Time series : an introduction using the likelihood as a guideline

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

Telecom Paris

BGDIA703 2023/2024

Outline

Example : USD vs EUR currency exchange rate

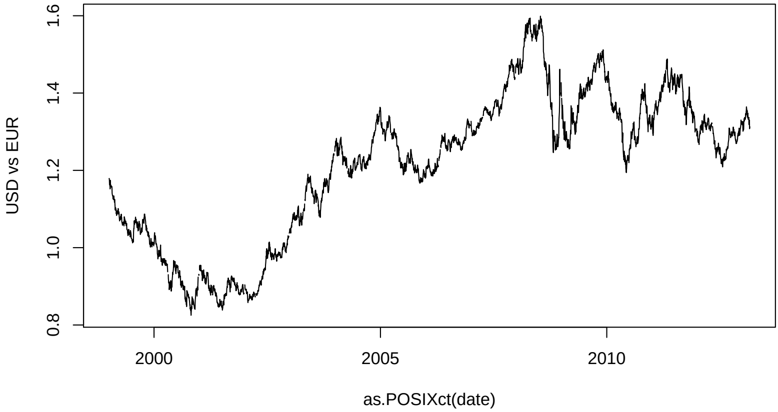
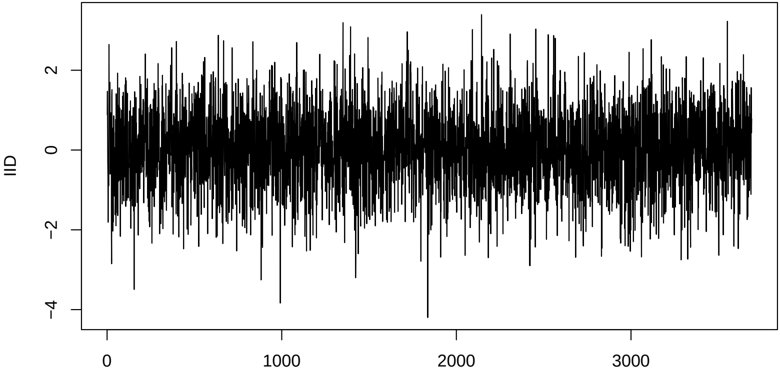


Figure: Daily currency exchange rate : price of 1 Euro in US Dollars.

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”.

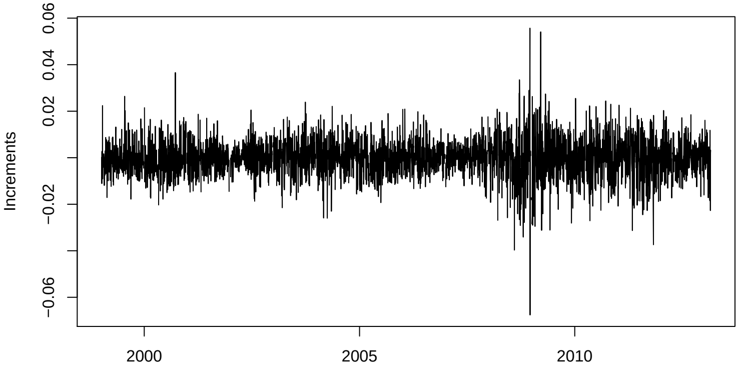


Figure: Increments of daily USD-EUR currency exchange rate.

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".

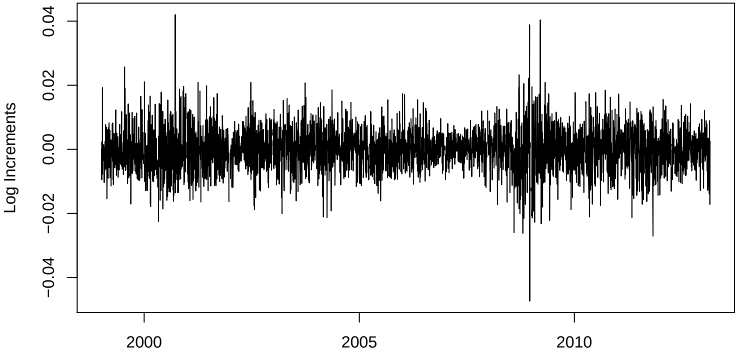


Figure: Log returns of daily USD-EUR currency exchange rate.

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

Looking at things “locally” ...

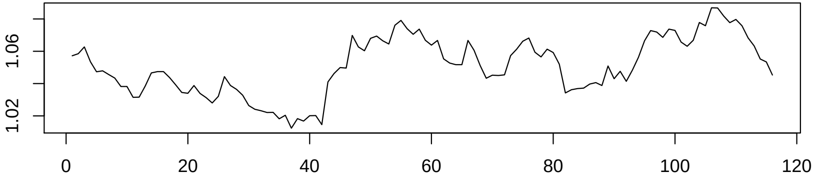


Figure: Daily currency exchange rate : price of 1 Euro in US Dollars, on a shorter observation window: between 1999-05-21 and 1999-12-17.

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

1. Examples of financial time series
2. Reminders: i.i.d. models
   - Univariate models
   - Multivariate models
   - Regression model
   - Hidden variables
3. Introducing dynamics

Discrete observations

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

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

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

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

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

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

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):

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

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):

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

- 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):

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

- 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

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

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

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

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.

$$\mathbf{X}_t = X_t(i) \quad i = 1, \ldots, p \; ,$$

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.
$$\mathbf{X}_t = X_t(i) \quad i = 1, \dots, p \,,$$
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.

$$\mathbf{X}_t = X_t(i) \quad i = 1, \ldots, p \; ,$$

  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.
$$\mathbf{X}_t = X_t(i) \quad i = 1, \ldots, p \; ,$$
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 $\mathbf{X}_t = X_t(i) \quad i = 1, \ldots, p$.
- 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 $\theta = (\boldsymbol{\mu}, \Sigma)$,

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

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

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

- Maximizing the likelihood yields:
  - the empirical mean

$$\widehat{\mu}_n(i) = \frac{1}{n} \sum_{t=1}^{n} X_t(i) \; .$$

  - the empirical covariance matrix

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

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

- Maximizing the likelihood yields:
  - the empirical mean

$$\widehat{\mu}_n(i) = \frac{1}{n}\sum_{t=1}^{n} X_t(i) \ .$$

  - the empirical covariance matrix

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

- In high dimension (p and n are of similar order), it is required to introduce a penalty based on a sparse or low rank assumption.

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

- Maximizing the likelihood yields:
  - the empirical mean

$$\widehat{\mu}_n(i) = \frac{1}{n}\sum_{t=1}^{n} X_t(i) \ .$$

  - the empirical covariance matrix

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

- In high dimension (p and n are of similar order), it is required to introduce a penalty based on a sparse or low rank assumption.
- From a regression perspective, the sparsity of the precision matrix $M = \Sigma^{-1}$ is more meaningful: for all $i \neq j$, $\mathbb{P}_\theta^{X(i)|\mathbf{X}(-i)}$ depends on $X(j)$ if and only if $\Sigma^{-1}[i,j] \neq 0$.

From bivariate distribution to conditional distribution

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

From bivariate distribution to conditional distribution

- In a regression model, each multivariate observation $\mathbf{X}_t$ is split into a pair of variables : $\mathbf{X}_t = (\mathbf{Z}_t, Y_t)$, where, usually, $\mathbf{Z}_t$ itself is multivariate, say valued in $\mathbb{R}^p$, and Y_t 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 $\mathbf{X}_t$ is split into a pair of variables : $\mathbf{X}_t = (\mathbf{Z}_t, Y_t)$, where, usually, $\mathbf{Z}_t$ itself is multivariate, say valued in $\mathbb{R}^p$, and Y_t 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 $\mathbf{X}_t$ is split into a pair of variables : $\mathbf{X}_t = (\mathbf{Z}_t, Y_t)$, where, usually, $\mathbf{Z}_t$ itself is multivariate, say valued in $\mathbb{R}^p$, and Y_t 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 $\mathbf{Z}_t$ as an input (regression variable) and Y_t 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:

$$\log L_n(\theta) = \sum_{t=1}^{n} \log p_\theta(Y_t | \mathbf{Z}_t)$$

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:

$$\log L_n(\theta) = \sum_{t=1}^{n} \log p_\theta(Y_t|\mathbf{Z}_t)$$

- 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 $\mathbf{X}_t$ 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_t.

The mixture model

- In certain cases, each (univariate or multivariate) observation $\mathbf{X}_t$ 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_t.
- 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 $\mathbf{X}_t$ 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_t.
- 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 $\mathbf{X}_t$ 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_t.
- 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 $\mathbf{X}_t$ 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_t.
- 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):

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

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):

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

- 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)

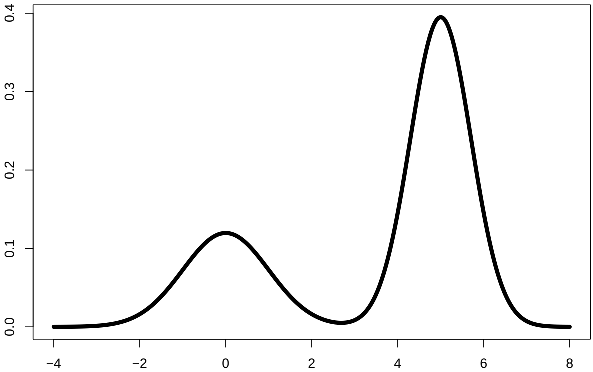


Figure: Density of the mixture of two Gaussian distributions

Two examples (cont)

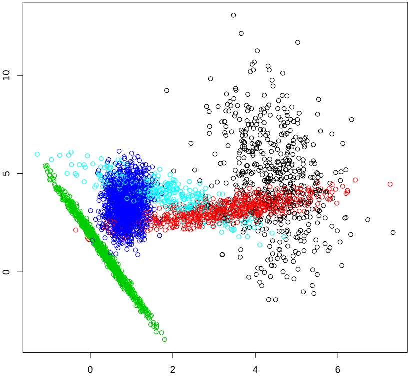


Figure: IID draws of the mixture of 5 bidimensional Gaussian distributions. Colors represent the (supposedly hidden) cluster variables.

1. Examples of financial time series

2. Reminders: i.i.d. models

3. Introducing dynamics
   - What's wrong with i.i.d. models ?
   - Univariate models
   - Multivariate models
   - Partially observed multivariate time series

Back to the USD vs EUR currency exchange rate.

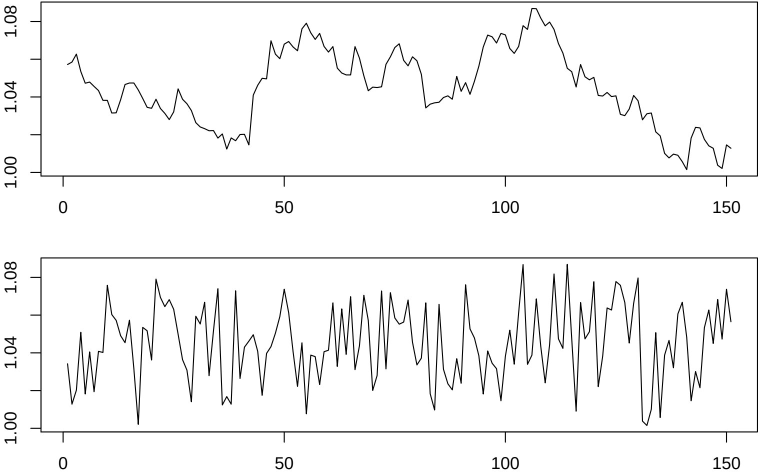


Figure: Top : price of 1 Euro in US Dollars between 1999-05-21 and 1999-12-17; Bottom : the same in randomly shuffled order.

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

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

  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

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

- 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 $\mathbf{X}$:

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

  - 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

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

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

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

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

Two important particular cases

- The i.i.d. case :

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

$$p_{\theta}^{X_t|X_{1:(t-1)}}(x_t|x_{1:t-1}) = p_{\theta}^{X_t}(x_t) \; .$$

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

$$p_{\theta}^{X_t|X_{1:(t-1)}}(x_t|x_{1:t-1}) = p_{\theta}^{X_t}(x_t) = p_{\theta}(x_t) \; ,$$

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

Two important particular cases (cont.)

- The homogeneous Markov case :

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

$$p_{\theta}^{X_t|X_{1:(t-1)}}(x_t|x_{1:t-1}) = p_{\theta}^{X_t|X_{t-1}}(x_t|x_{t-1}) \ .$$

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

$$p_{\theta}^{X_t|X_{1:(t-1)}}(x_t|x_{1:t-1}) = p_{\theta}^{X_t|X_{t-1}}(x_t|x_{t-1}) = q_{\theta}(x_t|x_{t-1}) \ .$$

Graphical representation of a homogeneous Markov chain

$$\cdots \dashrightarrow X_{t-1} \xrightarrow{q_\theta} X_t \xrightarrow{q_\theta} X_{t+1} \dashrightarrow \cdots$$

- 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

$$X_t = \phi X_{t-1} + \epsilon_t \ ,$$

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

$$X_t = \phi X_{t-1} + \epsilon_t \ ,$$

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

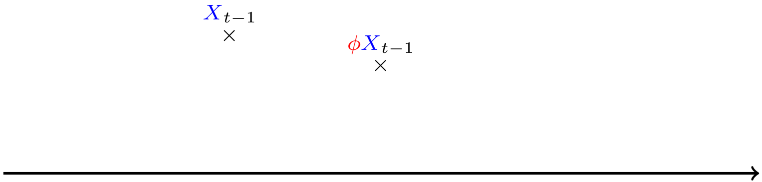
Examples of conditional density

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

$$X_t = \phi X_{t-1} + \epsilon_t \ ,$$

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



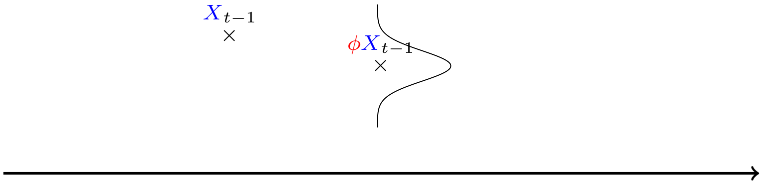
Examples of conditional density

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

$$X_t = \phi X_{t-1} + \epsilon_t \ ,$$

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



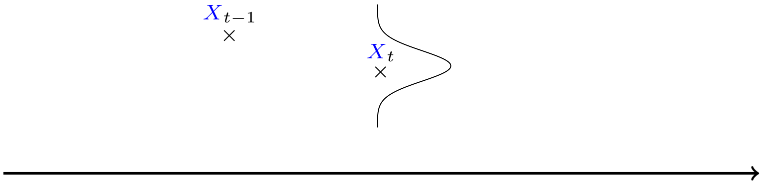
Examples of conditional density

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

$$X_t = \phi X_{t-1} + \epsilon_t \ ,$$

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



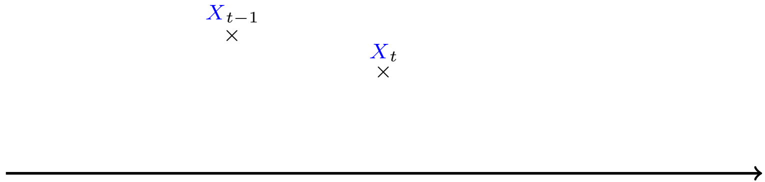
Examples of conditional density

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

$$X_t = \phi X_{t-1} + \epsilon_t \ ,$$

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



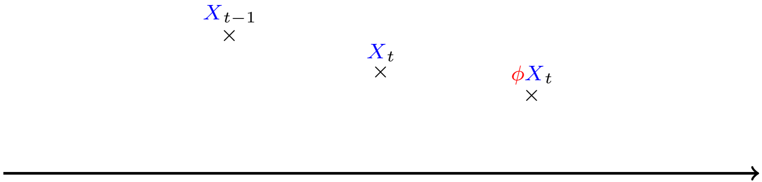
Examples of conditional density

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

$$X_t = \phi X_{t-1} + \epsilon_t \ ,$$

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



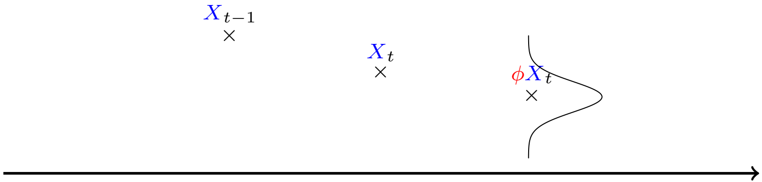
Examples of conditional density

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

$$X_t = \phi X_{t-1} + \epsilon_t \ ,$$

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



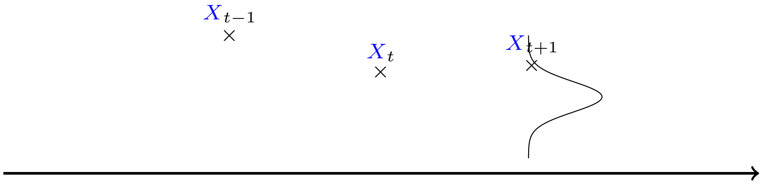
Examples of conditional density

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

$$X_t = \phi X_{t-1} + \epsilon_t \ ,$$

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



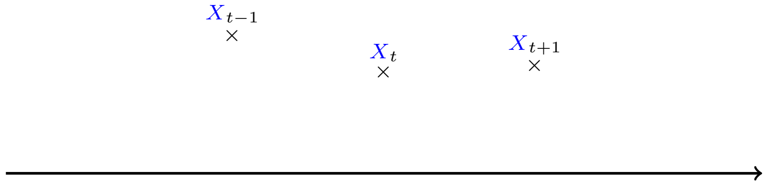
Examples of conditional density

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

$$X_t = \phi X_{t-1} + \epsilon_t \ ,$$

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



 BGDIA703 2023/2024

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

$$X_t = \sqrt{a + bX_{t-1}^2}\, \epsilon_t ,$$

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

$$X_t = \sqrt{a + bX_{t-1}^2}\, \epsilon_t \ ,$$

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

Examples of conditional density (cont.)

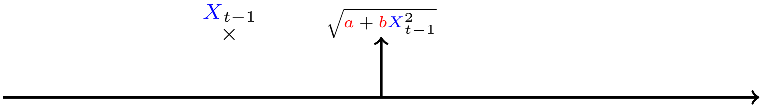
An heteroscedastic model : ARCH(1).

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

$$X_t = \sqrt{a + bX_{t-1}^2}\, \epsilon_t \ ,$$

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



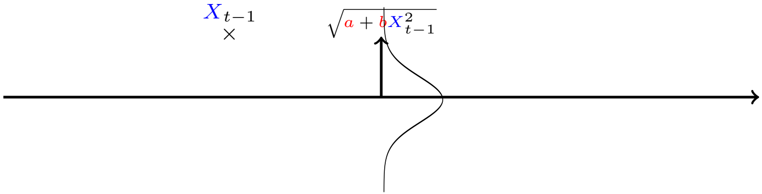
Examples of conditional density (cont.)

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

$$X_t = \sqrt{a + bX_{t-1}^2}\, \epsilon_t \ ,$$

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



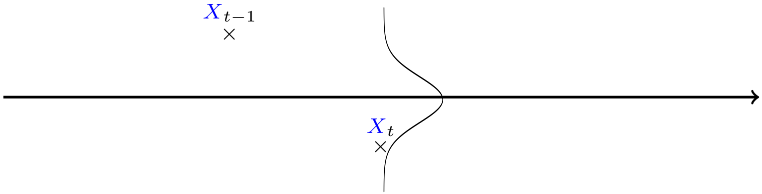
Examples of conditional density (cont.)

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

$$X_t = \sqrt{a + bX_{t-1}^2}\, \epsilon_t \ ,$$

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



Examples of conditional density (cont.)

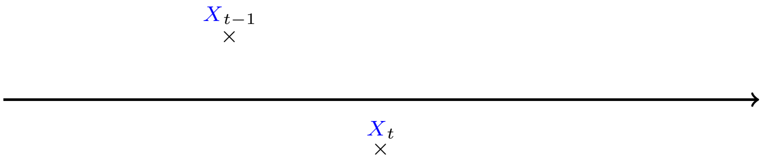
An heteroscedastic model : ARCH(1).

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

$$X_t = \sqrt{a + bX_{t-1}^2}\, \epsilon_t \ ,$$

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



Examples of conditional density (cont.)

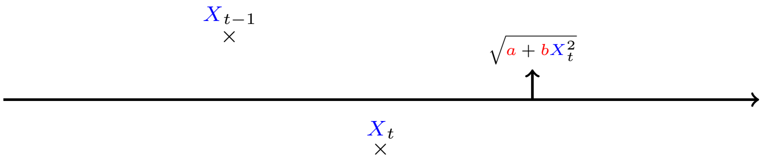
An heteroscedastic model : ARCH(1).

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

$$X_t = \sqrt{a + bX_{t-1}^2}\, \epsilon_t \,,$$

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



Examples of conditional density (cont.)

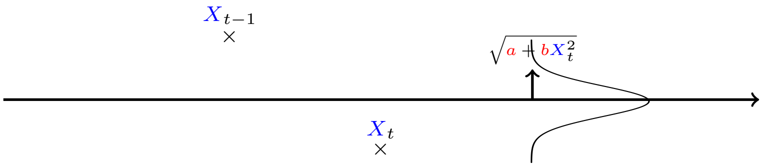
An heteroscedastic model : ARCH(1).

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

$$X_t = \sqrt{a + bX_{t-1}^2}\, \epsilon_t \ ,$$

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



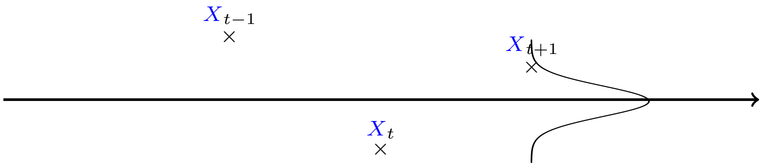
Examples of conditional density (cont.)

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

$$X_t = \sqrt{a + bX_{t-1}^2}\,\epsilon_t \ ,$$

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



Examples of conditional density (cont.)

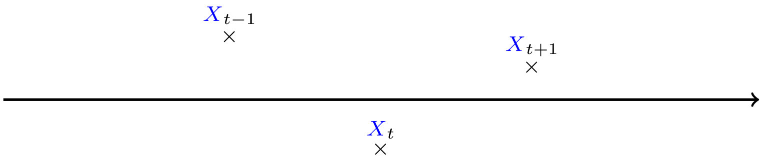
An heteroscedastic model : ARCH(1).

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

$$X_t = \sqrt{a + bX_{t-1}^2}\,\epsilon_t \;,$$

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

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

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

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

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

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

- The likelihood is no longer invariant by permutation.

Exemple: likelihood of the Gaussian AR(1) model

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

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

It follows that the (conditional) negated log likelihood reads

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

which leads to the estimators

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

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

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

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

It follows that the (conditional) negated log likelihood reads

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

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

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

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

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

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

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

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

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 $\mathbf{X}_t = \begin{bmatrix} X_t & X_{t-1} & \dots & X_{t-p+1} \end{bmatrix}^T$, 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

$$
\begin{array}{cccc}
\cdots & X_t(1) & X_{t+1}(1) & \cdots \\
 & \updownarrow & \updownarrow & \\
\cdots & X_t(2) & X_{t+1}(2) & \cdots
\end{array}
$$

BGDIA703 2023/2024

Exemple of Multivariate time series: general bivariate case

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

- ▷ IID case

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

- ▷ Markov case:

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

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.

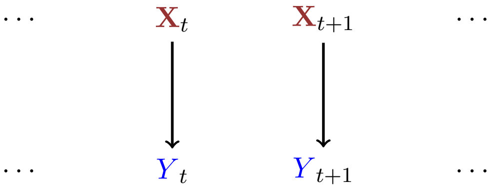
Examples of partially observed multivariate time series

- IID case

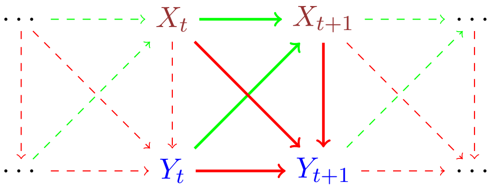
$$\begin{tikzcd}
\cdots & \mathbf{X}_t \arrow[d] & \mathbf{X}_{t+1} \arrow[d] & \cdots \\
\cdots & Y_t & Y_{t+1} & \cdots
\end{tikzcd}$$

Examples of partially observed multivariate time series

- IID case

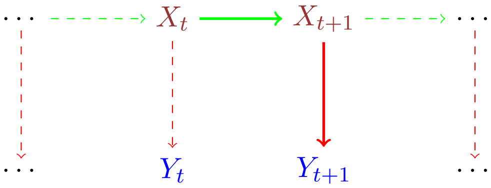


- Partially observed Markov model: general case.



Examples of partially observed multivariate time series (cont.)

- Hidden Markov model.



In this special case:

  - (X_t) alone is a Markov chain.

Examples of partially observed multivariate time series (cont.)

- Hidden Markov model.

$$\begin{tikzcd}
\cdots \arrow[r, dashed] \arrow[d, dashed] & X_t \arrow[r] \arrow[d, dashed] & X_{t+1} \arrow[r, dashed] \arrow[d] & \cdots \arrow[d, dashed] \\
\cdots & Y_t & Y_{t+1} & \cdots
\end{tikzcd}$$

In this special case:

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

BGDIA703 2023/2024

Examples of partially observed multivariate time series (cont.)

- Hidden Markov model.

$$
\begin{tikzcd}
\cdots \arrow[r, dashed] \arrow[d, dashed] & X_t \arrow[r] \arrow[d, dashed] & X_{t+1} \arrow[r, dashed] \arrow[d] & \cdots \arrow[d, dashed] \\
\cdots & Y_t & Y_{t+1} & \cdots
\end{tikzcd}
$$

  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.

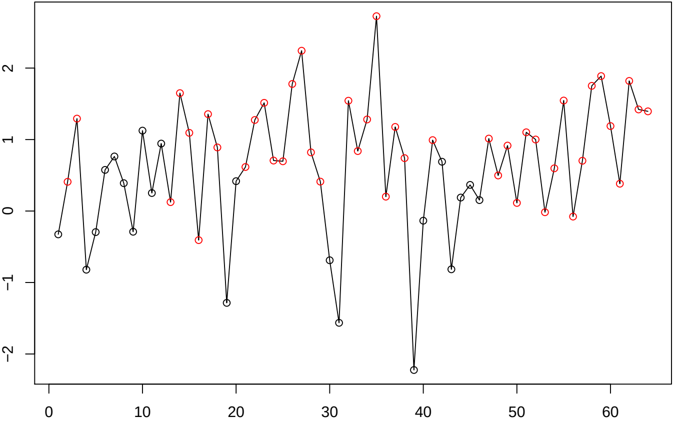


Figure: An HMM with two (supposedly) hidden states (red and black).

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

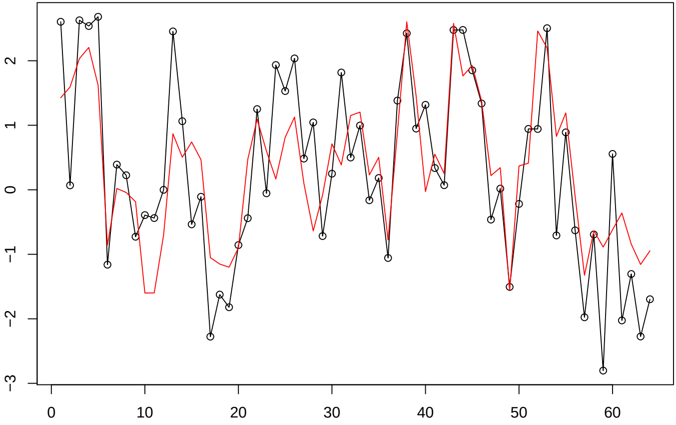


Figure: Observations (black 'o') obtained by adding noise to a (supposedly) hidden AR(1) process (red lines).

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:

$$\begin{tikzcd}
\cdots \arrow[r, dashed] \arrow[d, dashed] & X_t \arrow[r] \arrow[d, dashed] & X_{t+1} \arrow[r, dashed] \arrow[d] & \cdots \arrow[d, dashed] \\
\cdots \arrow[ur, dashed] & Y_t \arrow[ur] & Y_{t+1} \arrow[ur, dashed] & \cdots
\end{tikzcd}$$

With the additional property that the conditional distribution of X_{t+1} given (X_t, Y_t) 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

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

$$Y_t = \sigma_t \epsilon_t \ , \tag{2b}$$

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

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

$$Y_t = \sigma_t \epsilon_t \ , \tag{2b}$$

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 $\mathcal{N}(0, \sigma_t^2)$.

Exemple: GARCH(1,1) model, likelihood

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

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

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

Exemple: GARCH(1,1) model, likelihood

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

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

Note that $\psi^\theta < Y_{1:t-1} > (\sigma_1^2)$ 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

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

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