigma}_w. \end{aligned}$$

# VARMA($p$,$q$) model

- The VARMA (Vector Auto-Regressive and Moving Average) model is

$$\mathbf{r}_t = \phi_0 + \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{r}_{t-i} + \mathbf{w}_t - \sum_{j=1}^{q} \mathbf{\Theta}_j \mathbf{w}_{t-j},$$

  where $p$ and $q$ are nonnegative integers and

  - the vector $\phi_0 \in \mathbb{R}^N$ and the matrices $\mathbf{\Phi}_i, \mathbf{\Theta}_j \in \mathbb{R}^{N\times N}$ are parameters,
  - $\mathbf{w}_t$ is a white noise series with zero mean and constant covariance matrix $\mathbf{\Sigma}_w$.

- The conditional mean and covariance matrix are

$$\begin{aligned}\boldsymbol{\mu}_t &= \phi_0 + \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{r}_{t-i} - \sum_{j=1}^{q} \mathbf{\Theta}_j \mathbf{w}_{t-j},\\ \mathbf{\Sigma}_t &= \mathbf{\Sigma}_w.\end{aligned}$$

# Order selection of models

- All time series models have an order that is typically assumed to be known and given, e.g., the orders $p$ and $q$ in a VARMA($p$,$q$) model.
- In practice, the order of a model is unknown and also has to be determined from the observed data.
- Observe that the higher the order, the more parameters the model has to fit the data and, thus, the better the fit. So it seems the best model will be the one with higher order.
- However, this is completely wrong, because it will be doomed to overfit the data: one thing is to fit better the training data, a very different one is to fit better the future coming data.
- In practice, there are two common approaches:
  - cross-validation: splitting the data into a training part and a cross-validation part, the latter being used to test the model trained with the training data for different combinations of orders.
  - penalized estimation methods: penalizing the number of parameters of the model with a penalty term like: AIC, BIC, SIC, HQIC, etc.[9]

[9] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media, 2007.

# ARIMA vs ETS

- It is a commonly held myth that ARIMA models are more general than exponential smoothing.[10]
- While linear exponential smoothing models are all special cases of ARIMA models, the non-linear exponential smoothing models have no equivalent ARIMA counterparts.
- On the other hand, there are also many ARIMA models that have no exponential smoothing counterparts. In particular, all ETS models are non-stationary, while some ARIMA models are stationary.

[10] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2014.

# VARIMA($p$,$d$,$q$) model

- A multivariate time series $\mathbf{y}_t$ is said to be a VARIMA($p$,1,$q$) process if it is nonstationary but after differencing the series times $\mathbf{x}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$ then $\mathbf{x}_t$ follows a stationary VARMA($p$,$q$) model.
- More generally, a VARIMA($p$,$d$,$q$) process has to be differenced $d$ times to obtain a stationary VARMA($p$,$q$) process.
- In finance, price series $\mathbf{p}_t$ (or log-prices $\mathbf{y}_t = \log(\mathbf{p}_t)$) are believed to be nonstationary, but the log-return series $\mathbf{r}_t = \mathbf{y}_t - \mathbf{y}_{t-1} = \log(\mathbf{p}_t) - \log(\mathbf{p}_{t-1})$ is stationary.
- Thus, it is the same to talk about a VARIMA($p$,1,$q$) log-price series and about a VARMA($p$,$q$) log-return series.

# Vector Error Correction Model (VECM)

- Until now we have focused on modeling directly the log-returns $\mathbf{r}_t = \Delta \mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$, where $\mathbf{y}_t$ denotes the log-prices.
- The reason is that in general the log-price time series $\mathbf{y}_t$ is not weakly stationary (first and second-order moments are not constant).
- Example: think of Apple stock whose log-prices keep increasing.
- On the other hand, the log-return time series $\mathbf{r}_t$ is weakly stationary (at least over some time horizon), which is good.
- However, it turns out that differencing may destroy part of the structure in the relationship among the log-prices of the stocks which may be invaluable for forecasting.
- So it makes sense to analyze the original (probably non-stationary, be careful!) time series in $\mathbf{y}_t$ directly:

$$\mathbf{y}_t = \phi_0 + \sum_{i=1}^{p} \mathbf{\Phi}_i \mathbf{y}_{t-i} + \mathbf{w}_t.$$

# VECM

- The VECM[11] is better written as

$$\mathbf{r}_t = \phi_0 + \boldsymbol{\Pi}\mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \tilde{\boldsymbol{\Phi}}_i \mathbf{r}_{t-i} + \mathbf{w}_t,$$

where the term $\boldsymbol{\Pi}\mathbf{y}_{t-1}$ is called error correction term and

$$\tilde{\boldsymbol{\Phi}}_j = -\sum_{i=j+1}^{p} \boldsymbol{\Phi}_i$$

$$\boldsymbol{\Pi} = -\left(\mathbf{I} - \boldsymbol{\Phi}_1 - \cdots - \boldsymbol{\Phi}_p\right).$$

- The conditional mean and covariance matrix are

$$\boldsymbol{\mu}_t = \phi_0 + \boldsymbol{\Pi}\mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \tilde{\boldsymbol{\Phi}}_i \mathbf{r}_{t-i},$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_w.$$

[11] R. F. Engle and C. W. J. Granger, "Co-integration and error correction: Representation, estimation, and testing", *Econometrica: Journal of the Econometric Society*, pp. 251–276, 1987.

# VECM - Matrix $\boldsymbol{\Pi}$

- The matrix $\boldsymbol{\Pi}$ is of extreme importance.
- Notice that from the model $\mathbf{r}_t = \phi_0 + \boldsymbol{\Pi}\mathbf{y}_{t-1} + \sum_{i=1}^{p-1} \tilde{\boldsymbol{\Phi}}_i \mathbf{r}_{t-i} + \mathbf{w}_t$ one can conclude that $\boldsymbol{\Pi}\mathbf{y}_t$ must be stationary even though $\mathbf{y}_t$ is not!!!
- If that happens, it is said that $\mathbf{y}_t$ is cointegrated.
- There are three possibilities for $\boldsymbol{\Pi}$:
    - $\text{rank}\left(\boldsymbol{\Pi}\right) = 0$: This implies $\boldsymbol{\Pi} = \mathbf{0}$, thus $\mathbf{y}_t$ is not cointegrated (so no mystery here) and the VECM reduces to a VAR model on the log-returns.
    - $\text{rank}\left(\boldsymbol{\Pi}\right) = N$: This implies $\boldsymbol{\Pi}$ is invertible and thus $\mathbf{y}_t$ must be stationary already
    - $0 < \text{rank}\left(\boldsymbol{\Pi}\right) < N$: This is the interestinc case and $\boldsymbol{\Pi}$ can be decomposed as $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}^T$ with $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{N \times r}$ with full column rank. This means that $\mathbf{y}_t$ has $r$ linearly independent cointegrated components, i.e., $\boldsymbol{\beta}^T\mathbf{y}_t$, which can be used to design mean-reversion statistical arbitrage investment strategies (e.g., pairs trading).

# Outline

# Volatility clustering

- Recall that conditional on the past history $\mathcal{F}_{t-1}$, we can decompose the returns as follows:

$$\mathbf{r}_t = \boldsymbol{\mu}_t + \mathbf{w}_t$$

  where

  - $\boldsymbol{\mu}_t$ is the *conditional mean*

$$\boldsymbol{\mu}_t = \mathsf{E}[\mathbf{r}_t | \mathcal{F}_{t-1}]$$

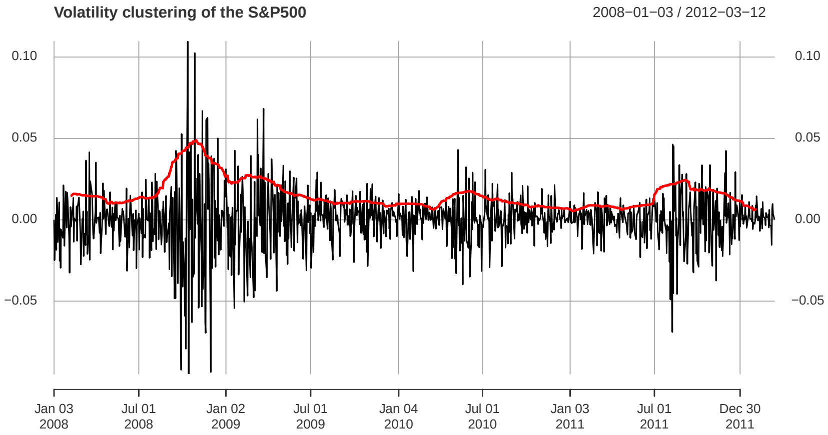
  - $\mathbf{w}_t$ is a white noise with zero mean and *conditional covariance*

$$\boldsymbol{\Sigma}_t = \mathsf{E}[(\mathbf{r}_t - \boldsymbol{\mu}_t)(\mathbf{r}_t - \boldsymbol{\mu}_t)^T | \mathcal{F}_{t-1}].$$

- We will focus now on the modeling of the term $\mathbf{w}_t$ and, more specifically, the covariance $\boldsymbol{\Sigma}_t$ (in the univariate case, it is just the variance $\sigma_t$).
- The previously models focus on modeling the conditional mean but assume that $\boldsymbol{\Sigma}_t$ is constant!

# Volatility clustering

- As we know from financial stylized facts, the volatility (i.e., the square root of conditional variance) is clustered:

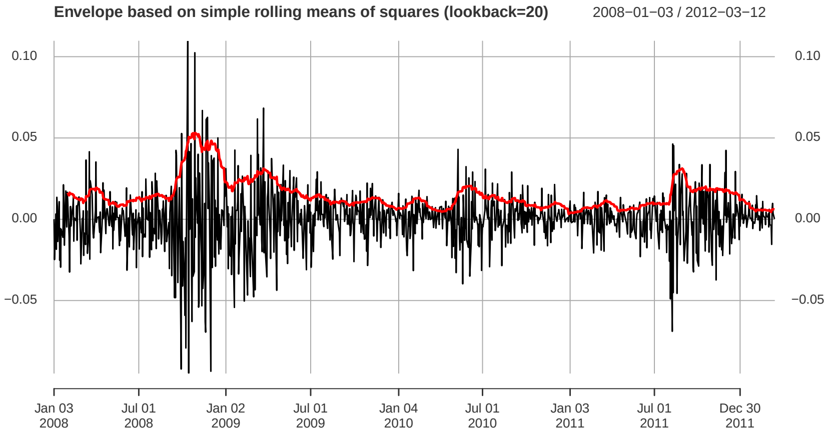


# Outline

# Moving average (MA) of squared returns

- Before we start with complicated models, we can consider a simple rolling means (aka moving average) of the squared returns:

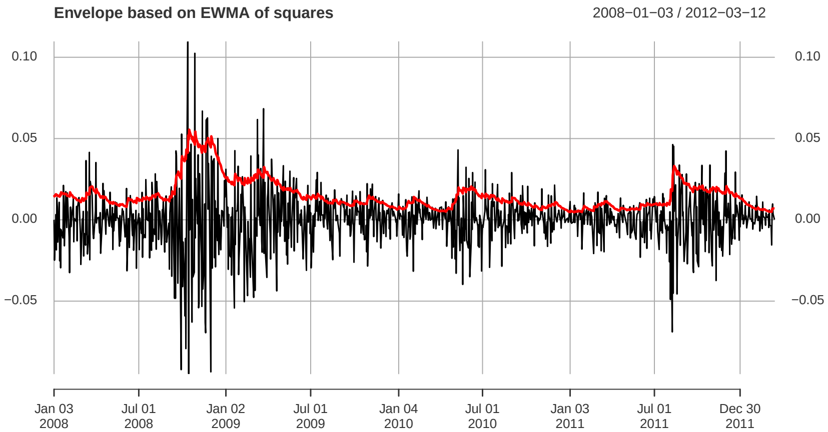
$$\sigma_t^2 = \frac{1}{m} \sum_{i=1}^{m} w_{t-i}^2$$



# Exponentially Weighted Moving average (EWMA) of squared returns

- We can now try an EWMA of the squared returns (after fitting, $\alpha = 0.097$):

$$\sigma_t^2 = \alpha w_{t-1}^2 + (1-\alpha)\sigma_{t-1}^2$$



# ARCH model

- The autoregressive conditional heteroskedasticity (ARCH) model is one of the earliest model to deal with the volatility clustering effect.
- The ARCH($m$) model[12] is

$$w_t = \sigma_t z_t,$$

where $z_t$ is a white noise series with zero mean and constant unit variance, and the conditional variance $\sigma_t^2$ is modeled by

$$\sigma_t^2 = \omega + \sum_{i=1}^{m} \alpha_i w_{t-i}^2$$

Here, $m$ is a nonnegative integer, $\omega > 0$, $\alpha_i \geq 0$ for all $i > 0$.

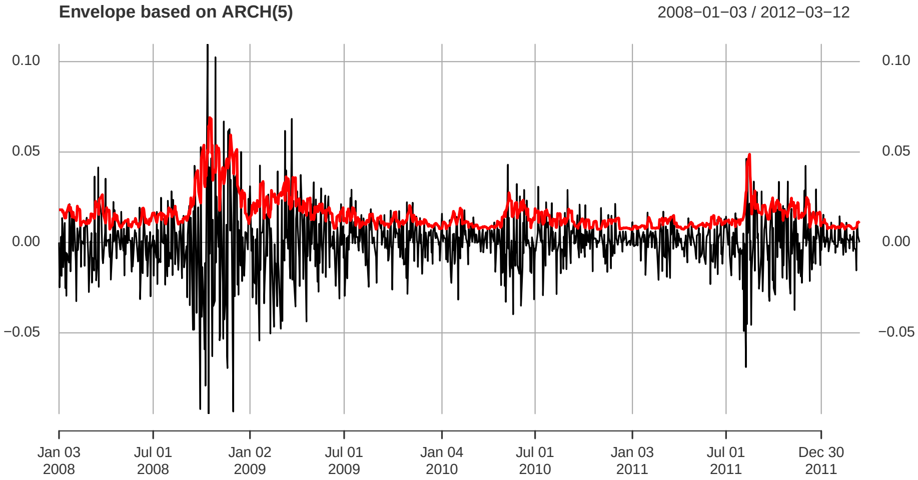
[12] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation", *Econometrica: Journal of the Econometric Society*, pp. 987–1007, 1982.

# ARCH model

Even though the ARCH model can model the conditional heteroskedasticity, it has several disadvantages:

- positive and negative noise have the same effects on volatility, but in practice they have different impact
- too restrictive to capture some patterns, e.g., excess kurtosis
- doesn't provide any new insight, just a mechanical way to describe the behavior of conditional variance
- tend to overpredict the volatility because it responds slowly to large isolated noise clusters.

# ARCH model example



# GARCH model

- A limitation of the ARCH model is that the high volatility is not persistent enough. This can be overcame by the Generalized ARCH (GARCH) model.[13]
- The GARCH$(m, s)$ model is

$$w_t = \sigma_t z_t,$$

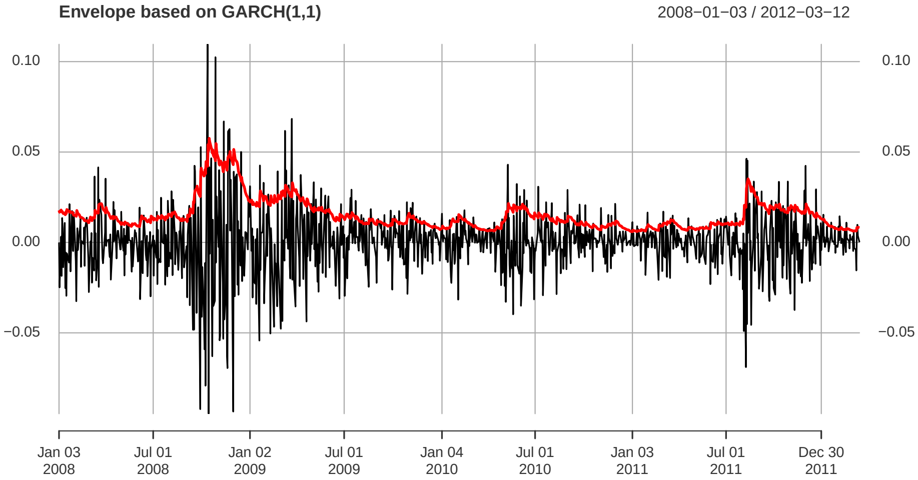
where $z_t$ is a white noise series with zero mean and constant unit variance, and the conditional variance $\sigma_t^2$ is modeled by

$$\sigma_t^2 = \omega + \sum_{i=1}^{m} \alpha_i w_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2.$$

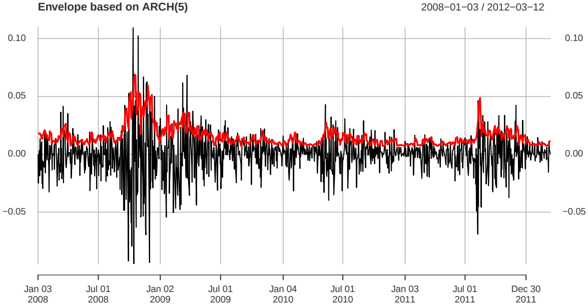
Here, $m$ and $s$ are nonnegative integers, $\omega > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$ for all $i > 0$ and $j > 0$ and $\sum_{i=1}^{m} \alpha_i + \sum_{j=1}^{s} \beta_j \leq 1$.

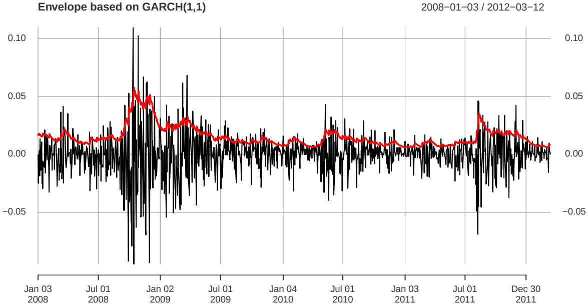
[13] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity", *Journal of Econometrics*, vol. 31, no. 3, pp. 307–327, 1986.

# GARCH example



# ARCH vs GARCH example

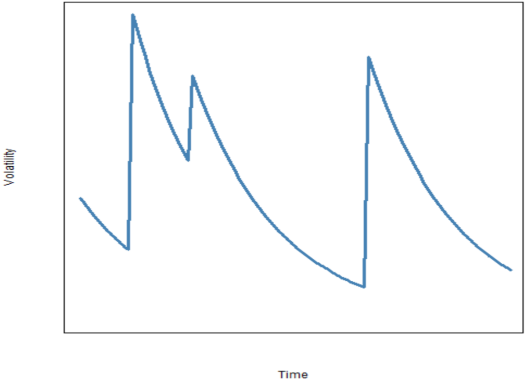


# Criticism of GARCH: Spike model

For criticism of GARCH see an insightful report by Patrick Burns:
https://www.burns-stat.com/pages/Present/3_realms_garch_modeling_annot.pdf

- GARCH thinks volatility is composed of exponentially decaying spikes:




- When the spikes happen is unpredictable.

# Criticism of GARCH: Spike model

- Consider a GARCH(1,1):

$$\sigma_t^2 = \omega + \alpha_1 w_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

with parameters: $\omega$, $\alpha_1$, and $\beta_1$.

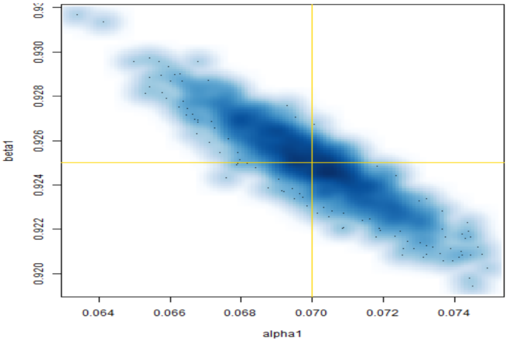
- If we set $\omega = 0$ and $\alpha_1 = 1 - \beta_1$, then we get an exponential smoothing:

$$\sigma_t^2 = (1 - \beta_1)\, w_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

- Thus we can say that GARCH is a glorified exponencial smoothing!
- So indeed GARCH thinks volatility is composed of an overlap of exponentially decaying spikes.

# Criticism of GARCH: Data hungry

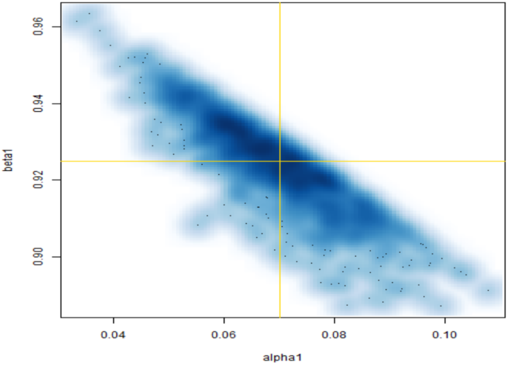
- Let's generate multiple synthetic realizations of a GARCH model with $\omega = 0$, $\alpha_1 = 0.07$, and $\beta_1 = 0.925$.
- If each realization has 100,000 observations and we estimate the parameters for each realization, we get the following scatter plot of the estimates:




- The range of estimates is about 0.01, which is good, but we used 100,000 observations (4 centuries of daily data).

# Criticism of GARCH: Data hungry

- If instead each realization has 2,000 observations (which is still a large number) and we estimate the parameters for each realization, we get the following scatter plot of the estimates:




- The estimates are not as good now, and 2,000 observations is still a lot (8 years of daily data).

# Outline

# Stochastic volatility model

- As an alternative to the GARCH class of models, Taylor porposed in a seminal work[14] to model the volatility probabilistically, i.e., through a state-space model where the logarithm of the squared volatilities follows an AR(1) process.
- This is called **stochastic volatility (SV)** model.
- The SV model has not enjoyed the popularity of the GARCH class models.
- There are very few software packages available to fit SV models.
- Fitting an SV model is computationally intensive.

[14] S. Taylor, "Financial returns modelled by the product of two stochastic processes: A study of daily sugar prices 1691â79", in *Time Series Analysis: Theory and Practice 1*, O. Anderson, Ed., North-Holland, Amsterdam, 1982, pp. 203–226.

# SV model

- Recall the decomposition of the returns as $r_t = \mu_t + w_t$, where $w_t$ is the innovation and can be understood as the demeaned return.
- The instantaneous variance of $w_t$, which before we denoted by the latent variable $\sigma_t^2$, here is modeled as $\sigma_t^2 = \exp(h_t)$ and $h_t$ is allowed to smoothly change following an AR(1) process:

$$\begin{aligned} w_t &= \exp(h_t/2)\, z_t \\ h_t - \bar{h} &= \phi\left(h_{t-1} - \bar{h}\right) + u_t \end{aligned}$$

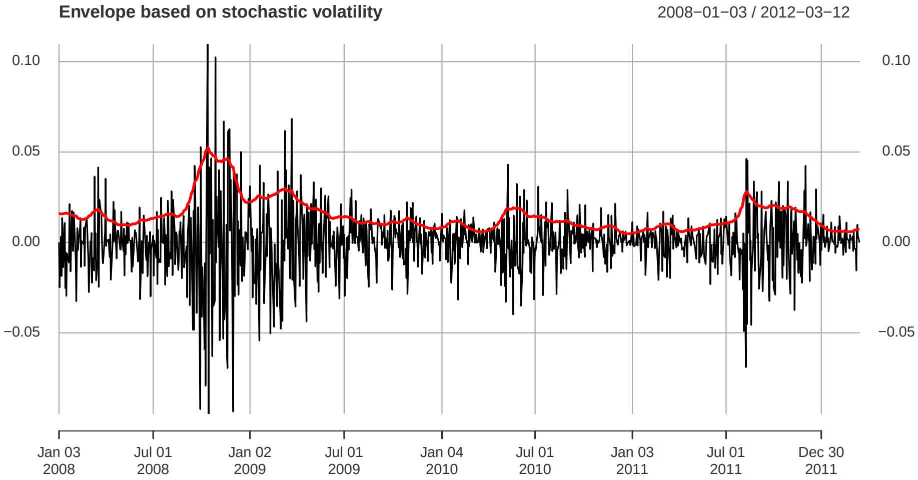
  where $z_t$ is white noise with zero mean and unit variance.
- Equivalently, we can write this model in terms of $\sigma_t$ as

$$\begin{aligned} w_t &= \sigma_t z_t \\ \log\left(\sigma_t^2\right) &= \bar{h} + \phi\left(\log\left(\sigma_{t-1}^2\right) - \bar{h}\right) + u_t \end{aligned}$$

- Compare with the GARCH(1,1) model:

$$\sigma_t^2 = \omega + \alpha_1 w_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

# SV example


Envelope based on stochastic volatility
2008−01−03 / 2012−03−12
0.10
0.05
0.00
−0.05
0.10
0.05
0.00
−0.05
Jan 03 2008
Jul 01 2008
Jan 02 2009
Jul 01 2009
Jan 04 2010
Jul 01 2010
Jan 03 2011
Jul 01 2011
Dec 30 2011

# Components model

- For completeness, it is worth mentioning that there are other possible models out there: The components model is one such example.
- Recall the GARCH(1,1) model:

$$\sigma_t^2 = \omega + \alpha_1 w_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

- The components model is

$$\begin{aligned} q_t =& \omega + \rho q_{t-1} + \phi \left( w_{t-1}^2 - q_{t-1} \right) \\ \sigma_t^2 =& q_t + \alpha_1 \left( w_{t-1}^2 - q_{t-1} \right) + \beta_1 \left( \sigma_{t-1}^2 - q_{t-1} \right). \end{aligned}$$

- The interpretation is that there is a smooth long-term trend in volatility, $q_t$, and then a short-term volatility that wiggles around the long-term trend.
- The parameter $\rho$ is the persistence in the components model.
- This model can have a high volatility long-term regime but low volatility short-term, and vice versa.

# Outline

# Multivariate GARCH model

- The multivariate noise (a vector) is modeled as

$$\mathbf{w}_t = \boldsymbol{\Sigma}_t^{1/2}\mathbf{z}_t,$$

  where $\mathbf{z}_t \in \mathbb{R}^N$ is an i.i.d. white noise series with zero mean and constant covariance matrix $\mathbf{I}$.

- The key is to model the conditional covariance matrix $\boldsymbol{\Sigma}_t$.
- But watch out as the number of parameters may quickly explode... and that will inevitably produce overfitting.

# VEC GARCH model

- One of the first extensions to the vector case is the vector (VEC) GARCH model:

$$\text{vech}\left(\boldsymbol{\Sigma}_t\right) = \mathbf{a}_0 + \sum_{i=1}^{m} \tilde{\mathbf{A}}_i \text{vech}(\mathbf{w}_{t-i}\mathbf{w}_{t-i}^T) + \sum_{j=1}^{s} \tilde{\mathbf{B}}_j \text{vech}\left(\boldsymbol{\Sigma}_{t-j}\right),$$

  where $m$ and $s$ are nonnegative integers, $\text{vech}(\cdot)$ is the half-vectorization operator that keeps the $N(N+1)/2$ lower triangular part of its $N \times N$ matrix argument, $\mathbf{a}_0$ is an $N(N+1)/2$ dimensional vector, and $\tilde{\mathbf{A}}_i, \tilde{\mathbf{B}}_j$ are $N(N+1)/2$ by $N(N+1)/2$ matrices.
- Advantage: This model is very flexible.
- Disadvantages: Does not guarantee $\boldsymbol{\Sigma}_t$ to be a positive definite covariance matrix and the number of parameters grows quickly as $O\left((m+s)\,N^4\right)$.

# Diagonal VEC (DVEC) model

- The DVEC model[15] is more parsimonious model assuming that $\tilde{\mathbf{A}}_i, \tilde{\mathbf{B}}_j$ are diagonal and can be simplified as

$$\boldsymbol{\Sigma}_t = \mathbf{A}_0 + \sum_{i=1}^{m} \mathbf{A}_i \odot \left(\mathbf{w}_{t-i}\mathbf{w}_{t-i}^T\right) + \sum_{j=1}^{s} \mathbf{B}_j \odot \boldsymbol{\Sigma}_{t-j},$$

  where $\mathbf{A}_i, \mathbf{B}_j$ are **symmetric** $N \times N$ matrix parameters. Here, the operator $\odot$ denotes the Hadamard (elementwise) product can be interpreted as moving weight matrices.
- Advantage: This is an element-wise GARCH model, so very simple.
- Disadvantages: Still $\boldsymbol{\Sigma}_t$ is not guaranteed to be positive-definite and the number of parameters grows more slowly but still fast as $O\left((m+s)\,N^2\right)$.

[15] T. Bollerslev, R. F. Engle, and J. M. Wooldridge, "A capital asset pricing model with time-varying covariances", *The Journal of Political Economy*, pp. 116–131, 1988.

# BEKK model

- To guarantee a positive-definite $\boldsymbol{\Sigma}_t$, the Baba-Engle-Kraft-Kroner (BEKK) model[16] was proposed as

$$\boldsymbol{\Sigma}_t = \mathbf{A}_0\mathbf{A}_0^T + \sum_{i=1}^{m} \mathbf{A}_i(\mathbf{w}_{t-i}\mathbf{w}_{t-i}^T)\mathbf{A}_i^T + \sum_{j=1}^{s} \mathbf{B}_j\boldsymbol{\Sigma}_{t-j}\mathbf{B}_j^T,$$

  where $\mathbf{A}_i, \mathbf{B}_j$ are $N \times N$ matrix parameters and $\mathbf{A}_0$ is lower triangular.

- Advantage: Guarantees positive definiteness of $\boldsymbol{\Sigma}_t$.
- Disadvantages:
  - The parameters $\mathbf{A}_i$ and $\mathbf{B}_j$ do not have direct interpretations.
  - Number of parameters still increases as $O\left((m+s)\,N^2\right)$ (although now roughly the number of parameters is twice as that in DVEC).

[16] R. F. Engle and K. F. Kroner, "Multivariate simultaneous generalized ARCH", *Econometric Theory*, vol. 11, no. 01, pp. 122–150, 1995.

# CCC model

- The constant conditional correlation (CCC) model[17] restricts the number of parameters while still guaranteeing the positive definite covariance.
- The idea is to model the conditional heteroskedasticity in each asset while having a constant correlation.
- Mathematically, the model is

$$\boldsymbol{\Sigma}_t = \mathbf{D}_t \mathbf{C} \mathbf{D}_t,$$

  where $\mathbf{D}_t = \mathrm{Diag}\left(\sigma_{1,t}, \ldots, \sigma_{N,t}\right)$ is the time-varying conditional volatilities of each stock and $\mathbf{C}$ is the CCC matrix of the standardized noise vector $\boldsymbol{\eta}_t = \mathbf{D}_t^{-1} \mathbf{w}_t$.

- Advantages: Guarantees positive definiteness of $\boldsymbol{\Sigma}_t$ and small number of parameters that grows as $O\left(\left(m+s\right)N + N^2\right)$.
- Disadvantages: Not too flexible due to constant asset correlations.

[17] T. Bollerslev, "Modelling the coherence in short-run nominal exchange rates: A multivariate generalized arch model", *The Review of Economics and Statistics*, pp. 498–505, 1990.

# DCC model

- The main limitation of the CCC model is that the correlation is constant.
- To overcome this drawback, the dynamic conditional correlation (DCC) was proposed[18] as

$$\boldsymbol{\Sigma}_t = \mathbf{D}_t \mathbf{C}_t \mathbf{D}_t,$$

where $\mathbf{C}_t$ contains diagonal elements equal to 1.

- In particular, Engle modeled it as follows:

$$C_{ij,t} = \frac{q_{ij,t}}{\sqrt{q_{ii,t} q_{jj,t}}}$$

with each $q_{ij,t}$ modeled by a simple GARCH(1,1) model:

$$q_{ij,t} = \alpha \eta_{i,t-1} \eta_{j,t-1} + (1 - \alpha)\, q_{ij,t-1}$$

[18] R. F. Engle, "Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models", *Journal of Business & Economic Statistics*, vol. 20, no. 3, pp. 339–350, 2002.

# DCC model

- More compactly, in matrix form:

$$\mathbf{Q}_t = \alpha \boldsymbol{\eta}_t \boldsymbol{\eta}_t^T + (1-\alpha)\,\mathbf{Q}_{t-1}.$$

  and

$$\mathbf{C}_t = \mathrm{Diag}^{-1/2}\left(\mathbf{Q}_t\right)\mathbf{Q}_t \mathrm{Diag}^{-1/2}\left(\mathbf{Q}_t\right).$$

- Advantages: Guarantees positive definiteness of $\boldsymbol{\Sigma}_t$ and small number of parameters that grows as $O\left(\left(m+s\right)N\right)$.
- Disadvantages: Good flexibility, although it forces all the correlation coefficients to have the same memory via the same $\alpha$.

# Beyond

- The previous models for the conditional mean and covariance matrix can be jointly combined to fit the financial data better.
- Limitations:
  - High-frequency data: when the sampling period becomes very small, say minutes, seconds, or even smaller, the previous models become invalid and one reaches a "quantum regime" where things are not fluid anymore but quantized into the limit order book. Not only the models have to be properly modified, but also the computer and internet communication speed matter (e.g., co-location of computers).
  - Heavy tails: most models assume a Gaussian distribution for simplicity, but they can be easily extended to deal with heavy-tailed distributions.
  - Lack of stationarity: financial data is only stationary for some time horizon, this produces a tradeoff between having enough data to properly estimate the parameters of the model but still within the stationarity time horizon.
  - Other practical details: different stocks may have a different historical length and some days the prices may be missing due to no trading or bad quality of data.

# Great References on Time Series Models

- H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Springer Science & Business Media, 2007
- R. S. Tsay, *Analysis of Financial Time Series*, 3rd. John Wiley & Sons, 2010
- R. S. Tsay, *Multivariate Time Series Analysis: With R and Financial Applications*. John Wiley & Sons, 2014

# Outline

# Model fitting

- Each model has some parameters that need to be fitted/estimated/calibrated to fit the observed data.
- About the fitting process:
  - In some cases, this fitting can be as simple as a least squares (LS) problem.
  - In some other cases, it can be more involved but still doable with closed-form expressions or fixed-point solutions that can be solved iteratively.
  - In some extreme cases, no analytical expressions can be found and one has to resort to numerical Monte-Carlo based methods that require intensive computational power (e.g., to approximate an integral).
- About the model itself:
  - Some models are stable in the sense that the parameter estimation is reliable and not too sensitive to each data realization.
  - However, other models are extremely sensitive to the data: different realizations of the estimation errors may give you very different values for the parameters.

# General estimation methodologies: LS

**Least squares (LS) estimator:**

- The idea is to define an error between the observed financial data and the model under consideration, and then minimize the $\ell_2$-norm of the error.
- For example, for a VAR(1) model $\mathbf{r}_t = \phi_0 + \mathbf{\Phi}_1 \mathbf{r}_{t-1} + \mathbf{w}_t$, where we have $T$ observations, the problem to solve would be

$$\underset{\phi_0, \mathbf{\Phi}_1}{\text{minimize}} \quad \sum_{t=2}^{T} \left\| \mathbf{r}_t - \phi_0 - \mathbf{\Phi}_1 \mathbf{r}_{t-1} \right\|^2 \quad .$$

# General estimation methodologies: MLE

**Maximum likelihood estimator (MLE):**

- The idea is to assume some distribution for the residual of the model $\mathbf{w}_t$, typically Gaussian for mathematical simplicity and tractability:

$$f(\mathbf{r}) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{r}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{r}-\boldsymbol{\mu})}$$

- Then, given the $T$ samples, the negative log-likelihood function is formed

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{T}{2} \log |\boldsymbol{\Sigma}| + \frac{1}{2} \sum_{t=1}^{T} (\mathbf{r}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{r}_t - \boldsymbol{\mu}) + \mathrm{const}.$$

- Finally, one can minimize the negative log-likelihood with respect to the parameters to be estimated, like $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

# Estimation of i.i.d. model: Sample estimators

- i.i.d. model:
$$\mathbf{r}_t = \boldsymbol{\mu} + \mathbf{w}_t,$$
where $\boldsymbol{\mu} \in \mathbb{R}^N$ is the mean and $\mathbf{w}_t \in \mathbb{R}^N$ is a white noise series with zero mean and constant covariance matrix $\boldsymbol{\Sigma}$.
- Good old sample estimators (sample mean and sample covariance matrix):
$$\hat{\boldsymbol{\mu}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_t,$$
$$\hat{\boldsymbol{\Sigma}} = \frac{1}{T-1}\sum_{t=1}^{T}(\mathbf{r}_t - \hat{\boldsymbol{\mu}})(\mathbf{r}_t - \hat{\boldsymbol{\mu}})^T.$$
- In practice: they are horrible!
- They can be improved with heavy-tail estimators and shrinkage.

# Estimation of i.i.d. model: LS estimator

- Minimize the least-square error in the $T$ observed i.i.d. samples:

$$\underset{\boldsymbol{\mu}}{\text{minimize}} \quad \frac{1}{T}\sum_{t=1}^{T}\left\|\mathbf{r}_t-\boldsymbol{\mu}\right\|_2^2.$$

- The optimal solution is the sample mean:

$$\hat{\boldsymbol{\mu}}=\frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_t$$

- The sample covariance of the residuals is the sample covariance estimator:

$$\hat{\boldsymbol{\Sigma}}=\frac{1}{T}\sum_{t=1}^{T}(\mathbf{r}_t-\hat{\boldsymbol{\mu}})(\mathbf{r}_t-\hat{\boldsymbol{\mu}})^T.$$

# Estimation of i.i.d. model: Gaussian Maximum Likelihood Estimator (MLE)

- Assume $\mathbf{r}_t$ are i.i.d. Gaussian distributed:

$$f(\mathbf{r}) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{r}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{r}-\boldsymbol{\mu})}$$

- Given the $T$ i.i.d. samples, the negative log-likelihood function is

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{T}{2} \log |\boldsymbol{\Sigma}| + \frac{1}{2} \sum_{t=1}^{T} (\mathbf{r}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{r}_t - \boldsymbol{\mu}) + \text{const}.$$

- Setting the derivative of $\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ w.r.t. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{-1}$ to zero and solving the equations yield:

$$\boldsymbol{\mu} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{r}_t,$$

$$\boldsymbol{\Sigma} = \frac{1}{T} \sum_{t=1}^{T} (\mathbf{r}_t - \boldsymbol{\mu})(\mathbf{r}_t - \boldsymbol{\mu})^T.$$

# Estimation of Factor Model: MLE

**Likelihood of the factor model:**

The log-likelihood of the parameters $(\boldsymbol{\alpha}, \boldsymbol{\Sigma})$ given $T$ i.i.d. observations $\mathbf{r}_t = \boldsymbol{\alpha} + \mathbf{B}\mathbf{f}_t + \mathbf{w}_t$ is

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\Sigma}) &= \log p(\mathbf{x}_1, \ldots, \mathbf{x}_T \mid \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \\ &= -\frac{TN}{2}\log(2\pi) - \frac{T}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{t=1}^{T}(\mathbf{x}_t - \boldsymbol{\alpha})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_t - \boldsymbol{\alpha}) \end{aligned}$$

**Maximum likelihood estimation (MLE):**

$$\begin{array}{ll} \underset{\boldsymbol{\alpha}, \boldsymbol{\Sigma}, \mathbf{B}, \boldsymbol{\Psi}}{\text{minimize}} & \frac{T}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{t=1}^{T}(\mathbf{x}_t - \boldsymbol{\alpha})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_t - \boldsymbol{\alpha}) \\ \text{subject to} & \boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^T + \boldsymbol{\Psi} \end{array}$$

- Without constraint $\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^T + \boldsymbol{\Psi}$, the solution is trivially the sample mean and sample covariance matrix as we have seen before.
- However, with such difficult nonconvex constraint, the problem becomes very involved and sophisticated methods are necessary.

# Estimation of VAR model with sparsity

- Sparsity refers to parameters having zero entries, so effectively reducing the number of parameters and the danger of overfitting.
- Mathematically, the number of nonzero entries of a vector or matrix is expressed via the $\ell_0$-pseudo norm $\|\cdot\|_0$.
- In practice, the $\ell_0$-pseudo norm is tough to manage and optimize and it is commonly approximated with the $\ell_1$-norm $\|\cdot\|_1$.
- Countless of examples where sparsity naturally arises in finance:
  - Sparse PCA for factor modeling: computation of sparse eigenvectors is key in the high-dimensional setting for automatic feature selection.[19,20]
  - Sparse parameters in all the multivariate models are required for parameter reduction (feature selection) such as VAR:

$$\begin{array}{ll} \underset{\boldsymbol{\phi_0}, \boldsymbol{\Phi_1}}{\text{minimize}} & \sum_{t=1}^{T} \left\| \mathbf{r}_t - \phi_0 - \boldsymbol{\Phi}_1 \mathbf{r}_{t-1} \right\|^2 \\ \text{subject to} & \left\| \boldsymbol{\Phi}_1 \right\|_0 \leq P \end{array}$$

[19] J. Song, P. Babu, and D. P. Palomar, "Sparse generalized eigenvalue problem via smooth optimization", *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1627–1642, 2015.
[20] K. Benidis, Y. Sun, P. Babu, and D. P. Palomar, "Orthogonal sparse PCA and covariance estimation via procrustes reformulation", *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6211–6226, 2016.

# Estimation of models with low rank

- Low-rank matrices are also useful to effectively reduce the number of parameters to be estimated and the danger of overfitting.
- For example, a VAR(1) model has parameters $\phi_0$ and $\mathbf{\Phi}_1$, which amounts to $N + N^2$ parameters. If the matrix has, say, rank $r \ll N$, then the number of parameters becomes $N + 2Nr$, which can be much smaller. If $N = 100$ and $r = 5$, then we go from $10,100$ parameters to $1,100$, which is one order of magnitude smaller.
- Low-rank naturally arises in finance:
  - Low-rank matrices are required to discover the low-dimensional structure in models like VAR:

$$\begin{array}{ll} \underset{\phi_0, \mathbf{\Phi}_1}{\text{minimize}} & \sum_{t=1}^{T} \|\mathbf{r}_t - \phi_0 - \mathbf{\Phi}_1 \mathbf{r}_{t-1}\|^2 \\ \text{subject to} & \text{rank}\left(\mathbf{\Phi}_1\right) \leq K \end{array}$$

  - Low-rank matrices are necessary in multivariate GARCH models for dimensionality reduction:

$$\begin{array}{ll} \underset{\{\mathbf{\Sigma}_t\}, \{\mathbf{B}_i\}}{\text{minimize}} & \frac{T}{2} \log |\mathbf{\Sigma}| + \frac{1}{2} \sum_{t=1}^{T} \mathbf{w}_t{}^T \mathbf{\Sigma}^{-1} \mathbf{w}_t \\ \text{subject to} & \mathbf{\Sigma}_t = \mathbf{B}_0 \mathbf{B}_0^T + \sum_{j=1}^{s} \mathbf{B}_j \mathbf{\Sigma}_{t-j} \mathbf{B}_j^T \\ & \text{rank}\left(\mathbf{B}_j\right) \leq K. \end{array}$$

# Estimation of VECM with low rank matrix $\mathbf{\Pi}$

- Another example where a low-rank matrix is required is in VECM modeling, in particular for matrix $\mathbf{\Pi}$:

$$\begin{array}{ll} \underset{\phi_0, \mathbf{\Phi}_1, \mathbf{\Pi}}{\text{minimize}} & \sum_{t=1}^{T} \left\| \mathbf{r}_t - \phi_0 - \mathbf{\Pi} \mathbf{y}_{t-1} - \mathbf{\Phi}_1 \mathbf{r}_{t-1} \right\|^2 \\ \text{subject to} & \left\| \mathbf{\Phi}_1 \right\|_0 \leq P \\ & \text{rank}\left(\mathbf{\Pi}\right) \leq K. \end{array}$$

# Estimation of SV model

- Recall the SV model:

$$\begin{aligned} w_t &= \exp\left(h_t/2\right) z_t \\ h_t - \bar{h} &= \phi\left(h_{t-1} - \bar{h}\right) + u_t \end{aligned}$$

- This is reminiscent of the popular linear state-space model under a Gaussian distribution easily estimated with Kalman[21]:

$$\begin{aligned} y_t &= a x_t + z_t \\ x_t &= b x_{t-1} + u_t. \end{aligned}$$

- However, Kalman filter cannot be used here since the SV model is not additive, not linear, and not Gaussian.
- Solutions:
  - extended Kalman filter: simple estimation but it's just an approximation;
  - Markow Chain Monte Carlo (MCMC) methods: computationally intensive but accurate (e.g., the R package stochvol).

[21] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods*, 2nd Ed. Oxford University Press, 2012.

# Outline

# Summary

- The returns can be expressed as $\mathbf{r}_t = \boldsymbol{\mu}_t + \mathbf{w}_t$ where $\boldsymbol{\mu}_t = \mathsf{E}[\mathbf{r}_t|\mathcal{F}_{t-1}]$ is the conditional mean on the history $\mathcal{F}_{t-1}$ and $\mathbf{w}_t$ is the residual with conditional covariance matrix $\boldsymbol{\Sigma}_t = \mathsf{E}[(\mathbf{r}_t - \boldsymbol{\mu}_t)(\mathbf{r}_t - \boldsymbol{\mu}_t)^T|\mathcal{F}_{t-1}]$.
- We have overviewed many models for the conditional mean: i.i.d. model, factor model, VAR models, VMA models, VARMA models, VECM, etc.
- We have overviewed the two basic models for the univariate conditional volatility that attemps to model the volatility clustering: ARCH and GARCH.
- The volatility clustering models can be extended to the multivariate case: VEC, DVEC, BEKK, CCC, DCC.
- The estimation of these models can be simple in some cases but also very difficult in other cases like the volatility clustering models.
- Many packages available in R for the fitting of these models.

# Thanks

For more information visit:

https://www.danielppalomar.com