MDI343
Time series : an introduction

François Roueff
http://perso.telecom-paristech.fr/~roueff/

Telecom ParisTech

Jan. 7, 2020

Outline

Example : USD vs EUR currency exchange rate

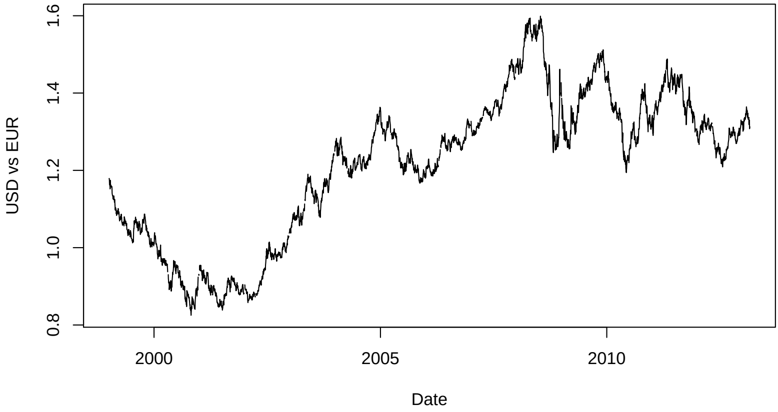
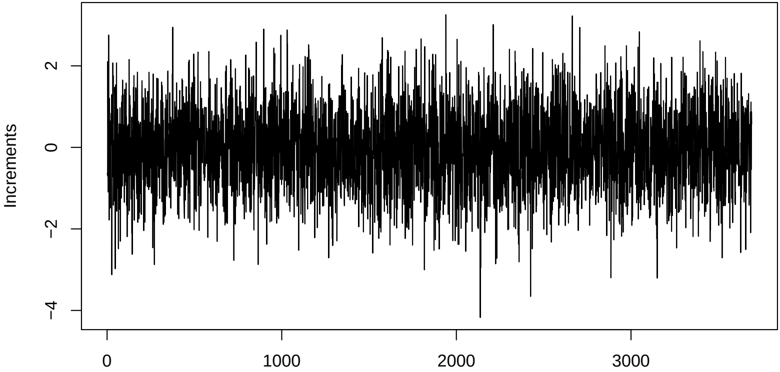


Figure: Daily currency exchange rate : price of 1 Euro in US Dollars.

Example : USD vs EUR currency exchange rate (cont.)

Compare with an IID $\mathcal{N}(0,1)$ sequence:



Example : USD vs EUR currency exchange rate (cont.)

Applying the differencing operator, we obtain the increment process

$$Y = \Delta X \quad \text{defined by} \quad Y_t = X_t - X_{t-1}, \quad t \in \mathbb{Z} .$$

Makes the “local” mean “more constant”.

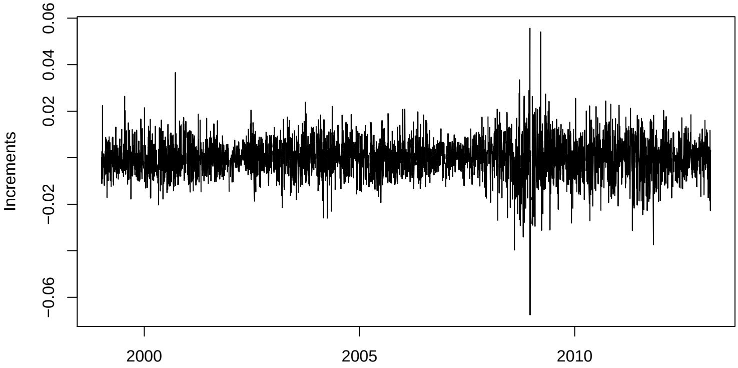


Figure: Increments of daily USD-EUR currency exchange rate.

Example : USD vs EUR currency exchange rate (cont.)

Applying the differencing operator of the logs, we obtain the log returns

$$Y = \Delta \log X \quad \text{defined by} \quad Y_t = \log X_t - \log X_{t-1}, \quad t \in \mathbb{Z} \ .$$

Makes the "local" mean and the variance "more constant".

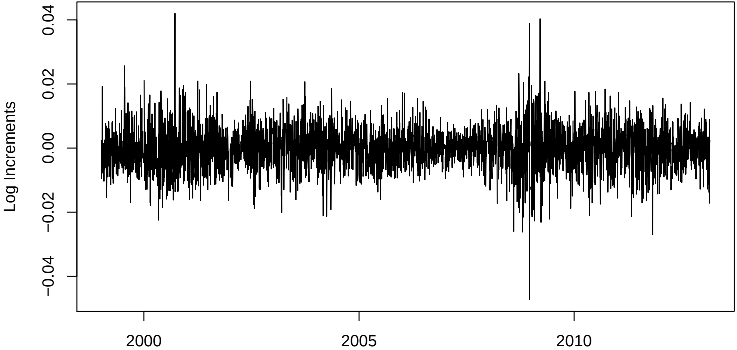


Figure: Log returns of daily USD-EUR currency exchange rate.

Example : USD vs EUR currency exchange rate (cont.)

Looking at things “locally” ...

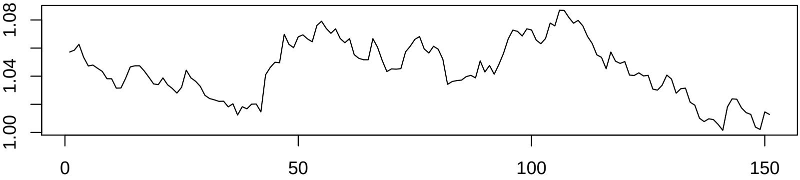


Figure: Daily currency exchange rate : price of 1 Euro in US Dollars, on a shorter observation window: between 1999-05-21 and 1999-12-17.

The mean and variance does not appear to vary too much, but still not i.i.d.

Discrete observations

- If we observe i.i.d. discrete observations $X_1, \dots, X_n$, then the log-likelihood can be defined as

$$L_n(\theta) = \sum_{k=1}^{n} \log p_\theta(X_k) \; ,$$

where, for all x in the discrete observation space and parameter θ

$$p_\theta(x) = \mathbb{P}_\theta(X_1 = x) \; .$$

Discrete observations

- If we observe i.i.d. discrete observations $X_1, \ldots, X_n$, then the log-likelihood can be defined as

$$L_n(\theta) = \sum_{k=1}^{n} \log p_\theta(X_k) \ ,$$

where, for all x in the discrete observation space and parameter θ

$$p_\theta(x) = \mathbb{P}_\theta(X_1 = x) \ .$$

- We denote the marginal distribution of X_1 under $\mathbb{P}_\theta$ by $\mathbb{P}_\theta^{X_1}$.

Discrete observations

- If we observe i.i.d. discrete observations $X_1, \ldots, X_n$, then the log-likelihood can be defined as

$$L_n(\theta) = \sum_{k=1}^{n} \log p_\theta(X_k) \,,$$

where, for all x in the discrete observation space and parameter θ

$$p_\theta(x) = \mathbb{P}_\theta(X_1 = x) \,.$$

- We denote the marginal distribution of X_1 under $\mathbb{P}_\theta$ by $\mathbb{P}_\theta^{X_1}$.
- Setting the definition of $\mathbb{P}_\theta^{X_1}$ or p_θ for all θ provides a statistical model for the observations $X_1, \ldots, X_n$.

Examples

- Bernoulli model:

$$p_\theta(x) = \theta^x (1-\theta)^{1-x}, \quad \theta \in (0,1), \quad x \in \{0,1\}.$$

Examples

- Bernoulli model:

$$p_\theta(x) = \theta^x (1-\theta)^{1-x}\ , \quad \theta \in (0,1)\ , \quad x \in \{0,1\}\ .$$

- Binomial model (with $p \geq 1$ known):

$$p_\theta(x) = \binom{p}{x} \theta^x (1-\theta)^{p-x}\ , \quad \theta \in (0\ , \quad 1)\ , \quad x \in \{0,\dots,p\}\ .$$

Examples

- Bernoulli model:

$$p_\theta(x) = \theta^x (1-\theta)^{1-x} , \quad \theta \in (0,1) , \quad x \in \{0,1\} .$$

- Binomial model (with $p \geq 1$ known):

$$p_\theta(x) = \binom{p}{x} \theta^x (1-\theta)^{p-x} , \quad \theta \in (0 , \quad 1) , \quad x \in \{0,\dots,p\} .$$

- Geometric model:

$$p_\theta(x) = \theta (1-\theta)^{x-1} , \quad \theta \in (0,1) , \quad x \in \mathbb{N}^* .$$

Examples

- Bernoulli model:

$$p_\theta(x) = \theta^x (1-\theta)^{1-x}\,, \quad \theta \in (0,1)\,, \quad x \in \{0,1\}\,.$$

- Binomial model (with $p \geq 1$ known):

$$p_\theta(x) = \binom{p}{x} \theta^x (1-\theta)^{p-x}\,, \quad \theta \in (0\,, \quad 1)\,, \quad x \in \{0,\ldots,p\}\,.$$

- Geometric model:

$$p_\theta(x) = \theta (1-\theta)^{x-1}\,, \quad \theta \in (0,1)\,, \quad x \in \mathbb{N}^*\,.$$

- Negative binomial, Poisson, ...

Continuous observations

- If we observe i.i.d. real valued observations $X_1, \ldots, X_n$, then the log-likelihood can be defined as

$$L_n(\theta) = \sum_{k=1}^{n} \log p_\theta(X_k) \;,$$

where, for all x in the discrete observation space and parameter θ, p_θ is the density of $\mathbb{P}_\theta^{X_1}$:

$$\mathbb{P}_\theta^{X_1}(A) = \mathbb{P}_\theta(X_1 \in A) = \int_A p_\theta(x)\, \mathrm{d}x \;.$$

Continuous observations

- If we observe i.i.d. real valued observations $X_1, \ldots, X_n$, then the log-likelihood can be defined as

$$L_n(\theta) = \sum_{k=1}^{n} \log p_\theta(X_k) \;,$$

where, for all x in the discrete observation space and parameter θ, p_θ is the density of $\mathbb{P}_\theta^{X_1}$:

$$\mathbb{P}_\theta^{X_1}(A) = \mathbb{P}_\theta(X_1 \in A) = \int_A p_\theta(x)\, \mathrm{d}x \;.$$

- Again, setting the definition of $\mathbb{P}_\theta^{X_1}$ or p_θ for all θ provides a statistical model for the observations $X_1, \ldots, X_n$.

Examples

- Exponential model:

$$p_\lambda(x) = \lambda \, \mathrm{e}^{-\lambda \, x} \, \mathbb{1}_{\mathbb{R}_+}(x) \; , \quad \lambda > 0 \; .$$

Examples

- Exponential model:

$$p_{\lambda}(x) = \lambda \, \mathrm{e}^{-\lambda \, x} \, \mathbb{1}_{\mathbb{R}_+}(x) \; , \quad \lambda > 0 \; .$$

- Gamma model

$$p_{\theta}(x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \, x^{\alpha-1} \mathrm{e}^{-\lambda \, x} \, \mathbb{1}_{\mathbb{R}_+}(x) \; , \quad \theta = (\alpha, \lambda) \in {\mathbb{R}_+^*}^2 \; .$$

Examples

- Exponential model:

$$p_{\lambda}(x) = \lambda \, \mathrm{e}^{-\lambda \, x} \, \mathbb{1}_{\mathbb{R}_+}(x) \;, \quad \lambda > 0 \;.$$

- Gamma model

$$p_{\theta}(x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \, x^{\alpha-1} \mathrm{e}^{-\lambda \, x} \, \mathbb{1}_{\mathbb{R}_+}(x) \;, \quad \theta = (\alpha, \lambda) \in {\mathbb{R}_+^*}^2 \;.$$

- Beta model :

$$p_{\theta}(x) = B(\alpha, \beta) \, x^{\alpha-1} (1-x)^{\beta-1} \, \mathbb{1}_{[0,1]}(x) \;, \quad \theta = (\alpha, \beta) \in {\mathbb{R}_+^*}^2 \;.$$

Examples

- Exponential model:

$$p_{\lambda}(x) = \lambda \, \mathrm{e}^{-\lambda \, x} \, \mathbb{1}_{\mathbb{R}_+}(x) \; , \quad \lambda > 0 \; .$$

- Gamma model

$$p_{\theta}(x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \, x^{\alpha-1} \mathrm{e}^{-\lambda \, x} \, \mathbb{1}_{\mathbb{R}_+}(x) \; , \quad \theta = (\alpha, \lambda) \in {\mathbb{R}_+^*}^2 \; .$$

- Beta model :

$$p_{\theta}(x) = B(\alpha, \beta) \, x^{\alpha-1} (1-x)^{\beta-1} \, \mathbb{1}_{[0,1]}(x) \; , \quad \theta = (\alpha, \beta) \in {\mathbb{R}_+^*}^2 \; .$$

- Gaussian model:

$$p_{\theta}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(x-\mu)^2/(2\sigma^2)} \; , \quad \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^* \; .$$

Multivariate data

- Most real life data is multivariate in the sense that it is doubly indexed, e.g.

$$\mathbf{X}_t = X_{i,t} \quad i = 1, \ldots, p \,,$$

  where t is the time index and i enumerates individuals, assets, features etc.

Multivariate data

- Most real life data is multivariate in the sense that it is doubly indexed, e.g.

$$\mathbf{X}_t = X_{i,t} \quad i = 1, \ldots, p \,,$$

  where t is the time index and i enumerates individuals, assets, features etc.
- Examples: portfolio returns, panel data (or longitudinal data), Risk indices, ...

Multivariate data

- Most real life data is multivariate in the sense that it is doubly indexed, e.g.

$$\mathbf{X}_t = X_{i,t} \quad i = 1, \ldots, p \,,$$

  where t is the time index and i enumerates individuals, assets, features etc.
- Examples: portfolio returns, panel data (or longitudinal data), Risk indices, ...
- To simplify the presentation, let us see the index i as a spatial index (as opposed to time index).

Multivariate data

- Most real life data is multivariate in the sense that it is doubly indexed, e.g.
$$\mathbf{X}_t = X_{i,t} \quad i = 1, \ldots, p \,,$$
where t is the time index and i enumerates individuals, assets, features etc.
- Examples: portfolio returns, panel data (or longitudinal data), Risk indices, ...
- To simplify the presentation, let us see the index i as a spatial index (as opposed to time index).
- A multivariate model will generally try to capture the *spatial* covariance structure through random vector models: e.g. Gaussian vectors, Ising model, or more general graphical models...

Example: i.i.d. Gaussian vectors

- Consider a portfolio of n asset returns $\mathbf{X}_t = X_{i,t} \quad i = 1, \ldots, p$.
- Suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where
  - $\boldsymbol{\mu} \in \mathbb{R}^p$ is the unknown mean.
  - $\Sigma \in \mathbb{R}^{p \times p}$ is the unknown covariance matrix
- Then the log-likelihood reads, for all $\theta = (\boldsymbol{\mu}, \Sigma)$,

$$
\begin{aligned}
L_n(\theta) &= \sum_{k=1}^{n} \log p_\theta(\mathbf{X}_k) \\
&= -\frac{1}{2n} \left(\log\det(2\pi\Sigma) + \sum_{k=1}^{n} (\mathbf{X}_k - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X}_k - \boldsymbol{\mu}) \right) .
\end{aligned}
$$

Example: i.i.d. Gaussian vectors, estimators

- Using a classical moment estimation method, we obtain the empirical estimators:
  - the empirical mean

$$\widehat{\mu}_{n,i} = \frac{1}{n} \sum_{t=1}^{n} \mathbf{X}_{i,t} \; .$$

  - the empirical covariance matrix

$$\widehat{\Sigma}_n[i,j] = \frac{1}{n} \sum_{t=1}^{n} (\mathbf{X}_{i,t} - \widehat{\mu}_{n,i})(\mathbf{X}_{j,t} - \widehat{\mu}_{n,j}) \; .$$

Example: i.i.d. Gaussian vectors, estimators

- Using a classical moment estimation method, we obtain the empirical estimators:
  - the empirical mean

$$\widehat{\mu}_{n,i} = \frac{1}{n} \sum_{t=1}^{n} \mathbf{X}_{i,t} \; .$$

  - the empirical covariance matrix

$$\widehat{\Sigma}_n[i,j] = \frac{1}{n} \sum_{t=1}^{n} (\mathbf{X}_{i,t} - \widehat{\mu}_{n,i})(\mathbf{X}_{j,t} - \widehat{\mu}_{n,j}) \; .$$

- Maximizing the likelihood yields to much more complicated estimators...

Example: i.i.d. Gaussian vectors, estimators

- Using a classical moment estimation method, we obtain the empirical estimators:
  - the empirical mean

$$\widehat{\mu}_{n,i} = \frac{1}{n} \sum_{t=1}^{n} \mathbf{X}_{i,t} \; .$$

  - the empirical covariance matrix

$$\widehat{\Sigma}_n[i,j] = \frac{1}{n} \sum_{t=1}^{n} (\mathbf{X}_{i,t} - \widehat{\mu}_{n,i})(\mathbf{X}_{j,t} - \widehat{\mu}_{n,j}) \; .$$

- Maximizing the likelihood yields to much more complicated estimators...
- In high dimension (p and n are of similar order), it is sometimes advantageous to make a sparse or low rank assumption.

Example: i.i.d. Gaussian vectors, estimators

- Using a classical moment estimation method, we obtain the empirical estimators:
  - the empirical mean
$$\widehat{\mu}_{n,i} = \frac{1}{n} \sum_{t=1}^{n} \mathbf{X}_{i,t} \; .$$
  - the empirical covariance matrix
$$\widehat{\Sigma}_n[i,j] = \frac{1}{n} \sum_{t=1}^{n} (\mathbf{X}_{i,t} - \widehat{\mu}_{n,i})(\mathbf{X}_{j,t} - \widehat{\mu}_{n,j}) \; .$$
- Maximizing the likelihood yields to much more complicated estimators...
- In high dimension (p and n are of similar order), it is sometimes advantageous to make a sparse or low rank assumption.
- From a regression perspective, it is easier to use sparsity of the precision matrix $M = \Sigma^{-1}$.

From bivariate distribution to conditional distribution

- In a regression model, each multivariate observation $\mathbf{X}_i$ is split into a pair of variables : $\mathbf{X}_i = (\mathbf{Z}_i, Y_i)$, where, usually, $\mathbf{Z}_i$ itself is multivariate, say valued in $\mathbb{R}^p$, and Y_i is univariate (discrete or continuous).

From bivariate distribution to conditional distribution

- In a regression model, each multivariate observation $\mathbf{X}_i$ is split into a pair of variables : $\mathbf{X}_i = (\mathbf{Z}_i, Y_i)$, where, usually, $\mathbf{Z}_i$ itself is multivariate, say valued in $\mathbb{R}^p$, and Y_i is univariate (discrete or continuous).
- Thus we can see the distribution of $\mathbf{X}_1$ as a bivariate distribution $\mathbb{P}_\theta^{\mathbf{X}_1} = \mathbb{P}_\theta^{(\mathbf{Z}_1, Y_1)}$.

From bivariate distribution to conditional distribution

- In a regression model, each multivariate observation $\mathbf{X}_i$ is split into a pair of variables : $\mathbf{X}_i = (\mathbf{Z}_i, Y_i)$, where, usually, $\mathbf{Z}_i$ itself is multivariate, say valued in $\mathbb{R}^p$, and Y_i is univariate (discrete or continuous).
- Thus we can see the distribution of $\mathbf{X}_1$ as a bivariate distribution $\mathbb{P}_\theta^{\mathbf{X}_1} = \mathbb{P}_\theta^{(\mathbf{Z}_1, Y_1)}$.
- Every bivariate distribution can be decomposed using
  - the marginal distribution of the first variable;
  - the conditional distribution of the second variable given the first variable.

From bivariate distribution to conditional distribution

- In a regression model, each multivariate observation $\mathbf{X}_i$ is split into a pair of variables : $\mathbf{X}_i = (\mathbf{Z}_i, Y_i)$, where, usually, $\mathbf{Z}_i$ itself is multivariate, say valued in $\mathbb{R}^p$, and Y_i is univariate (discrete or continuous).
- Thus we can see the distribution of $\mathbf{X}_1$ as a bivariate distribution $\mathbb{P}_\theta^{\mathbf{X}_1} = \mathbb{P}_\theta^{(\mathbf{Z}_1, Y_1)}$.
- Every bivariate distribution can be decomposed using
  - the marginal distribution of the first variable;
  - the conditional distribution of the second variable given the first variable.
- In a regression model, we see $\mathbf{Z}_i$ as an input (regression variable) and Y_i as an output (observation or response variable) and are only interested on the conditional distribution of the output given the input.

Likelihood of a regression model

- ▷ The decomposition of the bivariate distribution $\mathbb{P}_\theta^{\mathbf{X}_1} = \mathbb{P}_\theta^{(\mathbf{Z}_1, Y_1)}$ then yields

$$p_\theta(\mathbf{x}) = q(\mathbf{z}) p_\theta(y|\mathbf{z}) \;, \qquad \mathbf{x} = (\mathbf{z}, y) \;,$$

where $q(\mathbf{z})$ denotes the density of $\mathbf{Z}_1$ and $p_\theta(y|\mathbf{z})$ denotes the conditional density of Y_1 (or the conditional probability of $\mathbf{X}_1 = \mathbf{x}$) given $\mathbf{Z}_1 = \mathbf{z}$ under parameter θ.

Likelihood of a regression model

- The decomposition of the bivariate distribution $\mathbb{P}_\theta^{\mathbf{X}_1} = \mathbb{P}_\theta^{(\mathbf{Z}_1, Y_1)}$ then yields

$$p_\theta(\mathbf{x}) = q(\mathbf{z}) p_\theta(y|\mathbf{z}) \;, \qquad \mathbf{x} = (\mathbf{z}, y) \;,$$

  where $q(\mathbf{z})$ denotes the density of $\mathbf{Z}_1$ and $p_\theta(y|\mathbf{z})$ denotes the conditional density of Y_1 (or the conditional probability of $\mathbf{X}_1 = \mathbf{x}$) given $\mathbf{Z}_1 = \mathbf{z}$ under parameter θ.

- It follows that the log-likelihood takes the form, up to an additive constant:

$$\log L_n(\theta) = \sum_{k=1}^{n} \log p_\theta(Y_k | \mathbf{Z}_k)$$

Likelihood of a regression model

- The decomposition of the bivariate distribution $\mathbb{P}_\theta^{\mathbf{X}_1} = \mathbb{P}_\theta^{(\mathbf{Z}_1, Y_1)}$ then yields
$$p_\theta(\mathbf{x}) = q(\mathbf{z}) p_\theta(y|\mathbf{z}) \;, \qquad \mathbf{x} = (\mathbf{z}, y) \;,$$
where $q(\mathbf{z})$ denotes the density of $\mathbf{Z}_1$ and $p_\theta(y|\mathbf{z})$ denotes the conditional density of Y_1 (or the conditional probability of $\mathbf{X}_1 = \mathbf{x}$) given $\mathbf{Z}_1 = \mathbf{z}$ under parameter θ.
- It follows that the log-likelihood takes the form, up to an additive constant:
$$\log L_n(\theta) = \sum_{k=1}^{n} \log p_\theta(Y_k|\mathbf{Z}_k)$$
- Estimating θ allows one to propose a predictor of Y given a new input $\mathbf{Z}$, assuming that they are distributed according to the same bivariate distribution as the learning data set.

Two examples

- The linear regression model:

$$p_{\boldsymbol{\theta},\sigma^2}(y|\mathbf{z}) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(y-\boldsymbol{\theta}^T\mathbf{z})^2/(2\sigma^2)} , \quad (\boldsymbol{\theta},\sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^* , \quad y \in \mathbb{R} .$$

Optimizing the likelihood leads to the least mean square estimator.

Two examples

- The linear regression model:

$$p_{\boldsymbol{\theta},\sigma^2}(y|\mathbf{z}) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(y-\boldsymbol{\theta}^T\mathbf{z})^2/(2\sigma^2)} \;, \quad (\boldsymbol{\theta},\sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^* \;, \quad y \in \mathbb{R} \;.$$

  Optimizing the likelihood leads to the least mean square estimator.

- The logit regression model:

$$p_{\boldsymbol{\theta}}(y|\mathbf{z}) = \left(\frac{\mathrm{e}^{\boldsymbol{\theta}^T\mathbf{z}}}{1+\mathrm{e}^{\boldsymbol{\theta}^T\mathbf{z}}}\right)^y \left(\frac{1}{1+\mathrm{e}^{\boldsymbol{\theta}^T\mathbf{z}}}\right)^{1-y} \;, \quad \boldsymbol{\theta} \in \mathbb{R}^p \;, \quad y \in \{0,1\} \;.$$

The mixture model

- In certain cases, each (univariate or multivariate) observation $\mathbf{X}_i$ can be seen as a partial observation of a larger set of data. The unobserved part of the data is represented by a hidden variable V_i.

The mixture model

- In certain cases, each (univariate or multivariate) observation $\mathbf{X}_i$ can be seen as a partial observation of a larger set of data. The unobserved part of the data is represented by a hidden variable V_i.
- Again we can then decompose the bivariate distribution $\mathbb{P}_\theta^{(V_1,\mathbf{X}_1)}$ of the complete data $(V_1, \mathbf{X}_1)$ using
  - the marginal distribution of the hidden variable V_1;
  - the conditional distribution of the observed variable $\mathbf{X}_1$ given the hidden variable V_1.

The mixture model

- In certain cases, each (univariate or multivariate) observation $\mathbf{X}_i$ can be seen as a partial observation of a larger set of data. The unobserved part of the data is represented by a hidden variable V_i.
- Again we can then decompose the bivariate distribution $\mathbb{P}_\theta^{(V_1,\mathbf{X}_1)}$ of the complete data $(V_1, \mathbf{X}_1)$ using
  - the marginal distribution of the hidden variable V_1;
  - the conditional distribution of the observed variable $\mathbf{X}_1$ given the hidden variable V_1.
- The resulting marginal distribution $\mathbb{P}_\theta^{\mathbf{X}_1}$ is called a mixture model.

The mixture model

- In certain cases, each (univariate or multivariate) observation $\mathbf{X}_i$ can be seen as a partial observation of a larger set of data. The unobserved part of the data is represented by a hidden variable V_i.
- Again we can then decompose the bivariate distribution $\mathbb{P}_\theta^{(V_1,\mathbf{X}_1)}$ of the complete data $(V_1, \mathbf{X}_1)$ using
  - the marginal distribution of the hidden variable V_1;
  - the conditional distribution of the observed variable $\mathbf{X}_1$ given the hidden variable V_1.
- The resulting marginal distribution $\mathbb{P}_\theta^{\mathbf{X}_1}$ is called a mixture model.
- The simplest case is that of a finite mixture, where the hidden variable takes its values in a finite set $\{1, 2, \ldots, K\}$. This case amounts to see the data as being separated into K clusters, each of them following a different distribution, namely, the conditional distribution of $\mathbf{X}_1$ given $V_1 = k$, for $k = 1, 2, \ldots, K$.

The mixture model

- In certain cases, each (univariate or multivariate) observation $\mathbf{X}_i$ can be seen as a partial observation of a larger set of data. The unobserved part of the data is represented by a hidden variable V_i.
- Again we can then decompose the bivariate distribution $\mathbb{P}_\theta^{(V_1,\mathbf{X}_1)}$ of the complete data $(V_1, \mathbf{X}_1)$ using
  - the marginal distribution of the hidden variable V_1;
  - the conditional distribution of the observed variable $\mathbf{X}_1$ given the hidden variable V_1.
- The resulting marginal distribution $\mathbb{P}_\theta^{\mathbf{X}_1}$ is called a mixture model.
- The simplest case is that of a finite mixture, where the hidden variable takes its values in a finite set $\{1, 2, \ldots, K\}$. This case amounts to see the data as being separated into K clusters, each of them following a different distribution, namely, the conditional distribution of $\mathbf{X}_1$ given $V_1 = k$, for $k = 1, 2, \ldots, K$.
- A standard example of hidden variable for financial data is the (conditional) volatility.

Likelihood of a mixture model

- The natural decomposition of the bivariate distribution $\mathbb{P}_\theta^{(V_1,\mathbf{X}_1)}$ yields

$$p_\theta(v,\mathbf{x}) = q_\theta(v)p_\theta(\mathbf{x}|v) \ ,$$

where $q_\theta(v)$ denotes the density of V_1 (or the probability of $V_1 = v$) and $p_\theta(\mathbf{x}|v)$ denotes the conditional density of $\mathbf{X}_1$ (or the conditional probability of $\mathbf{X}_1 = \mathbf{x}$) given $V_1 = v$ under parameter θ.

Likelihood of a mixture model

- The natural decomposition of the bivariate distribution $\mathbb{P}_\theta^{(V_1, \mathbf{X}_1)}$ yields

$$p_\theta(v, \mathbf{x}) = q_\theta(v) p_\theta(\mathbf{x}|v) \ ,$$

  where $q_\theta(v)$ denotes the density of V_1 (or the probability of $V_1 = v$) and $p_\theta(\mathbf{x}|v)$ denotes the conditional density of $\mathbf{X}_1$ (or the conditional probability of $\mathbf{X}_1 = \mathbf{x}$) given $V_1 = v$ under parameter θ.
- It follows that the log-likelihood takes the form (in the case of continuous hidden variables):

$$\log L_n(\theta) = \sum_{k=1}^{n} \log \int q_\theta(v)\, p_\theta(\mathbf{X}_k|v)\, \mathrm{d}v \ .$$

Likelihood of a mixture model

- The natural decomposition of the bivariate distribution $\mathbb{P}_\theta^{(V_1,\mathbf{X}_1)}$ yields

$$p_\theta(v,\mathbf{x}) = q_\theta(v) p_\theta(\mathbf{x}|v) \;,$$

  where $q_\theta(v)$ denotes the density of V_1 (or the probability of $V_1 = v$) and $p_\theta(\mathbf{x}|v)$ denotes the conditional density of $\mathbf{X}_1$ (or the conditional probability of $\mathbf{X}_1 = \mathbf{x}$) given $V_1 = v$ under parameter θ.
- It follows that the log-likelihood takes the form (in the case of continuous hidden variables):

$$\log L_n(\theta) = \sum_{k=1}^{n} \log \int q_\theta(v)\, p_\theta(\mathbf{X}_k|v)\, \mathrm{d}v \;.$$

- For discrete mixtures, estimating θ allows one to clustering the data by identifying those who most likely share the same hidden variable.

Two examples

- Mixture of two Gaussian variables with parameter $\boldsymbol{\theta} = (\alpha, \mu_0, \mu_1, \sigma_0^2, \sigma_1^2) \in (0,1) \times \mathbb{R}^2 \times {\mathbb{R}_+^*}^2$: $V_1 \sim$Bernoulli(α) and given $V_1 = v$, $X_1 \sim \mathcal{N}(\mu_v, \sigma_v^2)$. Hence

$$q_{\boldsymbol{\theta}}(v) = \alpha^v (1-\alpha)^{1-v}$$
$$p_{\boldsymbol{\theta}}(x|v) = \left(2\pi\sigma_v^2\right)^{-1/2} \mathrm{e}^{-(x-\mu_v)^2/(2\sigma_v^2)} .$$

- Discrete mixture of Gaussian vectors with parameter $\boldsymbol{\theta} = (\alpha_k, \boldsymbol{\mu}_k, \Sigma_k)_{1 \le k \le K}$:

$$q_{\boldsymbol{\theta}}(v) = \alpha_v$$
$$p_{\boldsymbol{\theta}}(\mathbf{x}|v) = (\det(2\pi\Sigma_v))^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_v)^T \Sigma_v^{-1} (\mathbf{x}-\boldsymbol{\mu}_v)\right)$$

Optimizing the likelihood is a difficult question (related to the k-means algorithm).

Two examples (cont)

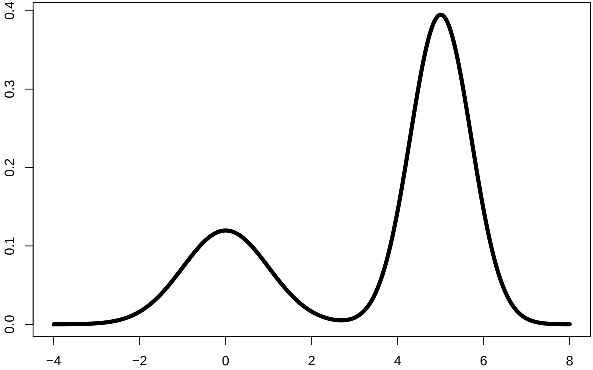


Figure: Density of the mixture of two Gaussian distributions

Two examples (cont)

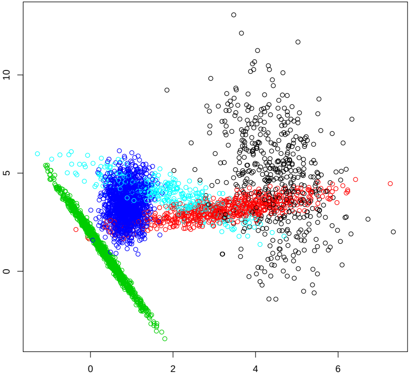


Figure: IID draws of the mixture of 5 bidimensional Gaussian distributions. Colors represent the (supposedly hidden) cluster variables.

Back to the USD vs EUR currency exchange rate.

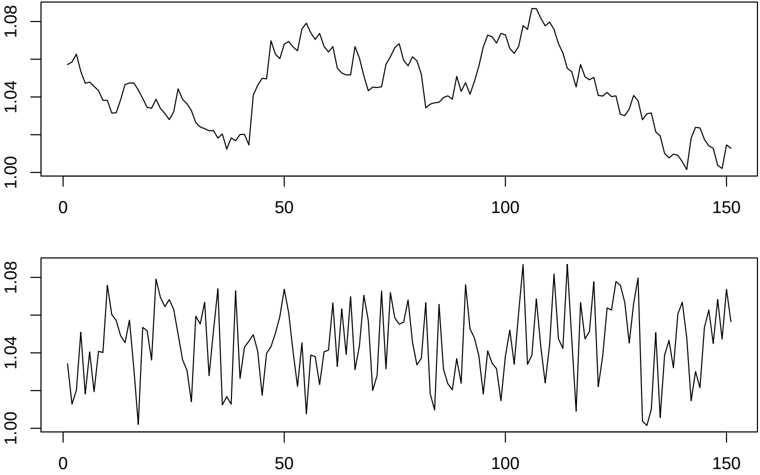


Figure: Top : price of 1 Euro in US Dollars between 1999-05-21 and 1999-12-17; Bottom : the same in randomly shuffled order.

Order of observations is not taken into account in i.i.d. models

- The log-likelihood of an i.i.d. model has the form

$$L_n(\theta) = \sum_{k=1}^{n} \log p_\theta(X_k) \ ,$$

  where $X_1, \ldots, X_n$ are the n observations, hence is invariant trough permutation of indices: $(X_1, \ldots, X_n) \mapsto (X_{\sigma(1)}, \ldots, X_{\sigma(n)})$, where $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ is a permutation.
- The two previous time series are the same up to a permutation of time indices.
- Hence they have the same likelihood for any i.i.d. model.

Some useful notation

- For any integers $k \geq l$ and sequence (x_t) we denote the subsample with indices between k and l by

$$x_{k:l} = (x_k, \ldots, x_l)$$

- If $(\mathbf{X}, \mathbf{Y})$ is valued in $\mathbb{R}^p \times \mathbb{R}^n$ and admits a density, we denote
  - by $p^{(\mathbf{X},\mathbf{Y})} : (x, y) \mapsto p^{(\mathbf{X},\mathbf{Y})}(x, y)$ the density of $(\mathbf{X}, \mathbf{Y})$,
  - by $p^{\mathbf{X}}$ the density of $\mathbf{X}$:

$$p^{\mathbf{X}}(\mathbf{x}) = \int_{\mathbb{R}^n} p^{(\mathbf{X},\mathbf{Y})}(\mathbf{x}, \mathbf{y})\,\mathrm{d}\mathbf{y} = \int \cdots \int p^{(\mathbf{X},\mathbf{Y})}(\mathbf{x}, y_{1:n})\,\mathrm{d}y_1 \ldots \mathrm{d}y_n \,.$$

  - by $p^{\mathbf{Y}|\mathbf{X}}(\cdot|x)$ the conditional density of $\mathbf{Y}$ given $\mathbf{X} = x$:

$$p^{\mathbf{Y}|\mathbf{X}}(y|x) = \frac{p^{(\mathbf{X},\mathbf{Y})}(x, y)}{p^{\mathbf{X}}(x)}$$

  - We add a subscript $_\theta$ if the density depends on the unknown parameter θ: $p_\theta^{(\mathbf{X},\mathbf{Y})}$, $p_\theta^{\mathbf{X}}$, $p_\theta^{\mathbf{Y}|\mathbf{X}}$ ...

General form of the likelihood

- How to generalize the product form of likelihood without the i.i.d. assumption ?

General form of the likelihood

- How to generalize the product form of likelihood without the i.i.d. assumption ?
- Suppose that $X_{1:n}$ admits a density $p_\theta^{X_{1:n}}$.

General form of the likelihood

- How to generalize the product form of likelihood without the i.i.d. assumption ?
- Suppose that $X_{1:n}$ admits a density $p_\theta^{X_{1:n}}$.
- Conditioning successively, we have

$$
\begin{aligned}
p_\theta^{X_{1:n}}(x_{1:n}) &= p_\theta^{X_n|X_{1:(n-1)}}(x_n|x_{1:n-1})p_\theta^{X_{1:n-1}}(x_{1:n}) \\
&\cdots \\
&= \prod_{k=2}^{n} p_\theta^{X_k|X_{1:(k-1)}}(x_k|x_{1:k-1})p_\theta^{X_1}(x_1) \;.
\end{aligned}
$$

General form of the likelihood

- How to generalize the product form of likelihood without the i.i.d. assumption ?
- Suppose that $X_{1:n}$ admits a density $p_\theta^{X_{1:n}}$.
- Conditioning successively, we have

$$
\begin{aligned}
p_\theta^{X_{1:n}}(x_{1:n}) &= p_\theta^{X_n|X_{1:(n-1)}}(x_n|x_{1:n-1})p_\theta^{X_{1:n-1}}(x_{1:n}) \\
&\dots \\
&= \prod_{k=2}^{n} p_\theta^{X_k|X_{1:(k-1)}}(x_k|x_{1:k-1})p_\theta^{X_1}(x_1) \; .
\end{aligned}
$$

- It is therefore of primary importance to understand the dynamics of the model through the conditional distribution of X_k given its past $X_{1:(k-1)}$.

Two important particular cases

- The i.i.d. case :

In this case, by independence of X_k and $X_{1:(k-1)}$, we have that $p_{\theta}^{X_k|X_{1:(k-1)}}(x_k|x_{1:k-1})$ does not depend on $x_{1:k-1}$, so that

$$p_{\theta}^{X_k|X_{1:(k-1)}}(x_k|x_{1:k-1}) = p_{\theta}^{X_k}(x_k) \; .$$

And, by the "i.d." property,

$$p_{\theta}^{X_k|X_{1:(k-1)}}(x_k|x_{1:k-1}) = p_{\theta}^{X_k}(x_k) = p_{\theta}(x_k) \; ,$$

where p_{θ} is the common density of all X_k's.

Two important particular cases (cont.)

- The homogeneous Markov case :

In this case, we have that $p_{\theta}^{X_k|X_{1:(k-1)}}(x_k|x_{1:k-1})$ only depends on x_{k-1}, so that

$$p_{\theta}^{X_k|X_{1:(k-1)}}(x_k|x_{1:k-1}) = p_{\theta}^{X_k|X_{k-1}}(x_k|x_{k-1}) \;.$$

And "homogeneous" means that $p_{\theta}^{X_k|X_{k-1}}$ does not depend on k and is given by a common conditional density, say $q_{\theta}(\cdot|\cdot)$, hence

$$p_{\theta}^{X_k|X_{1:(k-1)}}(x_k|x_{1:k-1}) = p_{\theta}^{X_k|X_{k-1}}(x_k|x_{k-1}) = q_{\theta}(x_k|x_{k-1}) \;.$$

Graphical representation of a homogeneous Markov chain

$$\cdots \dashrightarrow X_{k-1} \xrightarrow{q_\theta} X_k \xrightarrow{q_\theta} X_{k+1} \dashrightarrow \cdots$$

- Arrows indicate the dependence structure: given all other variables, a child can be generated using only its own parents.
- Here, each child only has 1 parent: the generation of the child is carried out through the conditional density q_θ.

Examples of conditional density

An homoscedastic model : AR(1).

In this case, $q_\theta(\cdot|x)$ is the density of $\mathcal{N}(\phi x, \sigma^2)$, with
$\theta = (\phi, \sigma^2) \in (-1, 1) \times \mathbb{R}_+^*$.
Equivalently, this model is given by the dynamical equation

$$X_k = \phi X_{k-1} + \epsilon_k \ ,$$

with $(\epsilon_t)_{t \in \mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$.

Examples of conditional density

An homoscedastic model : AR(1).

In this case, $q_{\theta}(\cdot|x)$ is the density of $\mathcal{N}(\phi x, \sigma^2)$, with $\theta = (\phi, \sigma^2) \in (-1, 1) \times \mathbb{R}_+^*$.
Equivalently, this model is given by the dynamical equation

$$X_k = \phi X_{k-1} + \epsilon_k \ ,$$

with $(\epsilon_t)_{t \in \mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$.

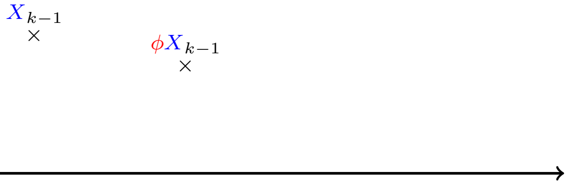
Examples of conditional density

An homoscedastic model : AR(1).

In this case, $q_\theta(\cdot|x)$ is the density of $\mathcal{N}(\phi x, \sigma^2)$, with $\theta = (\phi, \sigma^2) \in (-1, 1) \times \mathbb{R}_+^*$.
Equivalently, this model is given by the dynamical equation

$$X_k = \phi X_{k-1} + \epsilon_k \ ,$$

with $(\epsilon_t)_{t \in \mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$.



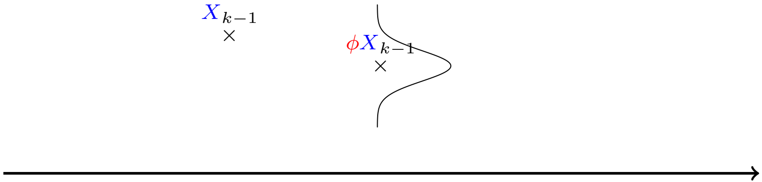
Examples of conditional density

An homoscedastic model : AR(1).

In this case, $q_\theta(\cdot|x)$ is the density of $\mathcal{N}(\phi x, \sigma^2)$, with $\theta = (\phi, \sigma^2) \in (-1, 1) \times \mathbb{R}_+^*$.
Equivalently, this model is given by the dynamical equation

$$X_k = \phi X_{k-1} + \epsilon_k \ ,$$

with $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$.



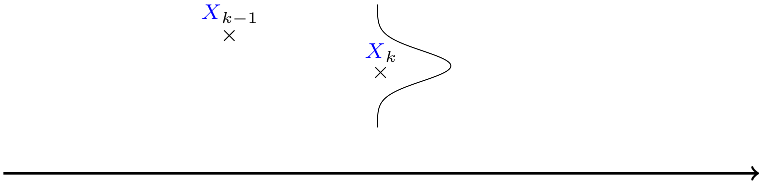
Examples of conditional density

An homoscedastic model : AR(1).

In this case, $q_\theta(\cdot|x)$ is the density of $\mathcal{N}(\phi x, \sigma^2)$, with $\theta = (\phi, \sigma^2) \in (-1, 1) \times \mathbb{R}_+^*$.
Equivalently, this model is given by the dynamical equation

$$X_k = \phi X_{k-1} + \epsilon_k \ ,$$

with $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$.



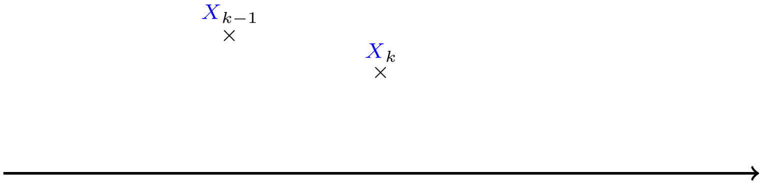
Examples of conditional density

An homoscedastic model : AR(1).

In this case, $q_{\theta}(\cdot|x)$ is the density of $\mathcal{N}(\phi x, \sigma^2)$, with $\theta = (\phi, \sigma^2) \in (-1, 1) \times \mathbb{R}_+^*$.
Equivalently, this model is given by the dynamical equation

$$X_k = \phi X_{k-1} + \epsilon_k \ ,$$

with $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$.



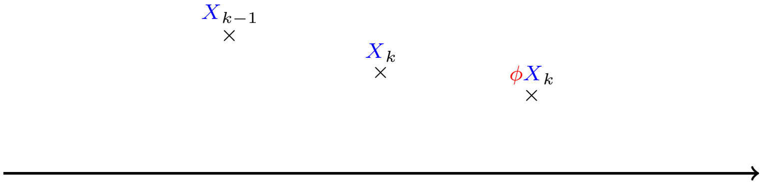
Examples of conditional density

An homoscedastic model : AR(1).

In this case, $q_\theta(\cdot|x)$ is the density of $\mathcal{N}(\phi x, \sigma^2)$, with
$\theta = (\phi, \sigma^2) \in (-1, 1) \times \mathbb{R}_+^*$.
Equivalently, this model is given by the dynamical equation

$$X_k = \phi X_{k-1} + \epsilon_k \ ,$$

with $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$.



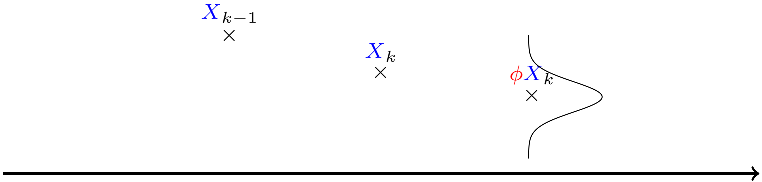
Examples of conditional density

An homoscedastic model : AR(1).

In this case, $q_{\theta}(\cdot|x)$ is the density of $\mathcal{N}(\phi x, \sigma^2)$, with $\theta = (\phi, \sigma^2) \in (-1, 1) \times \mathbb{R}_+^*$.
Equivalently, this model is given by the dynamical equation

$$X_k = \phi X_{k-1} + \epsilon_k \ ,$$

with $(\epsilon_t)_{t \in \mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$.



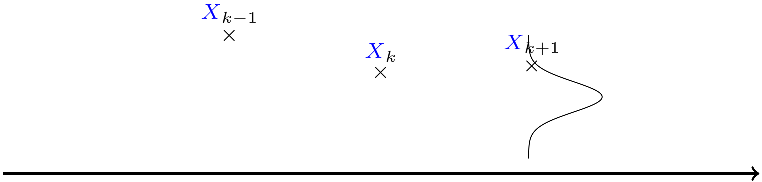
Examples of conditional density

An homoscedastic model : AR(1).

In this case, $q_{\theta}(\cdot|x)$ is the density of $\mathcal{N}(\phi x, \sigma^2)$, with $\theta = (\phi, \sigma^2) \in (-1, 1) \times \mathbb{R}_+^*$.
Equivalently, this model is given by the dynamical equation

$$X_k = \phi X_{k-1} + \epsilon_k \ ,$$

with $(\epsilon_t)_{t \in \mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$.



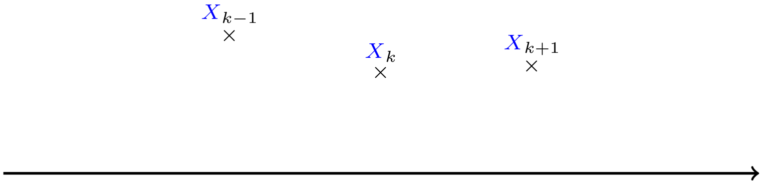
Examples of conditional density

An homoscedastic model : AR(1).

In this case, $q_{\theta}(\cdot|x)$ is the density of $\mathcal{N}(\phi x, \sigma^2)$, with
$\theta = (\phi, \sigma^2) \in (-1, 1) \times \mathbb{R}_+^*$.
Equivalently, this model is given by the dynamical equation

$$X_k = \phi X_{k-1} + \epsilon_k \; ,$$

with $(\epsilon_t)_{t \in \mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$.



Examples of conditional density (cont.)

An heteroscedastic model : ARCH(1).

In this case, $q_\theta(\cdot|x)$ is the density of $\mathcal{N}(0, a + bx^2)$, with $\theta = (a, b) \in \mathbb{R}_+^* \times \mathbb{R}_+$.
Equivalently, this model is given by the dynamical equation

$$X_k = \sqrt{a + bX_{k-1}^2}\, \epsilon_k \;,$$

with $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, 1)$.

Examples of conditional density (cont.)

An heteroscedastic model : ARCH(1).

In this case, $q_\theta(\cdot|x)$ is the density of $\mathcal{N}(0, a + bx^2)$, with $\theta = (a, b) \in \mathbb{R}_+^* \times \mathbb{R}_+$.

Equivalently, this model is given by the dynamical equation

$$X_k = \sqrt{a + bX_{k-1}^2}\, \epsilon_k ,$$

with $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, 1)$.

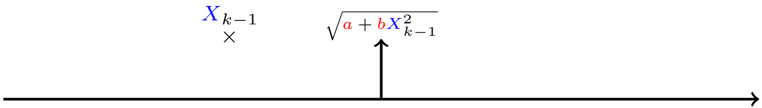
Examples of conditional density (cont.)

An heteroscedastic model : ARCH(1).

In this case, $q_\theta(\cdot|x)$ is the density of $\mathcal{N}(0, a + bx^2)$, with $\theta = (a, b) \in \mathbb{R}_+^* \times \mathbb{R}_+$.
Equivalently, this model is given by the dynamical equation

$$X_k = \sqrt{a + bX_{k-1}^2}\, \epsilon_k \ ,$$

with $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, 1)$.



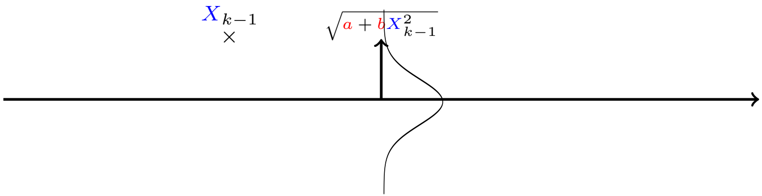
Examples of conditional density (cont.)

An heteroscedastic model : ARCH(1).

In this case, $q_\theta(\cdot|x)$ is the density of $\mathcal{N}(0, a + bx^2)$, with $\theta = (a, b) \in \mathbb{R}_+^* \times \mathbb{R}_+$.
Equivalently, this model is given by the dynamical equation

$$X_k = \sqrt{a + bX_{k-1}^2}\,\epsilon_k \ ,$$

with $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0,1)$.



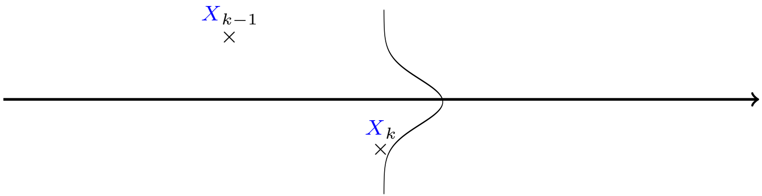
Examples of conditional density (cont.)

An heteroscedastic model : ARCH(1).

In this case, $q_\theta(\cdot|x)$ is the density of $\mathcal{N}(0, a + bx^2)$, with $\theta = (a, b) \in \mathbb{R}_+^* \times \mathbb{R}_+$.
Equivalently, this model is given by the dynamical equation

$$X_k = \sqrt{a + bX_{k-1}^2}\, \epsilon_k \ ,$$

with $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, 1)$.



Examples of conditional density (cont.)

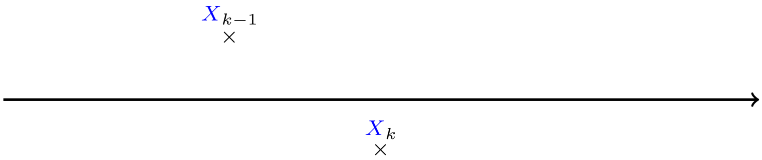
An heteroscedastic model : ARCH(1).

In this case, $q_\theta(\cdot|x)$ is the density of $\mathcal{N}(0, a + bx^2)$, with $\theta = (a, b) \in \mathbb{R}_+^* \times \mathbb{R}_+$.

Equivalently, this model is given by the dynamical equation

$$X_k = \sqrt{a + bX_{k-1}^2}\,\epsilon_k \ ,$$

with $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, 1)$.



Examples of conditional density (cont.)

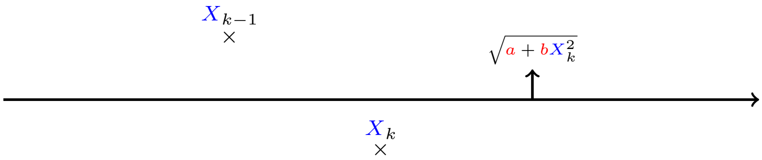
An heteroscedastic model : ARCH(1).

In this case, $q_\theta(\cdot|x)$ is the density of $\mathcal{N}(0, a + bx^2)$, with $\theta = (a, b) \in \mathbb{R}_+^* \times \mathbb{R}_+$.

Equivalently, this model is given by the dynamical equation

$$X_k = \sqrt{a + bX_{k-1}^2}\, \epsilon_k \ ,$$

with $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, 1)$.



Examples of conditional density (cont.)

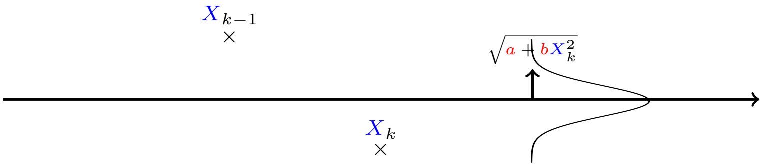
An heteroscedastic model : ARCH(1).

In this case, $q_\theta(\cdot|x)$ is the density of $\mathcal{N}(0, a + bx^2)$, with $\theta = (a, b) \in \mathbb{R}_+^* \times \mathbb{R}_+$.

Equivalently, this model is given by the dynamical equation

$$X_k = \sqrt{a + bX_{k-1}^2}\,\epsilon_k \ ,$$

with $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, 1)$.



Examples of conditional density (cont.)

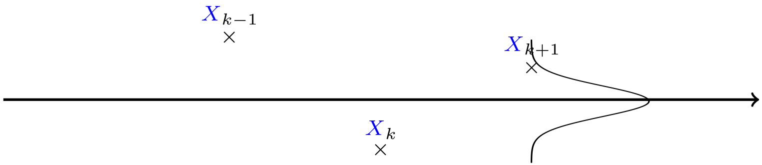
An heteroscedastic model : ARCH(1).

In this case, $q_\theta(\cdot|x)$ is the density of $\mathcal{N}(0, a + bx^2)$, with $\theta = (a, b) \in \mathbb{R}_+^* \times \mathbb{R}_+$.

Equivalently, this model is given by the dynamical equation

$$X_k = \sqrt{a + bX_{k-1}^2}\, \epsilon_k \ ,$$

with $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, 1)$.



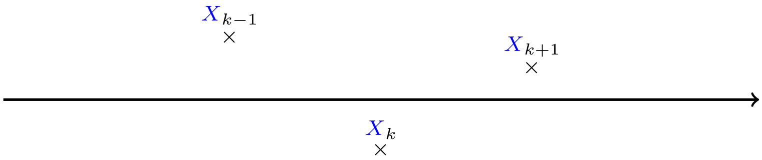
Examples of conditional density (cont.)

An heteroscedastic model : ARCH(1).

In this case, $q_\theta(\cdot|x)$ is the density of $\mathcal{N}(0, a + bx^2)$, with $\theta = (a, b) \in \mathbb{R}_+^* \times \mathbb{R}_+$.
Equivalently, this model is given by the dynamical equation

$$X_k = \sqrt{a + bX_{k-1}^2}\,\epsilon_k ,$$

with $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, 1)$.



Likelihood for univariate time series

- The simplest model beyond the i.i.d. model that contains some non-trivial dynamics is the homogeneous Markov model.

Likelihood for univariate time series

- The simplest model beyond the i.i.d. model that contains some non-trivial dynamics is the homogeneous Markov model.
- Its log-likelihood takes the form

$$L_n(\theta) = \sum_{k=2}^{n} \log q_\theta(X_k | X_{k-1}) + \log p^{X_1}(X_1) \ .$$

Likelihood for univariate time series

- The simplest model beyond the i.i.d. model that contains some non-trivial dynamics is the homogeneous Markov model.
- Its conditional log-likelihood takes the form

$$L_n(\theta) = \sum_{k=2}^{n} \log q_\theta(X_k|X_{k-1}) + \cancel{\log p^{X_1}(X_1)} .$$

Likelihood for univariate time series

- The simplest model beyond the i.i.d. model that contains some non-trivial dynamics is the homogeneous Markov model.
- Its conditional log-likelihood takes the form

$$L_n(\theta) = \sum_{k=2}^{n} \log q_\theta(X_k | X_{k-1}) + \cancel{\log p^{X_1}(X_1)} .$$

- The likelihood is no longer invariant by permutation.

Exemple: likelihood of the Gaussian AR(1) model

Consider the ▸ AR(1) model. Then we have

$$q_{\theta}(x_k|x_{k-1}) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{e}^{-(x_k-\phi x_{k-1})^2/(2\sigma^2)}\;.$$

It follows that the (conditional) negated log likelihood reads

$$-L_n(\theta) = \frac{n-1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{k=2}^{n}(X_k - \phi X_{k-1})^2\;,$$

which leads to the estimators

$$\widehat{\phi}_n = \frac{\sum_{k=2}^{n} X_{k-1}X_k}{\sum_{k=2}^{n} X_{k-1}^2} \quad \text{and} \quad \widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{k=2}^{n}(X_k - \widehat{\phi}_n X_{k-1})^2\;.$$

Exemple: likelihood of the conditionally Gaussian ARCH(1) model

Consider the ▸ ARCH(1) model. Then we have

$$q_\theta(x_k|x_{k-1}) = \frac{1}{\sqrt{2\pi(a + bx_{k-1}^2)}} \, \mathrm{e}^{-x_k^2/(2(a+bx_{k-1}^2))} \; .$$

It follows that the (conditional) negated log likelihood reads

$$-L_n(\theta) = \frac{1}{2} \sum_{k=2}^{n} \left(\log(2\pi(a + bX_{k-1}^2)) + \frac{X_k^2}{a + bX_{k-1}^2} \right) \; ,$$

which can be minimized in $\theta = (a, b)$ using a gradient descent algorithm.

Multivariate time series

- Exactly as in the IID case, a time series $(\mathbf{X}_t)$ can be multivariate, i.e. $\mathbf{X}_t$ is valued in $\mathbb{R}^p$ for some $p \geq 2$.

Multivariate time series

- Exactly as in the IID case, a time series $(\mathbf{X}_t)$ can be multivariate, i.e. $\mathbf{X}_t$ is valued in $\mathbb{R}^p$ for some $p \geq 2$.
- For instance, under the homogeneous Markov chain assumption, the (conditional) likelihood then reads

$$L_n(\theta) = \sum_{k=2}^{n} \log q_\theta(\mathbf{X}_k|\mathbf{X}_{k-1}) \; .$$

Multivariate time series

- Exactly as in the IID case, a time series $(\mathbf{X}_t)$ can be multivariate, i.e. $\mathbf{X}_t$ is valued in $\mathbb{R}^p$ for some $p \geq 2$.
- For instance, under the homogeneous Markov chain assumption, the (conditional) likelihood then reads

$$L_n(\theta) = \sum_{k=2}^{n} \log q_\theta(\mathbf{X}_k|\mathbf{X}_{k-1}) .$$

- In particular, consider a univariate p-order Markov time series with log likelihood

$$L_n(\theta) = \sum_{k=p+2}^{n} \log q_\theta(X_k|X_{k-p:k-1}) .$$

  To obtain a multivariate (first order) Markov time series, one can set $\mathbf{X}_k = X_{k-p+1:k}$.

Exemple of Multivariate time series: AR(p) time series

An AR(p) time series (X_t) satisfies the AR(p) equation

$$X_t = \sum_{k=1}^{p} \phi_k X_{t-k} + \epsilon_t , \qquad t \in \mathbb{Z} .$$

Setting $\mathbf{X}_k = \begin{bmatrix} X_k & X_{k-1} & \dots & X_{k-p+1} \end{bmatrix}^T$, this leads to the vector AR(1) equation:

$$\mathbf{X}_t = \Phi \mathbf{X}_{t-1} + \boldsymbol{\epsilon}_t , \qquad t \in \mathbb{Z} .$$

where

$$\Phi = \begin{bmatrix} \phi_1 & \phi_2 & \dots & \phi_p \\ 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \dots & 1 & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon}_t = \begin{bmatrix} \epsilon_t \\ 0 \vdots \\ 0 \end{bmatrix} .$$

Exemple of Multivariate time series: general bivariate case

Consider the bivariate case $\mathbf{X}_t = (X_t(1), X_t(2))$.

- ▷ IID case

$$\begin{tikzcd} \cdots & X_k(1) \arrow[d, leftrightarrow] & X_{k+1}(1) \arrow[d, leftrightarrow] & \cdots \\ \cdots & X_k(2) & X_{k+1}(2) & \cdots \end{tikzcd}$$

Exemple of Multivariate time series: general bivariate case

Consider the bivariate case $\mathbf{X}_t = (X_t(1), X_t(2))$.

- IID case

$$\begin{tikzcd}
\cdots & X_k(1) \arrow[d, leftrightarrow] & X_{k+1}(1) \arrow[d, leftrightarrow] & \cdots \\
\cdots & X_k(2) & X_{k+1}(2) & \cdots
\end{tikzcd}$$

- Markov case:

$$\begin{tikzcd}
\cdots \arrow[r, dashed] \arrow[dr, dashed] & X_k(1) \arrow[r] \arrow[d, leftrightarrow] \arrow[dr] & X_{k+1}(1) \arrow[r, dashed] \arrow[d, leftrightarrow] \arrow[dr, dashed] & \cdots \\
\cdots \arrow[r, dashed] \arrow[ur, dashed] & X_k(2) \arrow[r] \arrow[ur] & X_{k+1}(2) \arrow[r, dashed] \arrow[ur, dashed] & \cdots
\end{tikzcd}$$

Partially observed multivariate time series

- Exactly as in the IID case, adding hidden variables allows one to build a wild variety of models, while allowing the practitioner to provide intuitive interpretations of the model.

Partially observed multivariate time series

- Exactly as in the IID case, adding hidden variables allows one to build a wild variety of models, while allowing the practitioner to provide intuitive interpretations of the model.
- The most widely used such time series model is the linear state-space model, or dynamic linear model, defined through two linear equations

$$\mathbf{X}_t = \Phi \mathbf{X}_{t-1} + \mathbf{U}_t \quad \text{(State Equation)} \tag{1a}$$

$$\mathbf{Y}_t = A \mathbf{X}_t + \mathbf{V}_t \quad \text{(Observation Equation)}\ , \tag{1b}$$

where $(\mathbf{Y}_t)$ is the observed time series, and $(\mathbf{X}_t)$ is the hidden time series (also called the state variables), and $(\mathbf{U}_t)$ and $(\mathbf{V}_t)$ are IID noise sequences.

Partially observed multivariate time series

- Exactly as in the IID case, adding hidden variables allows one to build a wild variety of models, while allowing the practitioner to provide intuitive interpretations of the model.
- The most widely used such time series model is the linear state-space model, or dynamic linear model, defined through two linear equations

$$\mathbf{X}_t = \Phi \mathbf{X}_{t-1} + \mathbf{U}_t \quad \text{(State Equation)} \tag{1a}$$
$$\mathbf{Y}_t = A \mathbf{X}_t + \mathbf{V}_t \quad \text{(Observation Equation)}\,, \tag{1b}$$

  where $(\mathbf{Y}_t)$ is the observed time series, and $(\mathbf{X}_t)$ is the hidden time series (also called the state variables), and $(\mathbf{U}_t)$ and $(\mathbf{V}_t)$ are IID noise sequences.
- This is a articular instance of the general class of the partially observed Markov models, where one has a bivariate Markov chain $((\mathbf{X}_t, \mathbf{Y}_t))$, where only the component (Y_t) is observed.

Examples of partially observed multivariate time series

- IID case

$$
\begin{tikzcd}
\cdots & \mathbf{X}_k \arrow[d] & \mathbf{X}_{k+1} \arrow[d] & \cdots \\
\cdots & Y_k & Y_{k+1} & \cdots
\end{tikzcd}
$$

Examples of partially observed multivariate time series

- IID case

$$
\begin{tikzcd}
\cdots & \mathbf{X}_k \arrow[d] & \mathbf{X}_{k+1} \arrow[d] & \cdots \\
\cdots & Y_k & Y_{k+1} & \cdots
\end{tikzcd}
$$

- Partially observed Markov model: general case.

$$
\begin{tikzcd}
\cdots \arrow[r, dashed] \arrow[d, dashed] \arrow[dr, dashed] & X_k \arrow[r] \arrow[d, dashed] \arrow[dr] & X_{k+1} \arrow[r, dashed] \arrow[d] \arrow[dr, dashed] & \cdots \arrow[d, dashed] \\
\cdots \arrow[r, dashed] \arrow[ur, dashed] & Y_k \arrow[r] \arrow[ur] & Y_{k+1} \arrow[r, dashed] \arrow[ur, dashed] & \cdots
\end{tikzcd}
$$

Examples of partially observed multivariate time series (cont.)

- Hidden Markov model.

$$
\begin{tikzcd}
\cdots \arrow[r, dashed] \arrow[d, dashed] & X_k \arrow[r] \arrow[d, dashed] & X_{k+1} \arrow[r, dashed] \arrow[d] & \cdots \arrow[d, dashed] \\
\cdots & Y_k & Y_{k+1} & \cdots
\end{tikzcd}
$$

In this special case:

- (X_t) alone is a Markov chain.

Examples of partially observed multivariate time series (cont.)

- Hidden Markov model.

$$
\begin{tikzcd}
\cdots \arrow[r, dashed] \arrow[d, dashed] & X_k \arrow[r] \arrow[d, dashed] & X_{k+1} \arrow[r, dashed] \arrow[d] & \cdots \arrow[d, dashed] \\
\cdots & Y_k & Y_{k+1} & \cdots
\end{tikzcd}
$$

In this special case:

- (X_t) alone is a Markov chain.
- Given (X_t), the observations (Y_t) are conditionally independent.

Examples of partially observed multivariate time series (cont.)

- Hidden Markov model.

$$
\begin{tikzcd}
\cdots \arrow[r, dashed] \arrow[d, dashed] & X_k \arrow[r] \arrow[d, dashed] & X_{k+1} \arrow[r, dashed] \arrow[d] & \cdots \arrow[d, dashed] \\
\cdots & Y_k & Y_{k+1} & \cdots
\end{tikzcd}
$$

  In this special case:

  - (X_t) alone is a Markov chain.
  - Given (X_t), the observations (Y_t) are conditionally independent.

- Two highly popular special cases:
  - HMM with finite state space : when X_t takes values in $\{1, \ldots, K\}$.
  - The dynamic linear model, see (1).

Example : an HMM with two hidden states.

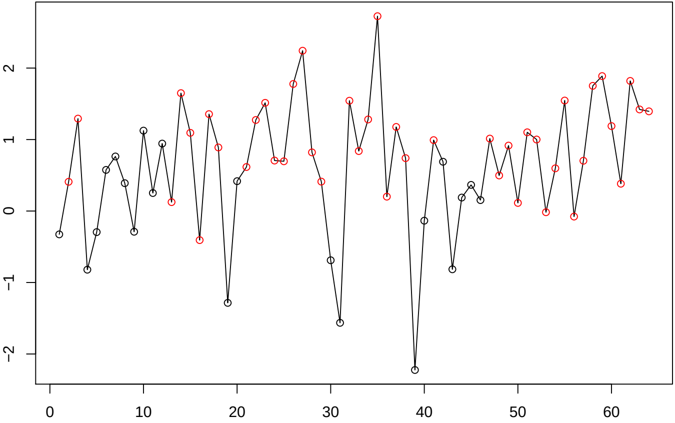


Figure: An HMM with two (supposedly) hidden states (red and black).

Example : Noisy observations of an hidden AR(1) state variables.

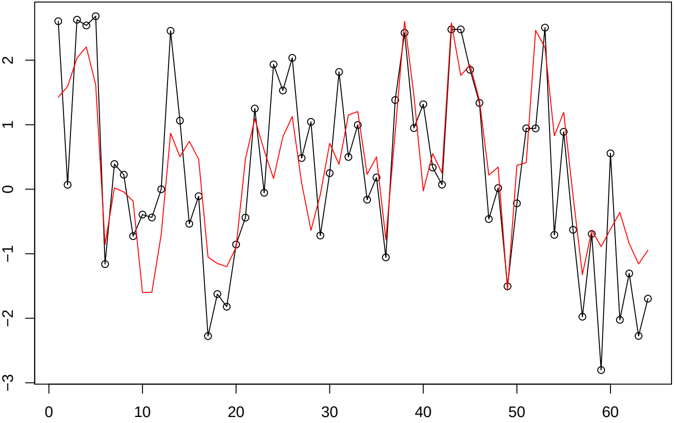


Figure: Observations (black 'o') obtained by adding noise to a (supposedly) hidden AR(1) process (red lines).

Observation driven models

- For most of the partially observed Markov models, there are no closed form formula for the likelihood and computational cost of L_n can be very high as n increases.
- Observation driven models stand as a popular exception. Their dependence structure takes the following form:

$$\begin{tikzcd}
\cdots \arrow[r, dashed] \arrow[d, dashed] & X_k \arrow[r] \arrow[d, dashed] & X_{k+1} \arrow[r, dashed] \arrow[d] & \cdots \arrow[d, dashed] \\
\cdots \arrow[ur, dashed] & Y_k \arrow[ur] & Y_{k+1} \arrow[ur, dashed] & \cdots
\end{tikzcd}$$

With the additional property that the conditional distribution of X_{k+1} given (X_k, Y_k) is degenerate.

Exemple: GARCH(1,1) model

GARCH(1,1) model

For parameter $\theta = (a, b, c) \in (0, \infty)^3$, (Y_t) satisfies the GARCH(1,1) equation

$$\sigma_t^2 = a + b\,Y_{t-1} + c\,\sigma_{t-1}^2 \tag{2a}$$
$$Y_t = \sigma_t \epsilon_t \;, \tag{2b}$$

where $(\epsilon_t)_{t \in \mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0, 1)$.
Moreover it is assumed that (σ_t) is non-anticipative solution, in the sense that, for all $t \in \mathbb{Z}$, σ_t only depends on $(\epsilon_s)_{s<t}$

Exemple: GARCH(1,1) model

GARCH(1,1) model

For parameter $\theta = (a, b, c) \in (0, \infty)^3$, (Y_t) satisfies the GARCH(1,1) equation

$$\sigma_t^2 = a + b\,Y_{t-1} + c\,\sigma_{t-1}^2 \tag{2a}$$
$$Y_t = \sigma_t \epsilon_t \ , \tag{2b}$$

where $(\epsilon_t)_{t\in\mathbb{Z}}$ i.i.d. $\sim \mathcal{N}(0,1)$.
Moreover it is assumed that (σ_t) is non-anticipative solution, in the sense that, for all $t \in \mathbb{Z}$, σ_t only depends on $(\epsilon_s)_{s<t}$

The fact that (σ_t) is non-anticipative ensures that, for all $t \in \mathbb{Z}$, given $(\epsilon_s)_{s<t}$, the conditional distribution of Y_t is $\mathcal{N}(0, \sigma_t)$.

Exemple: GARCH(1,1) model, likelihood

Iterating (2a) with a given θ, for all $k = 2, \dots, n$, one can express σ_k^2 as a deterministic function of $Y_{1:k-1}$ and σ_1^2, say

$$\sigma_k^2 = \psi^\theta < Y_{1:k-1} > (\sigma_1^2). \tag{3}$$

Note that $\psi^\theta < Y_{1:k-1} > (\sigma_1^2)$ is easy to compute iteratively.

Exemple: GARCH(1,1) model, likelihood

Iterating (2a) with a given θ, for all $k = 2, \ldots, n$, one can express σ_k^2 as a deterministic function of $Y_{1:k-1}$ and σ_1^2, say

$$\sigma_k^2 = \psi^\theta < Y_{1:k-1} > (\sigma_1^2). \tag{3}$$

Note that $\psi^\theta < Y_{1:k-1} > (\sigma_1^2)$ is easy to compute iteratively.

Using (3) and (2b), the (conditional) negated log likelihood (given $\sigma_1^2 = s_1^2$ and Y_1 for some arbitrary s_1^2) is given by

$$-L_n(\theta) = \frac{1}{2} \sum_{k=2}^{n} \left(\log \left(2\pi \, \psi^\theta < Y_{1:k-1} > (s_1^2) \right) + \frac{Y_k^2}{\psi^\theta < Y_{1:k-1} > (s_1^2)} \right) ,$$

which can be minimized in $\theta = (a, b, c)$ using a gradient descent algorithm.

Basic (important) definitions

Definition : Data set

A data set is a collection of values, say $X_{1:n} = X_1, \ldots, X_n$. Time series data sets are usually sampled from recorded measurements.

Basic (important) definitions

Definition : Data set

A data set is a collection of values, say $X_{1:n} = X_1, \ldots, X_n$. Time series data sets are usually sampled from recorded measurements.

Definition : Model

A model is a collection of probability distributions. The data set is assumed to be distributed according to one of them.

Basic (important) definitions

Definition : Data set

A data set is a collection of values, say $X_{1:n} = X_1, \ldots, X_n$. Time series data sets are usually sampled from recorded measurements.

Definition : Model

A model is a collection of probability distributions. The data set is assumed to be distributed according to one of them.

Definition : Statistic

A statistic is any value which can be computed from the data.

Raw data set

A time series $X_1, \ldots, X_n$ is usually presented as

Raw data set

A time series $X_1, \ldots, X_n$ is usually presented as

- a list of real values $X_1, \ldots, X_n$ in a data or spreadsheet file,

Raw data set

A time series $X_1, \ldots, X_n$ is usually presented as

- a list of real values $X_1, \ldots, X_n$ in a data or spreadsheet file,
- a corresponding list of dates (days, years, seconds...)
- or, equivalently, a starting date (in some unit), and a frequency.

Raw data set

A time series $X_1, \ldots, X_n$ is usually presented as

- a list of real values $X_1, \ldots, X_n$ in a data or spreadsheet file,
- a corresponding list of dates (days, years, seconds...)
- or, equivalently, a starting date (in some unit), and a frequency. For instance :
  - a date in years and frequency$= 12$ corresponds to monthly data,

Raw data set

A time series $X_1, \ldots, X_n$ is usually presented as

- a list of real values $X_1, \ldots, X_n$ in a data or spreadsheet file,
- a corresponding list of dates (days, years, seconds...)
- or, equivalently, a starting date (in some unit), and a frequency. For instance :
  - a date in years and frequency$= 12$ corresponds to monthly data,
  - a date in years and frequency$= 4$ corresponds to quarterly data,

Raw data set

A time series $X_1, \ldots, X_n$ is usually presented as

- a list of real values $X_1, \ldots, X_n$ in a data or spreadsheet file,
- a corresponding list of dates (days, years, seconds...)
- or, equivalently, a starting date (in some unit), and a frequency. For instance :
  - a date in years and frequency$= 12$ corresponds to monthly data,
  - a date in years and frequency$= 4$ corresponds to quarterly data,
  - a date in days and frequency$= 1$ corresponds to daily data,
  - $\vdots$

Raw data set

A time series $X_1, \ldots, X_n$ is usually presented as

- a list of real values $X_1, \ldots, X_n$ in a data or spreadsheet file,
- a corresponding list of dates (days, years, seconds...)
- or, equivalently, a starting date (in some unit), and a frequency. For instance :
  - a date in years and frequency$= 12$ corresponds to monthly data,
  - a date in years and frequency$= 4$ corresponds to quarterly data,
  - a date in days and frequency$= 1$ corresponds to daily data,
  - $\vdots$
- Remarks :
  - There may be missing values (usually expressed as 'NA')

Raw data set

A time series $X_1, \ldots, X_n$ is usually presented as

- a list of real values $X_1, \ldots, X_n$ in a data or spreadsheet file,
- a corresponding list of dates (days, years, seconds...)
- or, equivalently, a starting date (in some unit), and a frequency. For instance :
  - a date in years and frequency$= 12$ corresponds to monthly data,
  - a date in years and frequency$= 4$ corresponds to quarterly data,
  - a date in days and frequency$= 1$ corresponds to daily data,
  - $\vdots$
- Remarks :
  - There may be missing values (usually expressed as 'NA')
  - In the case of multivariate time series, each variable usually corresponds to a column (so each row corresponds to a date).

Example : US GNP data set

```
# Title:                  Gross National Product
# Source:                 U.S. Department of Commerce
# Frequency:              Quarterly
DATE,VALUE
1947-01-01,238.1
1947-04-01,241.5
1947-07-01,245.6
1947-10-01,255.6
1948-01-01,261.7
1948-04-01,268.7
1948-07-01,275.3
1948-10-01,276.6
1949-01-01,271.3
1949-04-01,267.5
1949-07-01,268.9
```

⋮

First step : remove an additive trend

- Consider a time series $X_1, \ldots, X_n$.

First step : remove an additive trend

- Consider a time series $X_1, \ldots, X_n$.
- It often includes an additive trend, for instance:

First step : remove an additive trend

- Consider a time series $X_1, \dots, X_n$.
- It often includes an additive trend, for instance:
  - a polynomial trend,

First step : remove an additive trend

- Consider a time series $X_1, \dots, X_n$.
- It often includes an additive trend, for instance:
  - a polynomial trend,
  - a seasonal trend.

First step : remove an additive trend

- Consider a time series $X_1, \ldots, X_n$.
- It often includes an additive trend, for instance:
  - a polynomial trend,
  - a seasonal trend.
- Then $X = D + Y$ where D belongs to a finite dimensional space V and Y is a centered random process.

First step : remove an additive trend

- Consider a time series $X_1, \ldots, X_n$.
- It often includes an additive trend, for instance:
  - a polynomial trend,
  - a seasonal trend.
- Then $X = D + Y$ where D belongs to a finite dimensional space V and Y is a centered random process.
- Trends can be estimated or removed by
  - Fitting the trend using least squares,

$$\widehat{D} = \operatorname*{argmin}_{d \in V} \sum_t |X_t - d_t|^2 \; .$$

First step : remove an additive trend

- Consider a time series $X_1, \ldots, X_n$.
- It often includes an additive trend, for instance:
  - a polynomial trend,
  - a seasonal trend.
- Then $X = D + Y$ where D belongs to a finite dimensional space V and Y is a centered random process.
- Trends can be estimated or removed by
  - Fitting the trend using least squares,

$$\widehat{D} = \operatorname*{argmin}_{d \in V} \sum_t |X_t - d_t|^2 .$$

  - Or applying a well chosen filter F_ψ, such that $\mathrm{F}_\psi(D) = 0$ and thus

$$\mathrm{F}_\psi(X) = \mathrm{F}_\psi(Y) .$$

R code example: Johnson and Johnson trend adjustment

trend-adjustment.html

Second step : choose a stochastic model on the random part

In time series analysis, one is interested in modeling the time dependence in the trend adjusted data $Y_1, \dots, Y_n$.

Second step : choose a stochastic model on the random part

In time series analysis, one is interested in modeling the time dependence in the trend adjusted data $Y_1, \ldots, Y_n$.

This can be done by using

- a parametric model.

Second step : choose a stochastic model on the random part

In time series analysis, one is interested in modeling the time dependence in the trend adjusted data $Y_1, \ldots, Y_n$.

This can be done by using

- a parametric model.

Example

$Y_1, \ldots, Y_n$ is the sample of a Gaussian ARMA(p, q) model with (unknown) parameter $\vartheta = (\theta_1, \ldots, \theta_q, \phi_1, \ldots, \phi_p, \sigma^2)$.

Second step : choose a stochastic model on the random part

In time series analysis, one is interested in modeling the time dependence in the trend adjusted data $Y_1, \ldots, Y_n$.

This can be done by using

- a parametric model.

Example

$Y_1, \ldots, Y_n$ is the sample of a Gaussian ARMA(p, q) model with (unknown) parameter $\vartheta = (\theta_1, \ldots, \theta_q, \phi_1, \ldots, \phi_p, \sigma^2)$.

- a non-parametric model.

Second step : choose a stochastic model on the random part

In time series analysis, one is interested in modeling the time dependence in the trend adjusted data $Y_1, \ldots, Y_n$.

This can be done by using

- ▷ a parametric model.

Example

$Y_1, \ldots, Y_n$ is the sample of a Gaussian ARMA(p, q) model with (unknown) parameter $\vartheta = (\theta_1, \ldots, \theta_q, \phi_1, \ldots, \phi_p, \sigma^2)$.

- ▷ a non-parametric model.

Example

$Y_1, \ldots, Y_n$ is the sample of a centered stationary Gaussian process with (unknown) autocovariance γ (or spectral density f).

Third step : estimate parameters, test hypotheses

Once a model is fixed for $Y_1, \ldots, Y_n$, it can be used to

- Estimate a parameter of the model such as ϑ, $\gamma(t)$, σ^2, f, ...

Third step : estimate parameters, test hypotheses

Once a model is fixed for $Y_1, \ldots, Y_n$, it can be used to

- ▷ Estimate a parameter of the model such as ϑ, $\gamma(t)$, σ^2, f, ...

  $\rightarrow$ Define an estimator, say $\widehat{\vartheta}_n$, which is a statistic based on the sample $Y_1, \ldots, Y_n$.

Third step : estimate parameters, test hypotheses

Once a model is fixed for $Y_1, \ldots, Y_n$, it can be used to

- Estimate a parameter of the model such as ϑ, $\gamma(t)$, σ^2, f, ...

  $\rightarrow$ Define an estimator, say $\widehat{\vartheta}_n$, which is a statistic based on the sample $Y_1, \ldots, Y_n$.
- Test hypotheses, for instance

$$H_0 = \{Y \text{ is white noise}\} \quad \text{against} \quad H_1 = \{Y \text{ is ARMA}(p, q)\}$$

Third step : estimate parameters, test hypotheses

Once a model is fixed for $Y_1, \ldots, Y_n$, it can be used to

- Estimate a parameter of the model such as ϑ, $\gamma(t)$, σ^2, f, $\ldots$

  $\rightarrow$ Define an estimator, say $\widehat{\vartheta}_n$, which is a statistic based on the sample $Y_1, \ldots, Y_n$.

- Test hypotheses, for instance

$$H_0 = \{Y \text{ is white noise}\} \quad \text{against} \quad H_1 = \{Y \text{ is ARMA}(p,q)\}$$

  $\rightarrow$ Define a statistical test, say

$$\delta = \begin{cases} 1 & \text{if } T_n > t_n \,, \\ 0 & \text{otherwise} \,, \end{cases}$$

  where T_n is a statistic based on the sample $Y_1, \ldots, Y_n$ and t_n is a threshold.

Stationary and ergodic models

- We see $Y_1, \ldots, Y_n$ as a finite sample of a stochastic process $(Y_t)_{t \in T}$, with $T = \mathbb{N}$ or $\mathbb{Z}$, distributed according to a statistical model.

Stationary and ergodic models

- We see $Y_1, \dots, Y_n$ as a finite sample of a stochastic process $(Y_t)_{t \in T}$, with $T = \mathbb{N}$ or $\mathbb{Z}$, distributed according to a statistical model.
- It is crucial to work with stationary and ergodic models.

Stationary and ergodic models

- We see $Y_1, \ldots, Y_n$ as a finite sample of a stochastic process $(Y_t)_{t \in T}$, with $T = \mathbb{N}$ or $\mathbb{Z}$, distributed according to a statistical model.
- It is crucial to work with stationary and ergodic models.
- Stationary means that the model is shift invariant: for all $n \geq 1$, and all $t_1, \ldots, t_n \in T$, we have

$$(X_{t_1}, \ldots, X_{t_n}) \stackrel{\mathrm{d}}{=} (X_{t_1+1}, \ldots, X_{t_n+1}) \ .$$

Stationary and ergodic models

- We see $Y_1, \ldots, Y_n$ as a finite sample of a stochastic process $(Y_t)_{t \in T}$, with $T = \mathbb{N}$ or $\mathbb{Z}$, distributed according to a statistical model.
- It is crucial to work with stationary and ergodic models.
- Stationary means that the model is shift invariant: for all $n \geq 1$, and all $t_1, \ldots, t_n \in T$, we have

$$(X_{t_1}, \ldots, X_{t_n}) \stackrel{\mathrm{d}}{=} (X_{t_1+1}, \ldots, X_{t_n+1}) .$$

- Ergodic means that observing one path $(Y_t)_{t \in T}$ allows one to recover the distribution entirely.

Examples

- An IID process is stationary and ergodic;

Examples

- An IID process is stationary and ergodic;
- A sequence of independent variables that are not identically distributed is not stationary;

Examples

- An IID process is stationary and ergodic;
- A sequence of independent variables that are not identically distributed is not stationary;
- some models are not stationary but can be made so through some basic transformations:

Examples

- An IID process is stationary and ergodic;
- A sequence of independent variables that are not identically distributed is not stationary;
- some models are not stationary but can be made so through some basic transformations:
  - removing trends,

Examples

- An IID process is stationary and ergodic;
- A sequence of independent variables that are not identically distributed is not stationary;
- some models are not stationary but can be made so through some basic transformations:
  - removing trends,
  - linear filters such as taking the increments (random walks),

Examples

- An IID process is stationary and ergodic;
- A sequence of independent variables that are not identically distributed is not stationary;
- some models are not stationary but can be made so through some basic transformations:
  - removing trends,
  - linear filters such as taking the increments (random walks),
  - non-linear filters such as taking the logs.

Examples

- An IID process is stationary and ergodic;
- A sequence of independent variables that are not identically distributed is not stationary;
- some models are not stationary but can be made so through some basic transformations:
  - removing trends,
  - linear filters such as taking the increments (random walks),
  - non-linear filters such as taking the logs.
- A sequence of variables $(Y_t)_{t\in\mathbb{Z}}$ that is constant, *i.e.* $Y_t = Y_0$ for all t, is stationary but is not ergodic;

Examples

- An IID process is stationary and ergodic;
- A sequence of independent variables that are not identically distributed is not stationary;
- some models are not stationary but can be made so through some basic transformations:
  - removing trends,
  - linear filters such as taking the increments (random walks),
  - non-linear filters such as taking the logs.
- A sequence of variables $(Y_t)_{t\in\mathbb{Z}}$ that is constant, *i.e.* $Y_t = Y_0$ for all t, is stationary but is not ergodic;
- A Markov chain on a finite state space can be made stationary by choosing the initial state adequately. If it is irreducible, then it is ergodic.

R code example: dependent data

non-iid-data.html

L^2 space

We denote

$$L^2(\Omega, \mathcal{F}, \mathbb{P}) = \left\{X \ \mathbb{C}\text{-valued r.v. such that } \mathbb{E}\left[|X|^2\right] < \infty\right\} .$$

$(L^2, \langle, \rangle)$ is a Hilbert space with

$$\langle X, Y \rangle = \mathbb{E}\left[X\overline{Y}\right] .$$

Definition : L^2 Processes

The process $X = (X_t)_{t \in T}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $\mathbb{C}$ is an L^2 process if $X_t \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ for all $t \in T$.

Mean and covariance functions

Let $X = (X_t)_{t \in T}$ be an L^2 process.

Mean and covariance functions

Let $X = (X_t)_{t \in T}$ be an L^2 process.

- ▷ Its mean function is defined by $\mu(t) = \mathbb{E}[X_t]$,

Mean and covariance functions

Let $X = (X_t)_{t \in T}$ be an L^2 process.

- Its mean function is defined by $\mu(t) = \mathbb{E}[X_t]$,
- Its covariance function is defined by

$$\gamma(s,t) = \mathrm{cov}(X_s, X_t) = \mathbb{E}\left[X_s \overline{X_t}\right] - \mathbb{E}[\mathbf{X}_s]\,\mathbb{E}\left[\overline{X_t}\right] .$$

Mean and covariance functions

Let $X = (X_t)_{t \in T}$ be an L^2 process.

- Its mean function is defined by $\mu(t) = \mathbb{E}[X_t]$,
- Its covariance function is defined by

$$\gamma(s,t) = \mathrm{cov}(X_s, X_t) = \mathbb{E}\left[X_s \overline{X_t}\right] - \mathbb{E}[\mathbf{X}_s]\,\mathbb{E}\left[\overline{X_t}\right] .$$

Hermitian symmetry, non-negative definiteness

For all finite subset $I \subset T$, $\Gamma_I = \mathrm{Cov}([X(t)]_{t \in I}) = [\gamma(s,t)]_{s,t \in I}$ is a hermitian non-negative definite matrix.

Examples

- L^2 independent random variables $(X_t)_{t\in\mathbb{Z}}$ have mean $\mu(t) = \mathbb{E}[X_t]$ and covariance

$$\gamma(s,t) = \begin{cases} \mathrm{var}(X_t) & \text{if } s = t, \\ 0 & \text{otherwise.} \end{cases}$$

Examples

- L^2 independent random variables $(X_t)_{t\in\mathbb{Z}}$ have mean $\mu(t) = \mathbb{E}[X_t]$ and covariance

$$\gamma(s,t) = \begin{cases} \mathrm{var}(X_t) & \text{if } s = t, \\ 0 & \text{otherwise.} \end{cases}$$

- A Gaussian process is an L^2 process whose law is entirely determined by its mean and covariance functions: for all $I = \{t_1, \dots, t_n\}$,

$$(X_s)_{s\in I} \sim \mathcal{N}\left((\mu_s)_{s\in I}, \Gamma_I\right) \ .$$

Weakly stationary processes

- Let $T = \mathbb{Z}$. Let X be an L^2 strictly stationary process with mean function μ and covariance function γ.

Weakly stationary processes

- Let $T = \mathbb{Z}$. Let X be an L^2 strictly stationary process with mean function μ and covariance function γ.
- Then $\mu(t) = \mu(0)$ and $\gamma(s,t) = \gamma(s-t,0)$ for all $s,t \in T$.

Weakly stationary processes

- Let $T = \mathbb{Z}$. Let X be an L^2 strictly stationary process with mean function μ and covariance function γ.
- Then $\mu(t) = \mu(0)$ and $\gamma(s,t) = \gamma(s-t,0)$ for all $s, t \in T$.

Definition : Weak stationarity

We say that a random process X is weakly stationary with mean $\mu \in \mathbb{C}$ and autocovariance function $\gamma : \mathbb{Z} \to \mathbb{C}$ if it is L^2 with mean function $t \mapsto \mu$ and covariance function $(s,t) \mapsto \gamma(s-t)$.

Weakly stationary processes

- Let $T = \mathbb{Z}$. Let X be an L^2 strictly stationary process with mean function μ and covariance function γ.
- Then $\mu(t) = \mu(0)$ and $\gamma(s,t) = \gamma(s-t,0)$ for all $s,t \in T$.

Definition : Weak stationarity

We say that a random process X is weakly stationary with mean $\mu \in \mathbb{C}$ and autocovariance function $\gamma : \mathbb{Z} \to \mathbb{C}$ if it is L^2 with mean function $t \mapsto \mu$ and covariance function $(s,t) \mapsto \gamma(s-t)$.

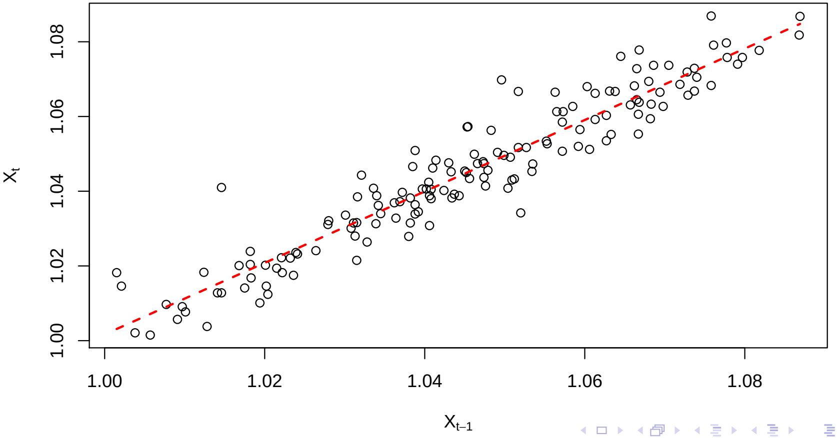
- The autocorrelation function is then defined (when $\gamma(0) > 0$) by

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} \in [-1, 1] \; .$$

Autocorrelation=slope of regression line

We have, for all $t \in \mathbb{Z}$ and $h = 1, 2, \ldots,$

$$X_t = \text{Constant} + \rho(h) X_{t-h} + \epsilon_{t,h} \quad \text{with} \quad \epsilon_{t,h} \perp \mathrm{Span}\left(1, X_{t-h}\right) \ .$$



Partial Autocorrelation

- We can also write, for all $t \in \mathbb{Z}$ and $h = 1, 2, \ldots,$

$$X_t = \text{Constant} + \sum_{k=1}^{h-1} \phi_k X_{t-k} + \kappa(h) X_{t-h} + \epsilon_{t,h}$$

where

$$\epsilon_{t,h} \perp \mathrm{Span}\left(1, X_{t-1}, \ldots, X_{t-h}\right) \ .$$

Partial Autocorrelation

- We can also write, for all $t \in \mathbb{Z}$ and $h = 1, 2, \ldots,$

$$X_t = \text{Constant} + \sum_{k=1}^{h-1} \phi_k X_{t-k} + \kappa(h) X_{t-h} + \epsilon_{t,h}$$

  where

$$\epsilon_{t,h} \perp \mathrm{Span}\left(1, X_{t-1}, \ldots, X_{t-h}\right) .$$

- In the following, we plot successively
  - X_t as a function of X_{t-h}, compared to the regression line;

Partial Autocorrelation

- We can also write, for all $t \in \mathbb{Z}$ and $h = 1, 2, \ldots,$

$$X_t = \text{Constant} + \sum_{k=1}^{h-1} \phi_k X_{t-k} + \kappa(h) X_{t-h} + \epsilon_{t,h}$$

  where

$$\epsilon_{t,h} \perp \mathrm{Span}\left(1, X_{t-1}, \ldots, X_{t-h}\right) \ .$$

- In the following, we plot successively
  - X_t as a function of X_{t-h}, compared to the regression line;
  - X_t as a function of X_{t-h}, compared to the predictor

    $\text{Constant} + \sum_{k=1}^{h-1} \phi_k X_{t-k} + \kappa(h) X_{t-h}$.

Partial Autocorrelation

- We can also write, for all $t \in \mathbb{Z}$ and $h = 1, 2, \ldots,$

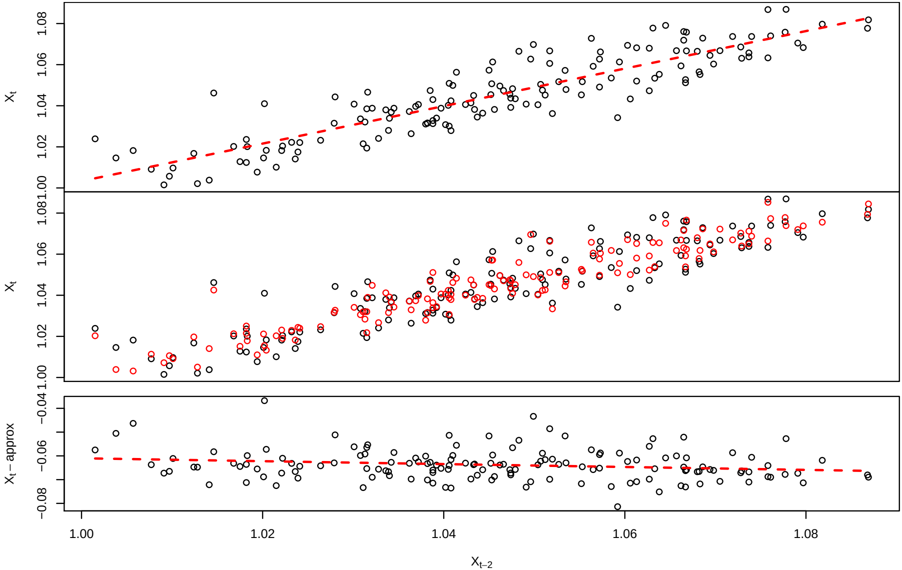
$$X_t = \text{Constant} + \sum_{k=1}^{h-1} \phi_k X_{t-k} + \kappa(h) X_{t-h} + \epsilon_{t,h}$$

  where

$$\epsilon_{t,h} \perp \mathrm{Span}\left(1, X_{t-1}, \ldots, X_{t-h}\right) \ .$$

- In the following, we plot successively
  - X_t as a function of X_{t-h}, compared to the regression line;
  - X_t as a function of X_{t-h}, compared to the predictor
    $$\text{Constant} + \sum_{k=1}^{h-1} \phi_k X_{t-k} + \kappa(h) X_{t-h}.$$
  - $X_t - \left(\text{Constant} + \sum_{k=1}^{h-1} \phi_k X_{t-k}\right)$ as a function of X_{t-h}, compared to the regression line $X_{t-h} \mapsto \kappa(h) X_{t-h}$.

Partial Autocorrelation=slope of partial regression



Examples

- An L^2 strictly stationary process is weakly stationary.

Examples

- An L^2 strictly stationary process is weakly stationary.
- The constant L^2 process has constant autocovariance function.

Examples

- An L^2 strictly stationary process is weakly stationary.
- The constant L^2 process has constant autocovariance function.

Strong and weak white noise

- A sequence of L^2 i.i.d. random variables is called a strong white noise, denoted by $X \sim \mathrm{IID}(\mu, \sigma^2)$.

Examples

- An L^2 strictly stationary process is weakly stationary.
- The constant L^2 process has constant autocovariance function.

Strong and weak white noise

- A sequence of L^2 i.i.d. random variables is called a strong white noise, denoted by $X \sim \mathrm{IID}(\mu, \sigma^2)$.
- An L^2 process X with constant mean μ and constant diagonal covariance function equal to σ^2 is called a weak white noise. It is denoted by $X \sim \mathrm{WN}(\mu, \sigma^2)$. (It does not have to be i.i.d.)

Examples based on stationarity preserving linear filters

- Let X be weakly stationary with mean μ and autocovariance γ.

Examples based on stationarity preserving linear filters

- Let X be weakly stationary with mean μ and autocovariance γ.
- Let $\psi \in \ell^1$. We define $Y = \mathrm{F}_{\psi}(X)$ by

$$Y_t = \sum_k \psi_k X_{t-k} \;, \qquad t \in \mathbb{Z} \;.$$

Examples based on stationarity preserving linear filters

- Let X be weakly stationary with mean μ and autocovariance γ.
- Let $\psi \in \ell^1$. We define $Y = \mathrm{F}_\psi(X)$ by

$$Y_t = \sum_k \psi_k X_{t-k} \;, \qquad t \in \mathbb{Z} \;.$$

- Then Y is weakly stationary with mean μ' and autocovariance γ' given by

$$\mu' = \mu \sum_k \psi_k$$

$$\gamma'(\tau) = \sum_{\ell,k} \psi_k \overline{\psi_\ell} \gamma(\tau + \ell - k) \tag{4}$$

Spectral measure

- Given a function $\gamma : \mathbb{Z} \to \mathbb{C}$, does there exist a weakly stationary process $(X_t)_{t \in \mathbb{Z}}$ with autocovariance γ?

Spectral measure

- Given a function $\gamma : \mathbb{Z} \to \mathbb{C}$, does there exist a weakly stationary process $(X_t)_{t\in\mathbb{Z}}$ with autocovariance γ?
- A necessary and sufficient condition is that $\Gamma_I = [\gamma(s-t)]_{s,t\in I}$ is a hermitian non-negative definite matrix for all finite subset $I \subset \mathbb{Z}$.

Spectral measure

- Given a function $\gamma : \mathbb{Z} \to \mathbb{C}$, does there exist a weakly stationary process $(X_t)_{t\in\mathbb{Z}}$ with autocovariance γ?
- A necessary and sufficient condition is that $\Gamma_I = [\gamma(s-t)]_{s,t\in I}$ is a hermitian non-negative definite matrix for all finite subset $I \subset \mathbb{Z}$.
- We say that γ is hermitian symmetric and non-negative definite.

Spectral measure

- Given a function $\gamma : \mathbb{Z} \to \mathbb{C}$, does there exist a weakly stationary process $(X_t)_{t \in \mathbb{Z}}$ with autocovariance γ?
- A necessary and sufficient condition is that $\Gamma_I = [\gamma(s-t)]_{s,t \in I}$ is a hermitian non-negative definite matrix for all finite subset $I \subset \mathbb{Z}$.
- We say that γ is hermitian symmetric and non-negative definite.

Herglotz Theorem

Let $\gamma : \mathbb{Z} \to \mathbb{C}$. Then the two following assertions are equivalent:

(i) γ is hermitian symmetric and non-negative definite.

(ii) There exists a finite non-negative measure ν on $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$ such that,

$$\text{for all } t \in \mathbb{Z}, \quad \gamma(t) = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}\lambda t} \, \nu(\mathrm{d}\lambda) \ . \tag{5}$$

When these two assertions hold, ν is uniquely defined by (5).

Spectral density

If moreover $\gamma \in \ell^1(\mathbb{Z})$, these assertions are equivalent to

$$f(\lambda) := \frac{1}{2\pi} \sum_{t \in \mathbb{Z}} \mathrm{e}^{-\mathrm{i}\lambda t} \gamma(t) \geq 0 \text{ for all } \lambda \in \mathbb{R} ,$$

and ν has density f (that is, $\nu(\mathrm{d}\lambda) = f(\lambda)\mathrm{d}\lambda$).

Spectral density

If moreover $\gamma \in \ell^1(\mathbb{Z})$, these assertions are equivalent to

$$f(\lambda) := \frac{1}{2\pi} \sum_{t \in \mathbb{Z}} \mathrm{e}^{-\mathrm{i}\lambda t} \gamma(t) \geq 0 \text{ for all } \lambda \in \mathbb{R} ,$$

and ν has density f (that is, $\nu(\mathrm{d}\lambda) = f(\lambda)\mathrm{d}\lambda$).

Definition : spectral measure and spectral density

If γ is the autocovariance of a weakly stationary process X, the corresponding measure ν is called the spectral measure of X. Whenever the spectral measure ν admits a density f, it is called the spectral density function.

Examples

- Let $X \sim \mathrm{WN}(\mu, \sigma^2)$. Then $f(\lambda) = \frac{\sigma^2}{2\pi}$.

Examples

- Let $X \sim \mathrm{WN}(\mu, \sigma^2)$. Then $f(\lambda) = \frac{\sigma^2}{2\pi}$.
- Let X be a weakly stationary process with covariance function γ/spectral measure ν. Define

$$Y_t = \sum_k \psi_k X_{t-k}$$

for a sequence $\psi \in \ell^1$.

Examples

- Let $X \sim \mathrm{WN}(\mu, \sigma^2)$. Then $f(\lambda) = \frac{\sigma^2}{2\pi}$.
- Let X be a weakly stationary process with covariance function γ/spectral measure ν. Define

$$Y_t = \sum_k \psi_k X_{t-k}$$

  for a sequence $\psi \in \ell^1$.
- Recall that Y is a weakly stationary process with covariance function

$$\gamma'(\tau) = \sum_{\ell,k} \psi_k \overline{\psi_\ell} \gamma(\tau + \ell - k) \; .$$

Examples

- Let $X \sim \mathrm{WN}(\mu, \sigma^2)$. Then $f(\lambda) = \frac{\sigma^2}{2\pi}$.
- Let X be a weakly stationary process with covariance function γ/spectral measure ν. Define

$$Y_t = \sum_k \psi_k X_{t-k}$$

for a sequence $\psi \in \ell^1$.
- Recall that Y is a weakly stationary process with covariance function

$$\gamma'(\tau) = \sum_{\ell,k} \psi_k \overline{\psi_\ell} \gamma(\tau + \ell - k) \; .$$

- Then Y is a weakly stationary process with spectral measure ν' having density $\lambda \mapsto \left|\sum_k \psi_k \mathrm{e}^{-\mathrm{i}\lambda k}\right|^2$ with respect to ν,

$$\nu'(\mathrm{d}\lambda) = \left|\sum_k \psi_k \mathrm{e}^{-\mathrm{i}\lambda k}\right|^2 \nu(\mathrm{d}\lambda) \; .$$

A special one : the harmonic process

Let $(A_k)_{1\leq k\leq N}$ be N real valued L^2 random variables. Denote $\sigma_k^2 = \mathbb{E}\left[A_k^2\right]$. Let $(\Phi_k)_{1\leq k\leq N}$ be N i.i.d. random variables with a uniform distribution on $[0, 2\pi]$, and independent of $(A_k)_{1\leq k\leq N}$. Define

$$X_t = \sum_{k=1}^{N} A_k \cos(\lambda_k t + \Phi_k) \;, \tag{6}$$

where $(\lambda_k)_{1\leq k\leq N} \in [-\pi, \pi]$ are N frequencies. The process (X_t) is called a harmonic process. It satisfies $\mathbb{E}[X_t] = 0$ and, for all $s, t \in \mathbb{Z}$,

$$\mathbb{E}[X_s X_t] = \frac{1}{2}\sum_{k=1}^{N} \sigma_k^2 \cos(\lambda_k(s-t)) \;.$$

Hence X is weakly stationary with autocovariance

$$\gamma(t) = \frac{1}{2}\sum_{k=1}^{N}\sigma_k^2\cos(\lambda_k t) = \int_{\mathbb{T}} \mathrm{e}^{\mathrm{i}\lambda t}\left(\frac{1}{4}\sum_{k=1}^{N}\sigma_k^2(\delta_{-\lambda_k}(\mathrm{d}\lambda) + \delta_{\lambda_k}(\mathrm{d}\lambda))\right) \;.$$

Empirical estimates

- Suppose you want to estimate the mean and the autocovariance from a sample $X_1, \dots, X_n$.

Empirical estimates

- Suppose you want to estimate the mean and the autocovariance from a sample $X_1, \ldots, X_n$.
- Define the empirical mean as

$$\widehat{\mu}_n = \frac{1}{n} \sum_{k=1}^{n} X_k \; ,$$

Empirical estimates

- Suppose you want to estimate the mean and the autocovariance from a sample $X_1, \ldots, X_n$.
- Define the empirical mean as

$$\widehat{\mu}_n = \frac{1}{n} \sum_{k=1}^{n} X_k \;,$$

- Define the empirical autocovariance and autocorrelation functions as

$$\widehat{\gamma}_n(h) = \frac{1}{n} \sum_{k=1}^{n-|h|} (X_k - \widehat{\mu}_n)(X_{k+|h|} - \widehat{\mu}_n) \quad \text{and}$$

$$\widehat{\rho}_n(h) = \frac{\widehat{\gamma}_n(h)}{\widehat{\gamma}_n(0)} \;.$$

Periodogram

- The previous formula only work for $h = -n + 1, \dots, n - 1$.

Periodogram

- The previous formula only work for $h = -n+1, \ldots, n-1$.
- Define $\widehat{\gamma}_n(h) = \widehat{\rho}_n(h) = 0$ for all $|h| \geq n$.

Periodogram

- The previous formula only work for $h = -n+1, \ldots, n-1$.
- Define $\widehat{\gamma}_n(h) = \widehat{\rho}_n(h) = 0$ for all $|h| \geq n$.
- Now $\widehat{\gamma}_n$ is defined on $\mathbb{Z}$ and satisfies

$$\widehat{\gamma}_n(h) = \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}\lambda h} \, I_n(\lambda) \, \mathrm{d}\lambda \;,$$

where I_n is called the (raw) periodogram and is defined by

$$I_n(\lambda) = \frac{1}{2\pi n} \left| \sum_{k=1}^{n} (X_k - \widehat{\mu}_n) \, \mathrm{e}^{-\mathrm{i}\lambda k} \right|^2 \;.$$

Periodogram

- The previous formula only work for $h = -n+1, \ldots, n-1$.
- Define $\widehat{\gamma}_n(h) = \widehat{\rho}_n(h) = 0$ for all $|h| \geq n$.
- Now $\widehat{\gamma}_n$ is defined on $\mathbb{Z}$ and satisfies

$$\widehat{\gamma}_n(h) = \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}\lambda h}\, I_n(\lambda)\, \mathrm{d}\lambda\ ,$$

  where I_n is called the (raw) periodogram and is defined by

$$I_n(\lambda) = \frac{1}{2\pi n} \left| \sum_{k=1}^{n} (X_k - \widehat{\mu}_n)\, \mathrm{e}^{-\mathrm{i}\lambda k} \right|^2 \ .$$

- $I_n(\lambda)$ can be seen as a (bad) estimator of the spectral density $f(\lambda)$.