Introduction à l'analyse des séries temporelles

Felix Cheysson
Chercheur postdoctoral en biostatistiques

Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation
felix.cheysson@sorbonne-universite.fr

Année universitaire 2021-2022

Modalités d'enseignement et d'évaluation

Enseignement

- 2 $\times$ 3h de cours théoriques (ici, faculté de médecine du KB).
- 4 $\times$ 3h de TPs pratiques sous R (ici, KB-TP2 ou TP3).
  - TPs dédoublés en demi-groupes ?
- Supports de cours, exercices corrigés, annales et TPs disponibles sur `https://fcheysson.github.io/serc.html`.

Evaluation

- Quiz (questions de cours) au premier TP : 15 minutes, $\sim$ 3 points.
  - Contenu indispensable signalé par l'icône 📖.
- Devoir collaboratif au dernier TP : 45 minutes, $\sim$ 7 points.
- Devoir maison : 1 mois, $\sim$ 10 points.

Rappels : variable aléatoire

- **Formellement**. Une variable aléatoire X est une application d'un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ dans un espace mesurable $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.
- **Intuitivement**. Une variable aléatoire est une variable dont la valeur dépend d'événements inconnus.
- On caractérise une variable aléatoire par :
  - les différentes valeurs qu'elle peut prendre ;
  - la probabilité qu'elle prenne ces valeurs.

- Fonction de répartition F_X :

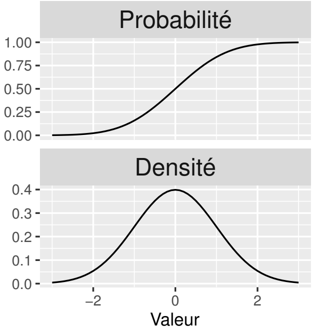
$$F_X(t) = \mathbb{P}(X \leq t);$$

- Fonction de densité f_X :

$$F_X(t) = \int_{-\infty}^{t} f_X(u)\mathrm{d}u.$$

- Fonction quantile :

$$Q(p) = \inf\{x \in \mathbb{R}; p \leq F_X(x)\} \approx F_X^{-1}(p)$$



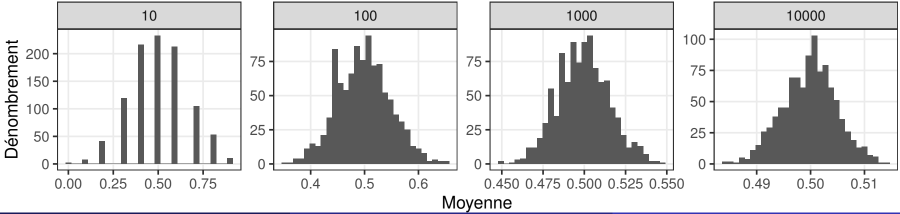
Rappels : loi normale $\mathcal{N}(\mu, \sigma^2)$

- Fonction de densité :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

- Caractérisée complètement par son espérance μ et sa variance σ^2.
- Apparait dans le **théorème central limite** : Soit une suite de variables aléatoires $X_1, X_2, \ldots$ indépendantes et identiquement distribuées (*i.i.d.*), d'espérance μ et de variance $\sigma^2 < \infty$. Alors

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2).$$



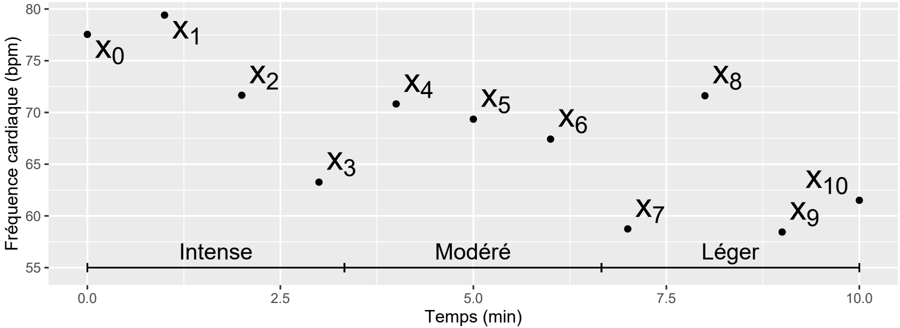
Définition : série temporelle

- **Définition.** On appelle série temporelle une collection de variables aléatoires indexées par le temps :

$$(x_t)_{t\in\mathbb{N}} = (x_0, x_1, x_2, \ldots).$$

- **Exemple.** On mesure la fréquence cardiaque d'un individu toutes les minutes pendant 10 minutes au cours d'un effort :

$$(x_t)_{0\leq t\leq 10} = (x_0, x_1, \ldots, x_{10}).$$



Problématique

- Lors d'un test d'effort, on cherche à exprimer la fréquence cardiaque d'un individu en fonction de l'effort réalisé par ce dernier.
- **Naïvement**. On effectue une ANOVA en fonction de l'effort :

$$x_t = \mu + \alpha_{\text{effort}} + \varepsilon_t,$$

où ε_t désigne une variable d'erreur gaussienne : $\varepsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

Question

Quel est le problème d'une telle modélisation ? Quelle hypothèse n'est pas vérifiée dans ce modèle ?

Motivation (application shiny)

PLOS ONE

RESEARCH ARTICLE

ICU admissions and in-hospital deaths linked to COVID-19 in the Paris region are correlated with previously observed ambient temperature

Mehdi Mejdoubi[1]*, Xavier Kyndt[2], Mehdi Djennaoui[2]

1 Department of Radiology, Hospital of Valenciennes, Valenciennes, France, 2 Department of Public Health, Hospital of Valenciennes, Valenciennes, France

* mejdoubi-m@ch-valenciennes.fr

Check for updates

Abstract

The purpose of this ecological study was to explore the association of weather with severity indicators of coronavirus disease 2019 (COVID-19). Daily COVID-19-related intensive care unit (ICU) admissions and in-hospital deaths in the Paris region and the daily weather characteristics of Paris midtown were correlated with a time lag. We assessed different study periods (41, 45, 50, 55, and 62 days) beginning from 31 March 2020. Daily ICU admissions and in-hospital deaths were strongly and negatively correlated to ambient temperatures (minimal, average, and maximal). The highest Pearson correlation coefficients and statistically significant p values were found 8 days before the occurrence of ICU admissions and 15 days before deaths. Partial correlations with adjustment on days since lockdown showed similar significant results. The study findings show a negative correlation of previously observed ambient temperature with severity indicators of COVID-19 that could partly explain the death toll discrepancies between and within countries.

OPEN ACCESS

Citation: Mejdoubi M, Kyndt X, Djennaoui M (2020) ICU admissions and in-hospital deaths linked to COVID-19 in the Paris region are correlated with previously observed ambient temperature. PLoS ONE 15(11): e0242268. https://doi.org/10.1371/journal.pone.0242268

Editor: Corstiaan den Uil, Erasmus Medical Center, NETHERLANDS

Received: August 3, 2020

Accepted: October 22, 2020

Published: November 20, 2020

Data Availability Statement: The data used in this study are public and freely accessible. Ambient temperature data are available from http://www.meteofrance.com and daily numbers of COVID-19 deaths are available from https://geodes.santepubliquefrance.fr.

Funding: The author(s) received no specific funding for this work.

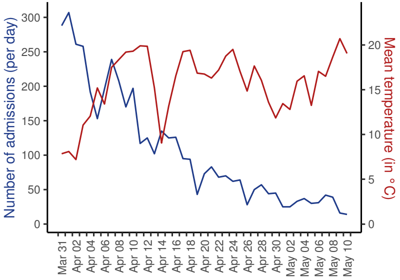
Competing interests: The authors have declared that no competing interests exist.

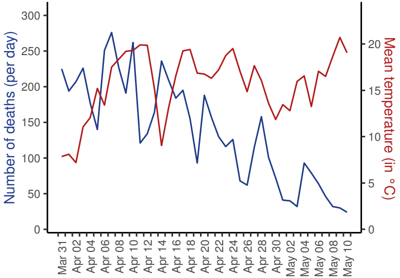
Introduction

The ongoing major epidemic of coronavirus disease 2019 (COVID-19) infection began in cold areas of temperate countries of the northern hemisphere [1], which raises the question of weather influence.

Continental France has been severely hit by this outbreak of COVID-19, with around 30 000 deaths, including more than 19 000 in-hospital deaths (as of 15 June 2020). The main affected regions were its eastern region (mainly Alsace) and the Paris region (Île-de-France), with more than 3500 and 7200 in-hospital deaths, respectively.

Monitoring the outbreak of COVID-19 with biological testing is imprecise as the infection can be asymptomatic [2]. Data such as deaths or intensive care unit (ICU) admissions (reflecting critically ill patients) are more reliable, especially in highly developed countries. In the natural course of infection with COVID-19, a pulmonary worsening may appear around 5 to 9 days after infection onset [2, 3], with an immunological mechanism known as "cytokine





Rappels : tests d'hypothèse

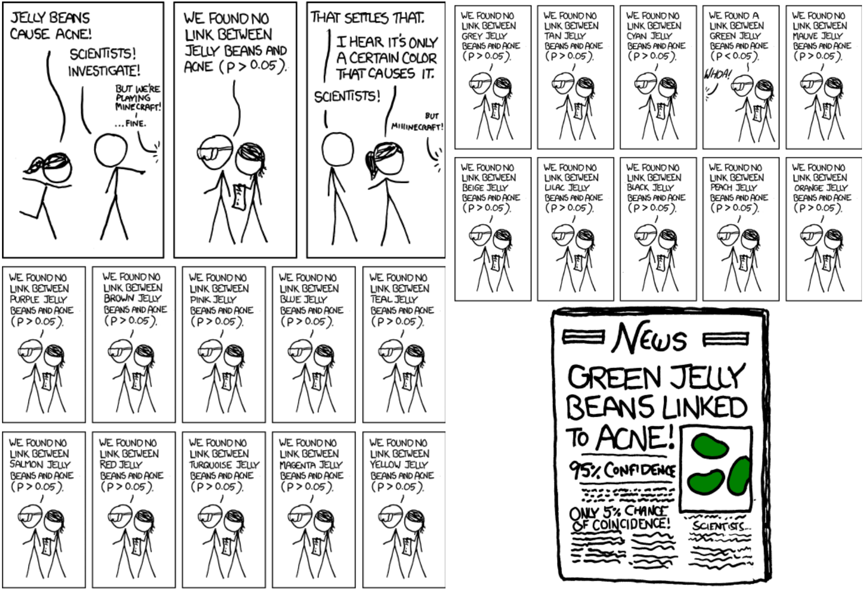
Procédure de décision permettant de rejeter une hypothèse sur la base de l'étude statistique d'un échantillon $(X_1, \dots, X_n)$.

- Hypothèse nulle $\mathcal{H}_0$:
  - En général choisie comme l'inverse de ce que l'on veut montrer.
  - Hypothèse pour laquelle les calculs statistiques sont explicites.
- Statistique de test et distribution sous H_0.
  - Détermine si l'échantillon a de fortes chances d'être observé sous $\mathcal{H}_0$.
  - Ne représente qu'un cliché des caractéristiques de l'échantillon.
- Région de rejet et p-valeur.
  - Permettent de conclure quant au rejet ou non de $\mathcal{H}_0$.

Risque de première espèce α

Probabilité de rejeter $\mathcal{H}_0$ à tort : attention aux tests multiples, aux fausses interprétations, aux idées reçues, etc.

p-value : erreurs d'interprétation (credits to xkcd)



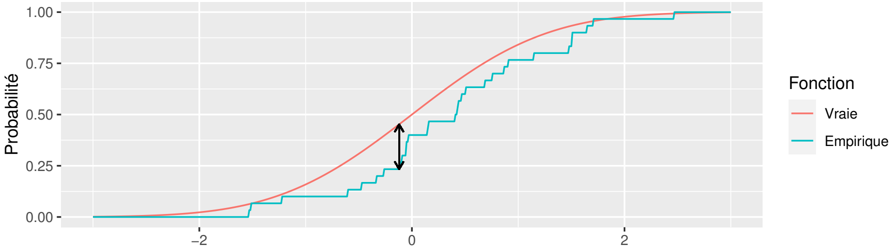
Exemple : test de Kolmogorov-Smirnov

Permet de tester si l'échantillon provient d'une distribution F déterminée.

- Hypothèse nulle $\mathcal{H}_0$: $\forall x \in \mathbb{R}, F_n(x) = F(x)$.
- Statistique de test : $D_n = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$.
- Loi sous $\mathcal{H}_0$:

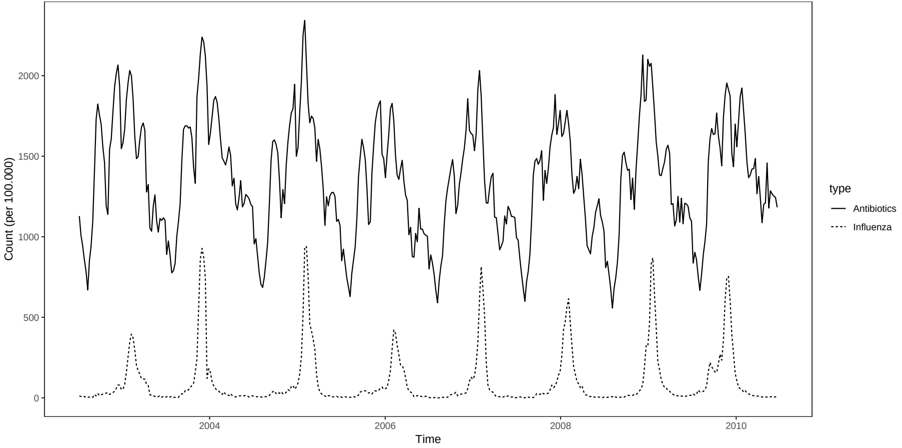
$$\sqrt{n} D_n \xrightarrow[n\to\infty]{d} K = \sup_{t\in[0,1]} |B(F(t))|,$$

  où $B(\cdot)$ correspond au pont brownien.
- Région de rejet au niveau de risque α : $\sqrt{n}D_n > K_{1-\alpha}$, où $K_{1-\alpha}$ représente le quantile de niveau $1-\alpha$ de la variable K.



Exemple de données de surveillance

- Nombre de remboursements des prescriptions d'antibiotiques en ville.
  - Données du SNIIRAM (CNAMTS + RSI) : 90% de la population.
- Nombre de syndromes grippaux.
  - Données du Réseau Sentinelles : représentatives à l'échelle de la France.



Surveillance épidémiologique

Qualité de l'indicateur

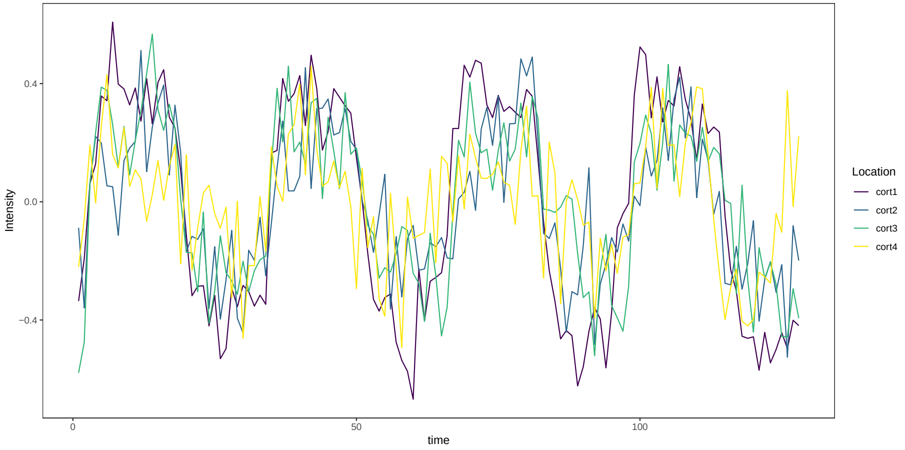
- **Représentatif ?**
  - Surveillance passive : sous-estimation.
  - Biais de recrutement.
- **Stabilité temporelle ?**
  - Evolution de l'indicateur dans le temps.

Qualité de la modélisation

- Caractérisation simple de la série (tendance, saison, variations accidentelles, points de rupture, ...).
- Représentation de la série par un modèle statistique (ARIMA).
- Prédiction de la série à partir de ses valeurs passées.

Exemple de données biologiques

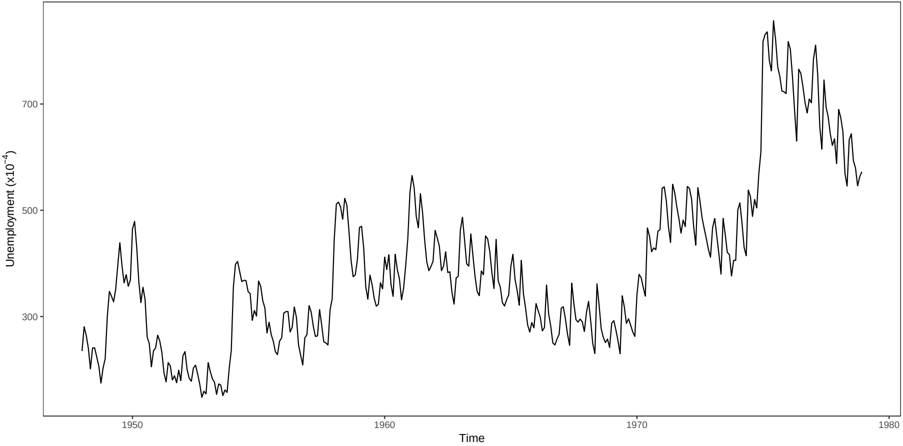
Intensité du signal BOLD (Blood Oxygenation-Level Dependent) dans le cortex d'un patient soumis à des stimuli, collectée par fMRI.



Exemple de données économiques

Nombre de personnes sans emploi aux Etats-Unis entre 1948 et 1978.

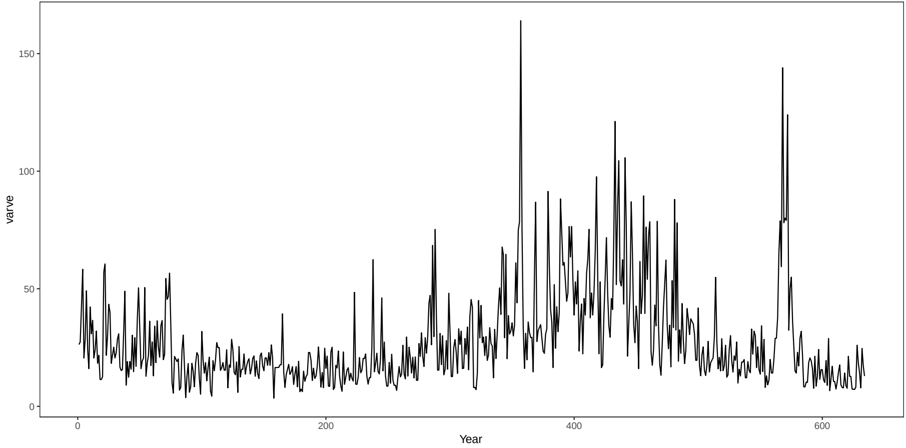
- Impact du choc pétrolier de 1973.



Exemple de données géologiques

Épaisseur de la couche sédimentaire.

- Recueillie au Massachussetts.
- Représentative de la température il y a 11.834 ans.



Motivation et enjeux

On dispose d'une ou de deux séries que l'on veut modéliser :

$$x_t = T_t + S_t + \nu_t$$
$$y_t = T'_t + S'_t + \beta x_t + \upsilon_t,$$

où

- T_t et T'_t désignent des **tendances** (évolution à long terme) ;
- S_t et S'_t désignent des **saisonnalités** (évolution périodique) ;
- ν_t et υ_t désignent des variables **résiduelles autodépendantes**, *i.e.*

$$\forall s \neq t, \quad \nu_s \not\perp \nu_t \quad \text{et} \quad \upsilon_s \not\perp \upsilon_t.$$

Remarque : On parle de variables d'**innovation**, ou de **bruit**.

Problématique

Prendre en compte la structure d'autodépendance des innovations.

Objectifs

<u>Applications</u>

- Identifier la nécessité d'un modèle de séries temporelles.
- Modéliser les séries temporelles simples (ARIMA).
  - Nomenclature, démarche, *etc.*
  - Programmation en R.
- Connaître ses limites.

<u>Théoriques</u>

- Connaître les notions statistiques sous-jacentes aux modèles de séries temporelles.
- Comprendre les résultats mathématiques permettant l'analyse des séries temporelles.
  - Stationnarité d'une série.
  - Théorème de représentation de Wold.

Contenu

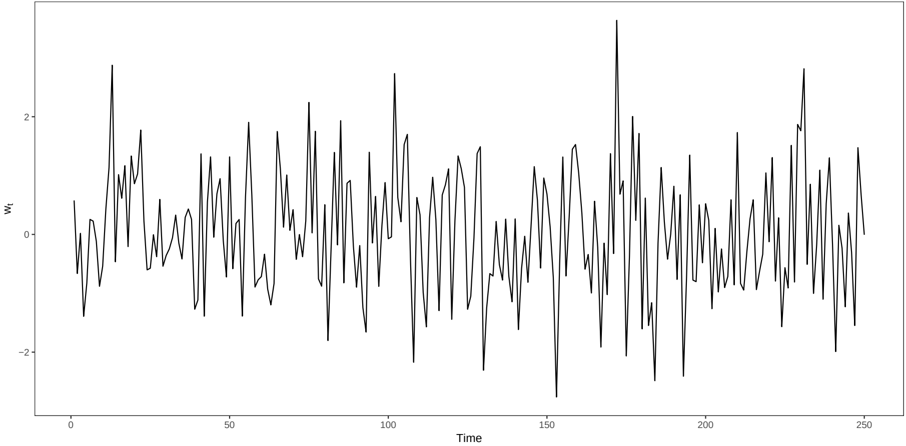
Outline

1. Exemples
   - Bruit blanc Gaussien
   - Moyenne mobile
   - Processus autorégressif
   - Marche aléatoire avec dérive
   - Signal bruité

Exemple 1.1 : Bruit blanc Gaussien

Collection de variables aléatoires normales *indépendantes* et *identiquement distribuées* :

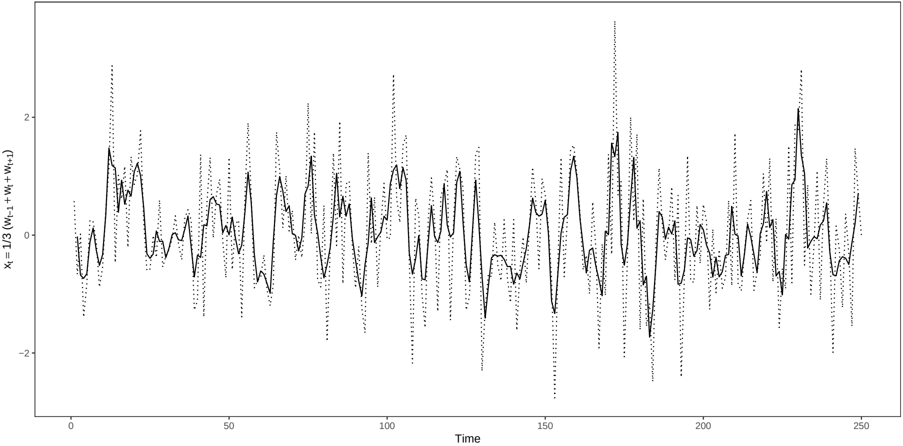
$$w_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_w^2)$$



Exemple 1.2 : Moyenne mobile

Combinaison linéaire d'un bruit blanc Gaussien :

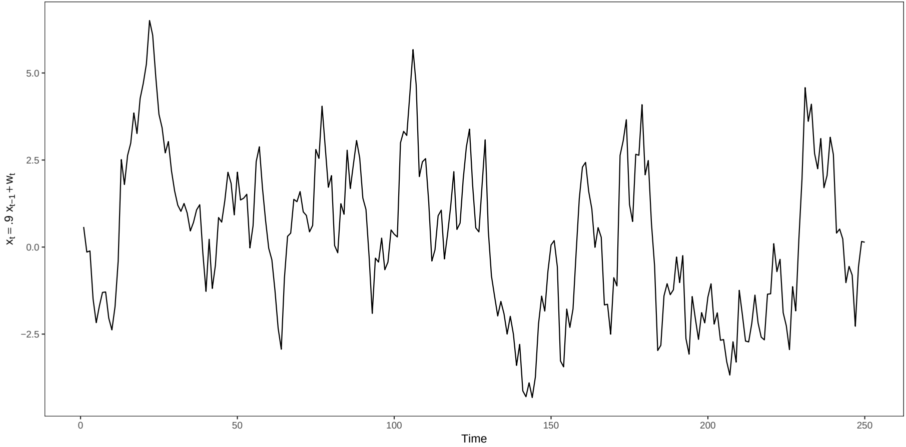
$$x_t = \frac{1}{3}\left(w_{t-1} + w_t + w_{t+1}\right)$$



Exemple 1.3 : Processus autorégressif

Régression sur les valeurs passées de la série :

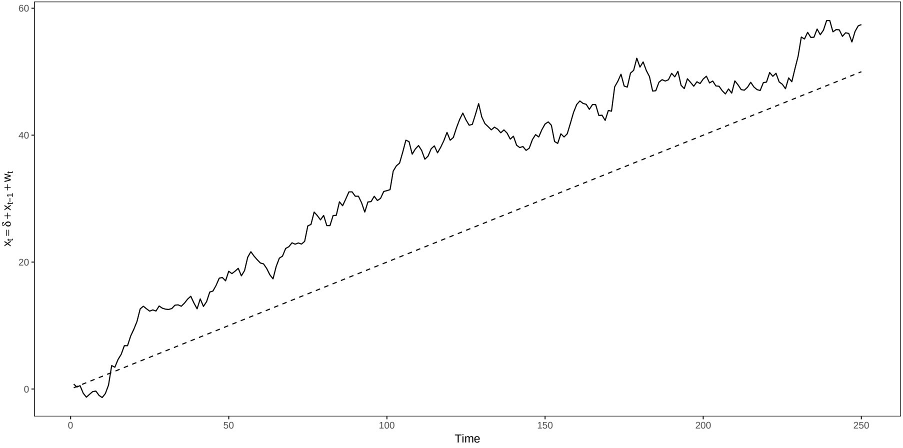
$$x_t = 0.9\, x_{t-1} + w_t$$



Exemple 1.4 : Marche aléatoire avec dérive

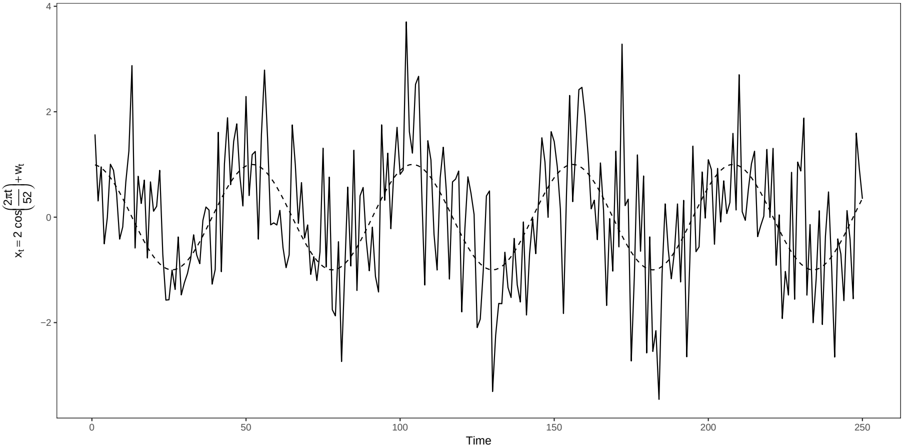
Succession de pas aléatoire avec dérive δ :

$$x_t = \delta + x_{t-1} + w_t$$



Signal périodique bruité :

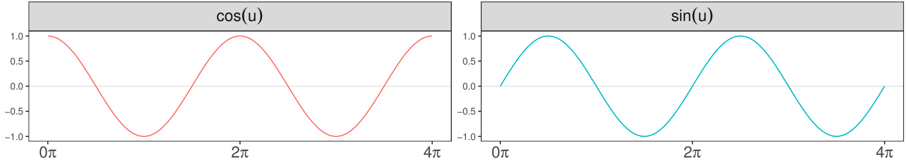
$$x_t = 2\cos\left(\frac{2\pi t}{52}\right) + w_t$$



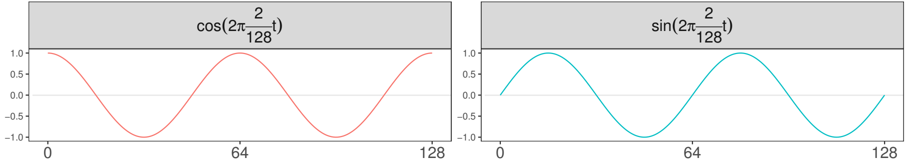
Sinus et cosinus

Les fonctions sin and cos sont 2π-périodiques : pour $u \in \mathbb{R}$ et $k \in \mathbb{Z}$,

$$\cos(u + 2k\pi) = \cos(u).$$



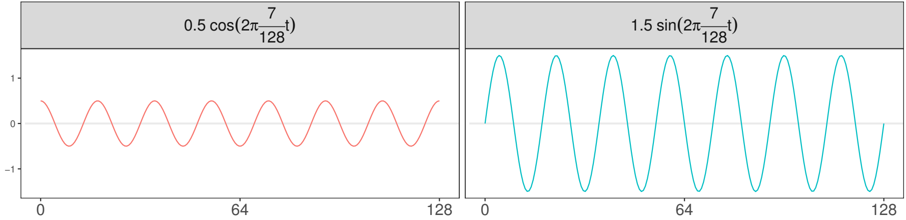
Soit $u = 2\pi \frac{2}{128} t$ pour $t = 1, 2, \ldots, 128$.



On peut interpréter $\frac{2}{128}$ comme 2 cycles pour 128 unités de temps.

Sinus et cosinus

Similairement, avec $u = 2\pi\frac{7}{128}t$ pour $t = 1, 2, \ldots, 128$.



- $\cos(2\pi\frac{k}{n}t)$ et $\sin(2\pi\frac{k}{n}t)$ possèdent k cycles pour n unités de temps.
- La quantité $f = \frac{k}{n}$ est appelée la *fréquence* du sinus ou du cosinus.
  - Elle représente le nombre (ou la fraction) de cycles par unité de temps.
  - Si f est faible (resp. élevée), le sinus est dit de basse (haute) fréquence.
- La *période* $T = \frac{1}{f}$ dénombre les unités de temps nécessaires à un cycle.
- L'*amplitude* est l'étendue maximale de la variation et est égale à 1 pour les fonctions $\sin$ et $\cos$.

Outline

2 Notions théoriques essentielles
- Fonction d'autocovariance
- Fonction d'autocorrélation

- **Espérance.** (ex. 2, 4, 5)

$$\mu_t = \mathbb{E}[x_t]$$

- **Fonction d'autocovariance** ou produit moment d'ordre 2. (ex. 1)

$$\begin{aligned} \gamma(s,t) &= \mathrm{Cov}(x_s, x_t) \\ &= \mathbb{E}\big[(x_s - \mu_s)(x_t - \mu_t)\big] \end{aligned}$$

- *Remarque*. La fonction d'autocovariance désigne la dépendance linéaire entre deux temps.
- *Remarque*. $\gamma(t,t) = \mathrm{Var}(x_t)$.

- **Propriétés**. Soient A, B et C trois variables aléatoires et α un réel. (ex. 2, 3, 4)

$$\begin{aligned} \operatorname{Cov}(A, B) &= \operatorname{Cov}(B, A) \\ \operatorname{Cov}(A + B, C) &= \operatorname{Cov}(A, C) + \operatorname{Cov}(B, C) \\ \operatorname{Cov}(\alpha A, B) &= \alpha \operatorname{Cov}(A, B) \end{aligned}$$

- **Fonction d'autocorrélation (ACF)**. (ex. 2, 3)

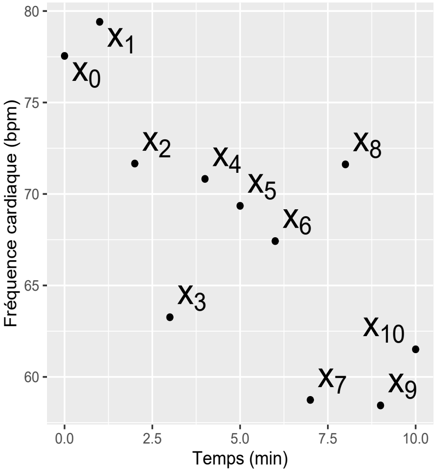
$$\rho(s, t) = \frac{\gamma(s, t)}{\sqrt{\gamma(s, s)\gamma(t, t)}}$$

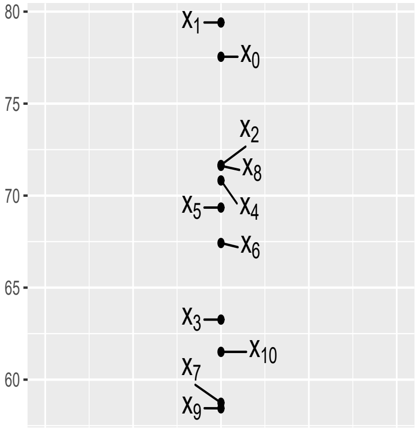
- *Remarque*. La fonction d'autocorrélation caractérise la prédictabilité linéaire de la série au temps t sachant la valeur de x_s.

Outline

Stationnarité : problématique

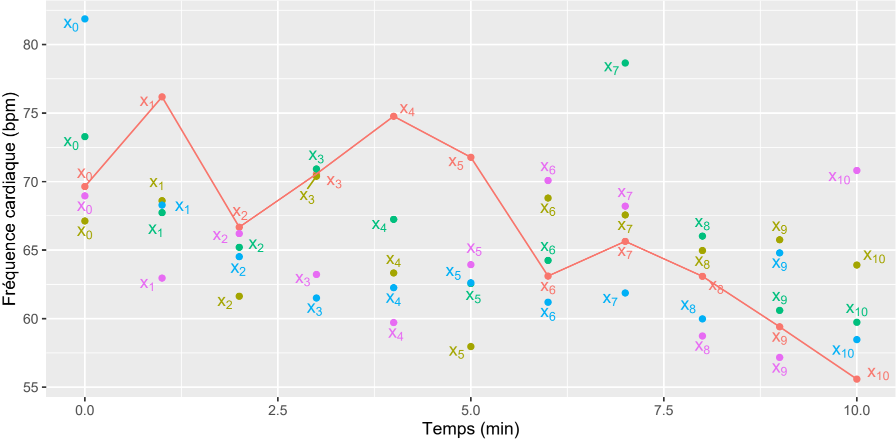
Pour la régression linéaire, toutes les observations sont indépendantes et identiquement distribuées.





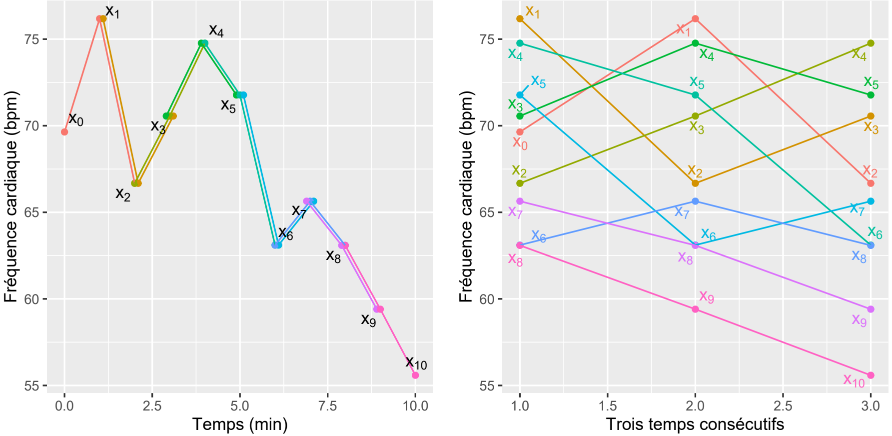
Stationnarité : problématique (cont'd)

Pour la plupart des séries temporelles, les observations ne sont pas indépendantes.



On dit que la série (x_t) est **stationnaire au sens strict** si le comportement probabiliste de toute collection $(x_{t_1}, \dots, x_{t_k})$ est invariant par translation.

$\rightarrow$ Pour tout h, le comportement probabiliste de $(x_{t_1}, \dots, x_{t_k})$ est identique à $(x_{t_1+h}, \dots, x_{t_k+h})$.



Stationnarité au sens faible (ou du 2ème ordre)

On dit que la série (x_t) est **stationnaire au sens faible** si ses deux premiers moments sont invariants par translation :

- $\mu_t = \mathbb{E}[x_t]$ ne dépend pas de t ;
- $\gamma(s,t) = \mathrm{Cov}(x_s, x_t)$ ne dépend que de $|s-t|$.

Il s'agit d'une propriété plus faible que la stationnarité stricte :

$$\begin{aligned}
\text{stationnarité stricte} &\quad\Longrightarrow\quad \text{stationnarité faible},\\
\text{stationnarité stricte} &\quad\not\Longleftarrow\quad \text{stationnarité faible},\\
\text{stationnarité stricte} &\quad\underset{x_t\ \text{gaussien}}{\Longleftarrow}\quad \text{stationnarité faible}.
\end{aligned}$$

Enjeu de la propriété de stationnarité

Pour une série stationnaire, il est possible d'identifier et d'estimer les caractéristiques statistiques de la série à partir d'une unique trajectoire.

Stationnarité au sens faible (cont'd)

- **Espérance.** $\mathbb{E}[x_t] = \mu$;
- **Fonction d'autocovariance.**

$$\begin{aligned}\gamma(t+h,t) &= \operatorname{Cov}(x_{t+h}, x_t) \\ &= \operatorname{Cov}(x_h, x_0) \\ &= \gamma(h,0) \\ &=: \gamma(h) \ ;\end{aligned}$$

- **Variance.** $\operatorname{Var}(x_t) = \gamma(t,t) = \gamma(0,0) = \gamma(0)$;
- **Fonction d'autocorrélation.** $\rho(h) := \rho(t+h,t) = \cdots = \gamma(h)/\gamma(0)$.
- *Exemples.* Les séries 1, 2, 4 et 5 sont-elles stationnaires ?
- *Exemple.* Sous l'hypothèse de stationnarité, calculer l'espérance et la variance pour la série 3.

Outline

4. Estimateurs de la corrélation
   - Définitions
   - Exemples

Estimateurs de la corrélation

- **Fonction d'autocovariance empirique.** Il s'agit de l'estimateur défini par :

$$\widehat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x}).$$

- **Fonction d'autocorrélation (ACF) empirique.** Il s'agit de l'estimateur précédent normalisé :

$$\widehat{\rho}(h) = \widehat{\gamma}(h)/\widehat{\gamma}(0).$$

Propriétés asymptotiques de l'ACF

Pour un bruit blanc Gaussien (w_t), si n est "grand", la loi de l'ACF empirique peut être approchée par :

$$\widehat{\rho}_w(h) \sim \mathcal{N}\left(0, \sigma^2_{\widehat{\rho}_w(h)}\right),$$

où $\sigma_{\widehat{\rho}_w(h)} = \frac{1}{\sqrt{n}}$.

Test d'hypothèse sur l'autocorrélation

Motivation : Permet de déterminer si l'autocorrélation $\rho(\cdot)$ à un décalage h fixé est différent de 0.

- Hypothèse nulle $\mathcal{H}_0$: $\rho(h) = 0$.
- Statistique de test :

$$Z = \frac{\widehat{\rho}(h) - \rho(h)}{\widehat{\sigma}_{\widehat{\rho}(h)}} \underset{\mathcal{H}_0}{=} \frac{\widehat{\rho}(h)}{\widehat{\sigma}_{\widehat{\rho}(h)}}.$$

- Loi *asymptotique* sous $\mathcal{H}_0$:

$$Z \sim \mathcal{N}(0, 1).$$

- Région de rejet au risque α : $|Z| > u_{1-\alpha/2}$, où $u_{1-\alpha/2}$ désigne le quantile de niveau $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$.

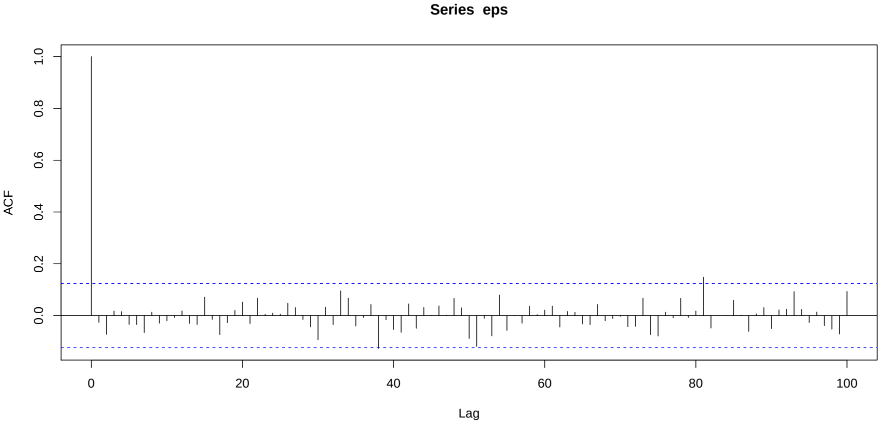
Rejet de $\mathcal{H}_0$

L'existence d'un pic d'autocorrélation $\rho(h)$ non nul représente une association statistique entre deux observations de la série séparées de h unités de temps.

ACF 4.1 : Bruit blanc Gaussien

Collection de variables aléatoires normales *indépendantes* et *identiquement distribuées* :

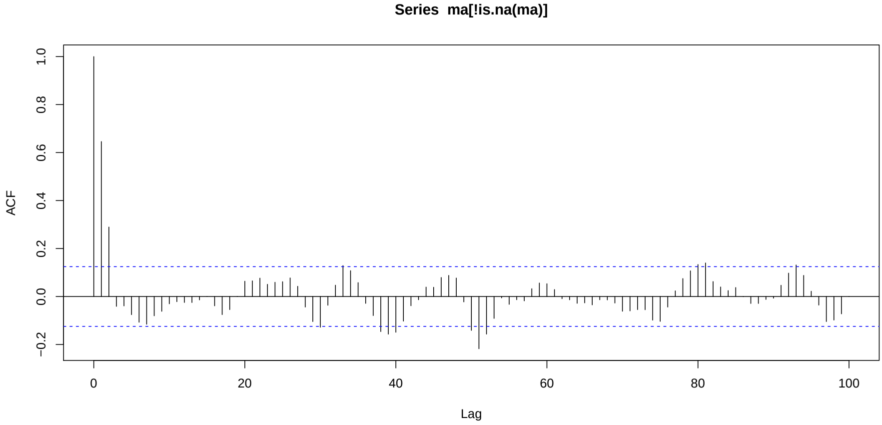
$$w_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_w^2)$$



ACF 4.2 : Moyenne mobile

Combinaison linéaire d'un bruit blanc Gaussien :

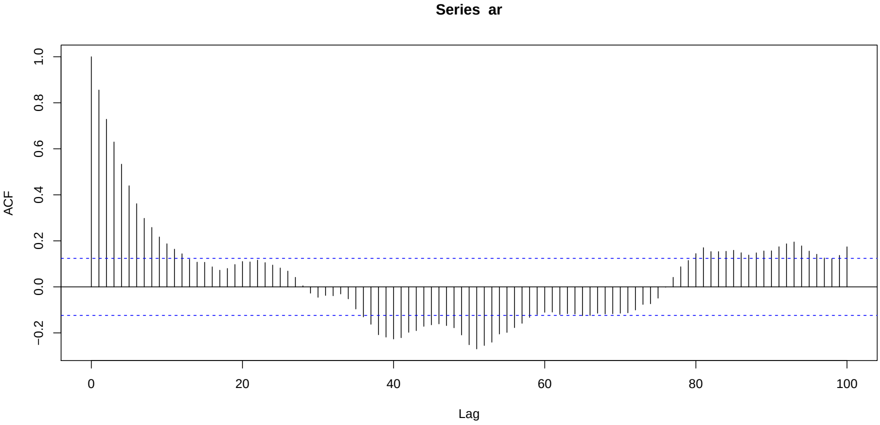
$$x_t = \frac{1}{3}\left(w_{t-1} + w_t + w_{t+1}\right)$$



ACF 4.3 : Processus autorégressif

Régression sur les valeurs passées de la série :

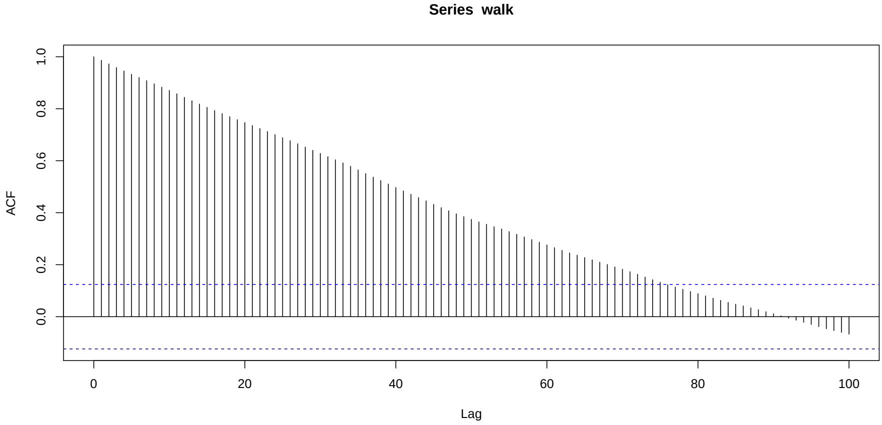
$$x_t = 0.9\,x_{t-1} + w_t$$



ACF 4.4 : Marche aléatoire avec dérive

Succession de pas aléatoire avec dérive δ :

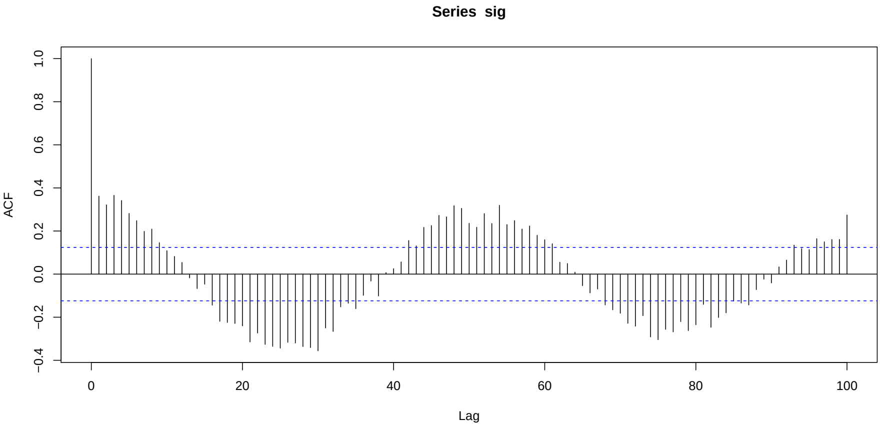
$$x_t = \delta + x_{t-1} + w_t$$



ACF 4.5 : Signal bruité

Signal périodique bruité :

$$x_t = 2\cos\left(\frac{2\pi t}{52}\right) + w_t$$



Outline

5. Données non stationnaires
   - Prise en compte d'une tendance par régression
   - Prise en compte d'une tendance par différenciation
   - Stabilisation de la variance

Données non stationnaires

Problématique

En pratique, la plupart des jeux de données ne sont pas issus de processus stationnaires.

Méthode. On va chercher à *stationnariser* (rendre stationnaire) la série de données afin de pouvoir analyser sa structure de dépendance.

- Prise en compte d'une tendance par régression ;
- Prise en compte d'une tendance par différentiation.
    - Tendance linéaire ;
    - Tendance stochastique ;
    - Tendance saisonnière.

Prise en compte d'une tendance par régression

On considére une série du type $x_t = \mu_t + \nu_t$, où ν_t est stationnaire.

Par exemple :

- Régression linéaire avec erreurs stationnaires : $x_t = \beta_0 + \beta_1 t + \nu_t$;
- Signal bruité : $x_t = 2\cos\left(\frac{2\pi t}{52}\right) + w_t$

Démarche de modélisation :

1. Modéliser μ_t, et effectuer la régression linéaire $x_t = \mu_t + \varepsilon_t$;
2. Etudier la structure de dépendance des résidus $\widehat{\nu}_t = x_t - \widehat{\mu}_t$;
3. Utiliser la structure identifiée pour estimer conjointement les paramètres de la régression et de la structure de dépendance.

Prise en compte d'une tendance par différenciation

- **Principe**. On retire la tendance en étudiant les différences de la série.
- On note B l'**opérateur retard** défini par :

$$
\begin{aligned}
B \ x_t &= x_{t-1}, \\
B^2 x_t &= B(Bx_t) = B(x_{t-1}) = x_{t-2}, \\
B^k x_t &= x_{t-k}.
\end{aligned}
$$

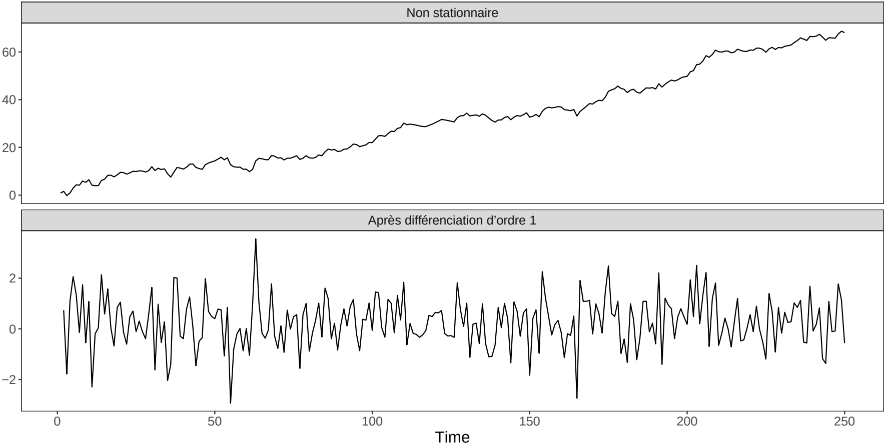
- On note $\nabla := 1 - B$ l'opérateur de différence d'ordre 1 :

$$
\begin{aligned}
\nabla \ x_t &= (1-B) \ x_t = x_t - Bx_t = x_t - x_{t-1}, \\
\nabla^2 x_t &= (1-B)^2 x_t = (1 - 2B + B^2) x_t = x_t - 2x_{t-1} + x_{t-2}.
\end{aligned}
$$

- *Exemple*. Etudier la stationnarité de la série 4 par différenciation.
- *Exemple*. Etudier la stationnarité de la régression linéaire avec erreurs stationnaires par différenciation.

Différentiation de la marche aléatoire

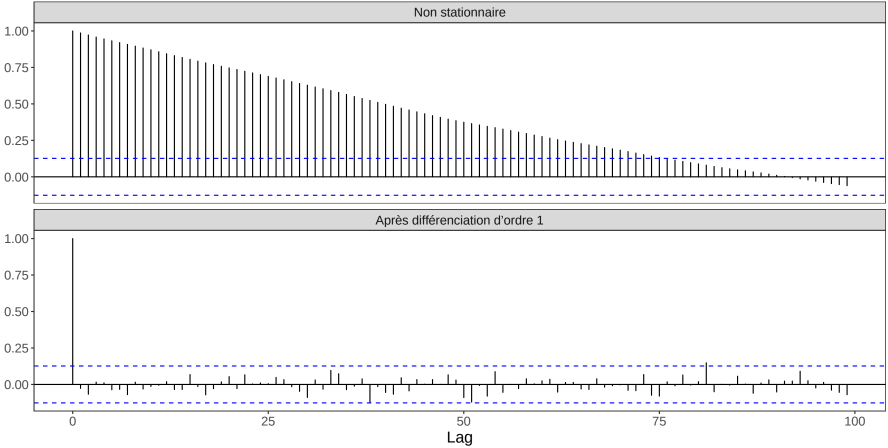
$$x_t = \delta + x_{t-1} + w_t$$



Stationnarité : Marche aléatoire avec dérive

Succession de pas aléatoire avec dérive δ :

$$x_t = \delta + x_{t-1} + w_t$$



Prise en compte de la tendance par différenciation

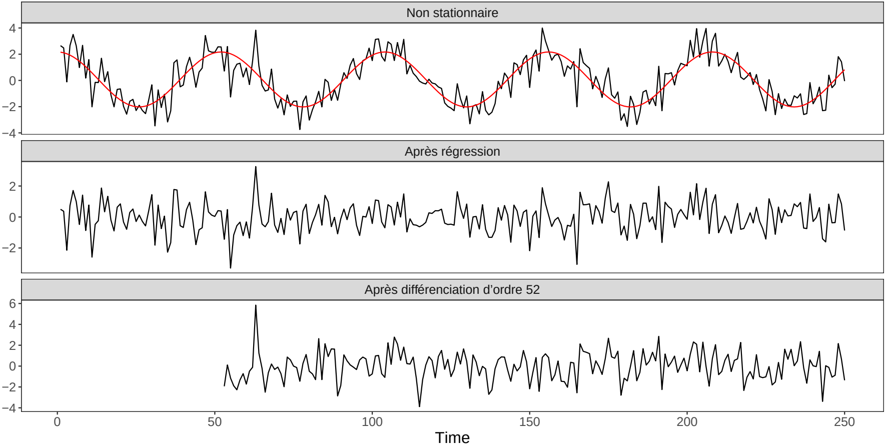
- Pour une série saisonnière, on pourra étudier une différenciation de plus grand ordre.
- On note $\nabla_S = 1 - B^S$ l'opérateur de différence d'ordre S :

$$\nabla_S x_t = (1 - B^S) x_t = x_t - x_{t-S}.$$

- *Exemple*. Etudier la stationnarité de la série 5 par différenciation.

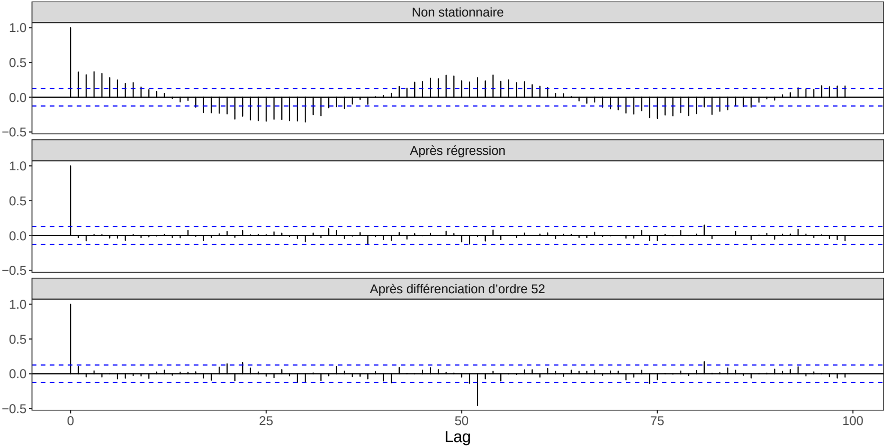
Différentiation ou régression du signal bruité

$$x_t = 2\cos\left(\frac{2\pi t}{52}\right) + w_t$$



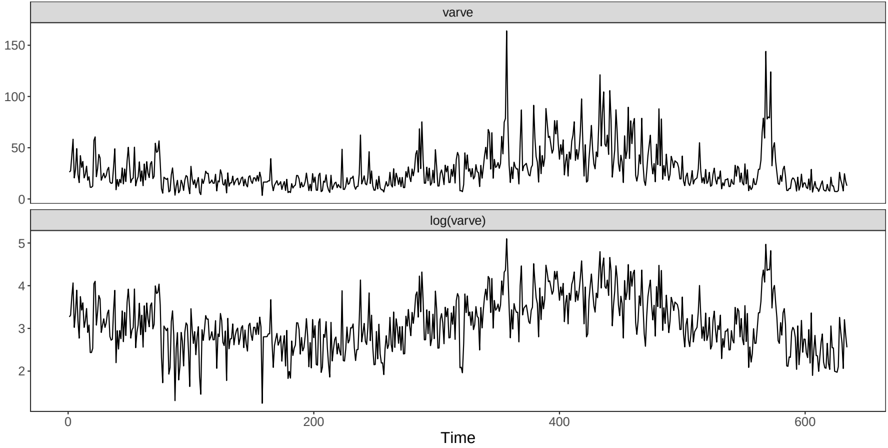
Signal périodique bruité :

$$x_t = 2\cos\left(\frac{2\pi t}{52}\right) + w_t$$



Stabilisation de la variance

- Pour des séries dont la variabilité augmente avec l'espérance, on peut stabiliser la variance en transformant la série par un logarithme.
- *Exemple.* $y_t = \log x_t$ avec x_t la série des couches sédimentaires.



Outline

- **Définition**. On appelle processus moyenne mobile d'ordre q, noté MA(q), le processus x_t défini par :

$$\begin{aligned} x_t &= w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q} \\ &= w_t + \sum_{j=1}^{q} \theta_j w_{t-j}, \end{aligned}$$

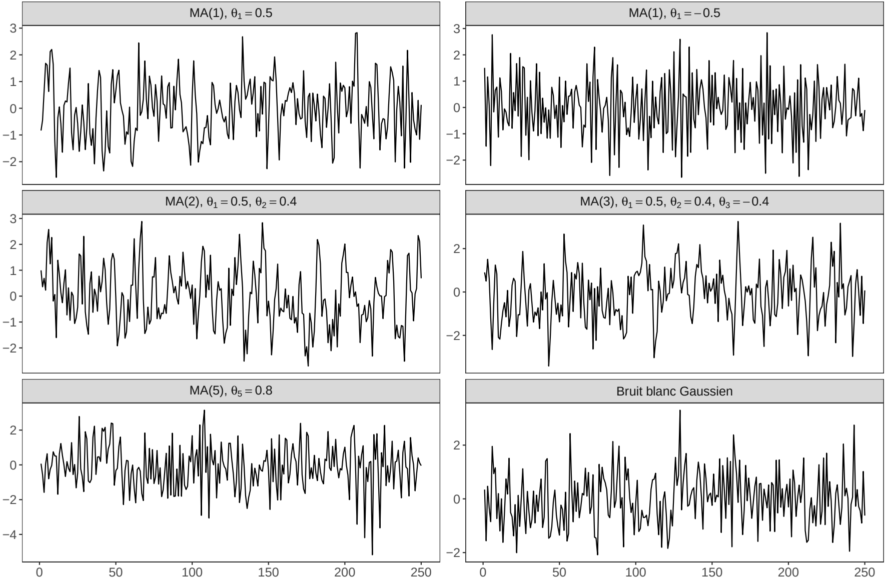
où $\theta_q \neq 0$ et $w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2)$.

- *Remarque*.

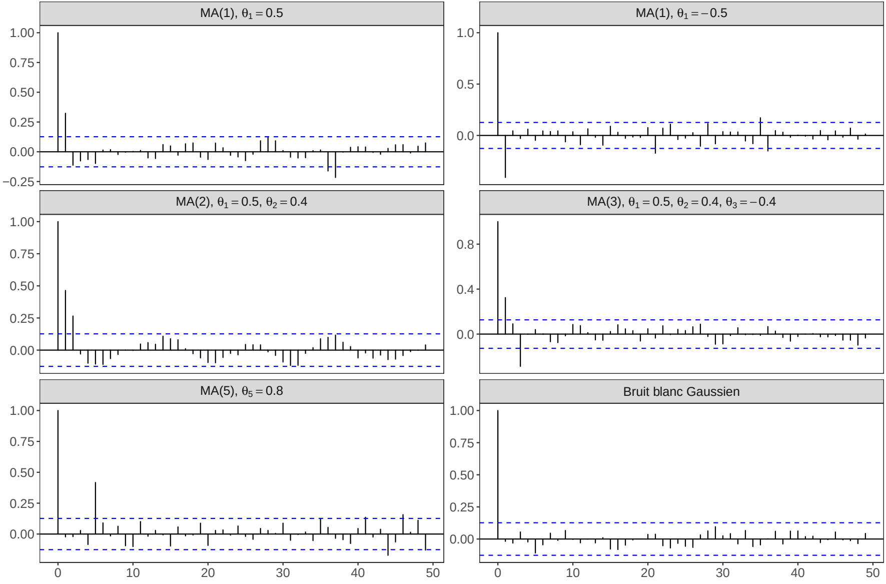
$$x_t = (1 + \theta_1 B + \ldots + \theta_q B^q) w_t$$
$$x_t = \Theta(B) w_t$$

- **Définition**. On appelle Θ le polynôme moyenne mobile.
- *Exemple*. MA(1) : $x_t = w_t + \theta w_{t-1}$. Que valent $\mathbb{E}[x_t]$ et $\rho(h)$?

Trajectoires de processus MA



ACF de processus MA



Modèle autorégressif

- **Définition**. On appelle processus autorégressif d'ordre p, noté AR(p), le processus stationnaire x_t vérifiant :

$$
\begin{aligned}
x_t &= w_t + \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} \\
&= w_t + \sum_{i=1}^{p} \phi_i x_{t-i},
\end{aligned}
$$

  où $\phi_p \neq 0$ et $w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2)$.
- *Remarque*. Si $\mathbb{E}[x_t] = \mu$, on considérera $y_t = x_t - \mu$.
- *Remarque*.

$$
\begin{aligned}
x_t &= (\phi_1 B + \ldots + \phi_p B^p) x_t + w_t \\
(1 - \phi_1 B - \ldots - \phi_p B^p) x_t &= w_t \\
\Phi(B) x_t &= w_t
\end{aligned}
$$

- **Définition**. On appelle Φ le polynôme autorégressif.

Ecriture d'un AR(1) comme processus linéaire

1. Soit x_t un AR(1). Prouver que :

$$x_t = \phi^k x_{t-k} + \sum_{j=0}^{k-1} \phi^j w_{t-j}.$$

2. Si $\mid \phi \mid < 1$ et x_t stationnaire, on peut donc représenter un AR(1) par un MA(∞) :

$$x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}.$$

3. En déduire :

$$\mathbb{E}[x_t] = 0,$$

$$\gamma(h) = \frac{\sigma_w^2 \phi^h}{1 - \phi^2},$$

$$\rho(h) = \phi^h.$$

Causalité

- **Définition**. Lorsque, pour un AR(p) de la forme $\Phi(B)x_t = w_t$, l'écriture

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} = \Psi(B)w_t$$

existe, on dira que le processus x_t est causal.

- *Remarque*. Dans ce cas, on a

$$\Phi^{-1}(B)\Phi(B)x_t = \Phi^{-1}(B)w_t,$$

soit

$$\Psi(B) = \Phi^{-1}(B).$$

Inversibilité

- **Définition**. Lorsque, pour un MA(q) de la forme $x_t = \Theta(B)w_t$, l'écriture

$$w_t = \sum_{j=0}^{\infty} \pi_j x_{t-j} = \Pi(B)x_t$$

existe, on dira que le processus x_t est inversible.

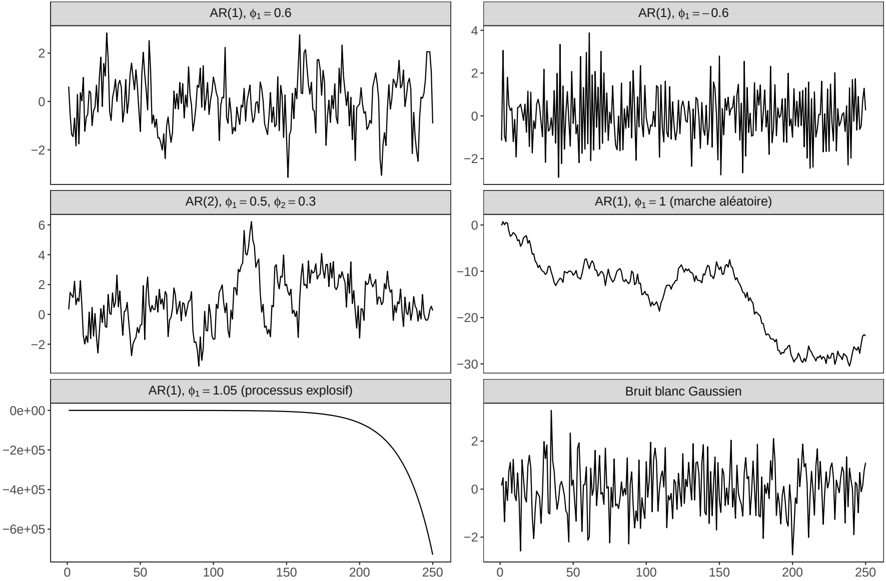
- *Remarque*. Dans ce cas, on a

$$\Theta^{-1}(B)x_t = \Theta^{-1}(B)\Theta(B)w_t,$$

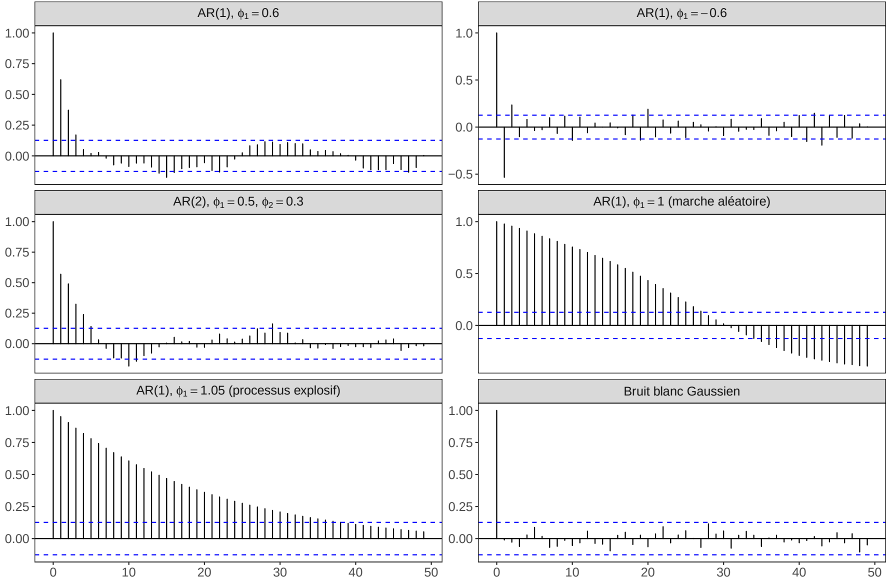
soit

$$\Pi(B) = \Theta^{-1}(B).$$

Trajectoires de processus AR



ACF de processus AR



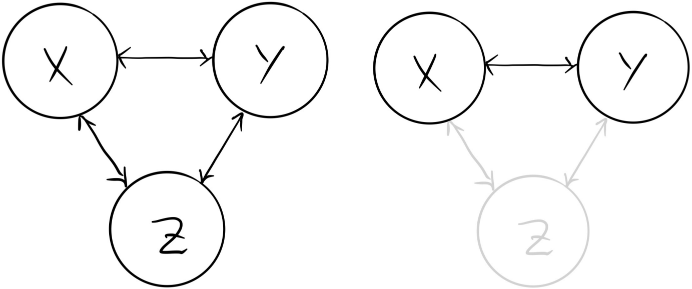
Corrélation partielle

- *Idée.* Contrôler les facteurs de confusion dans le calcul de corrélation.
- **Définition.** Le coefficient de corrélation partielle entre deux variables aléatoires X et Y sachant une troisième variable Z est défini par :

$$\mathrm{Corr}(X, Y|Z) = \mathrm{Corr}(E, F),$$

où E et F sont les résidus des régressions de X et Y sur Z :

$$X = a_1 + b_1 Z + E, \qquad Y = a_2 + b_2 Z + F.$$



Autocorrélation partielle

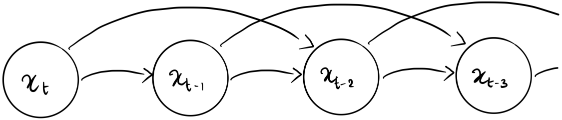
- **Définition**. La fonction d'autocorrélation partielle (**PACF**) d'une série stationnaire x_t, notée ϕ_{hh}, est définie par :

$$\begin{aligned} \phi_{11} &= \mathrm{Corr}(x_t, x_{t-1}) = \rho(1) \\ \phi_{hh} &= \mathrm{Corr}(x_t, x_{t-h} | x_{t-1}, \ldots, x_{t-h+1}), \quad h \geq 2. \end{aligned}$$

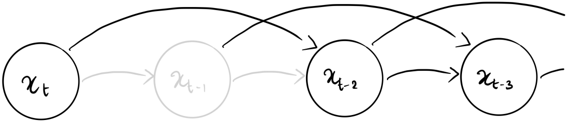
- *Idée*. On retire la dépendance linéaire des termes intermédiaires de la série.

Autocorrélation partielle : processus AR

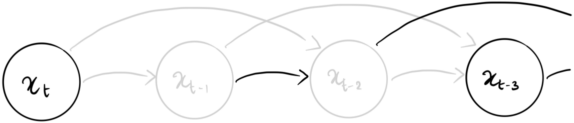
- On considère un processus AR(2) : $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$.



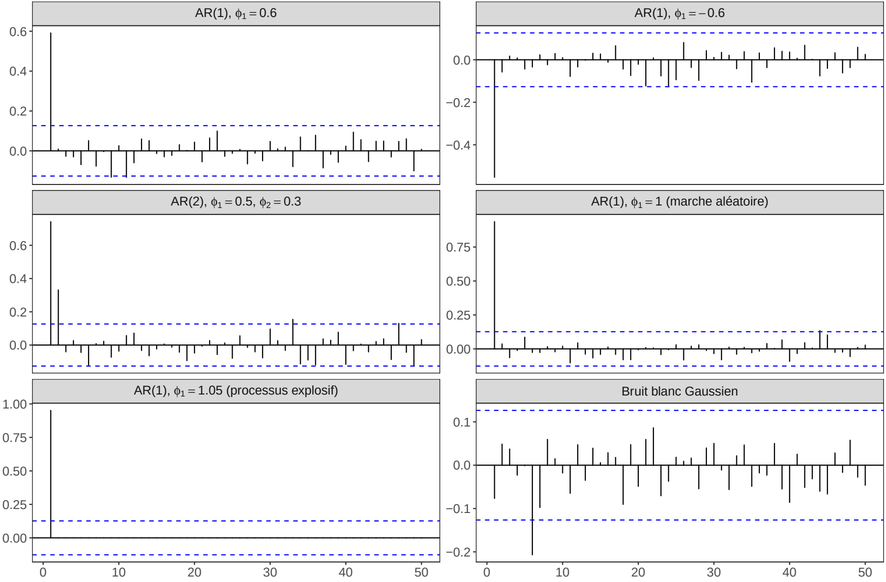
- PACF d'ordre 2 : $\phi_{22} = \mathrm{Corr}(x_t, x_{t-2} | x_{t-1})$.



- PACF d'ordre 3 : $\phi_{33} = \mathrm{Corr}(x_t, x_{t-3} | x_{t-1}, x_{t-2})$.

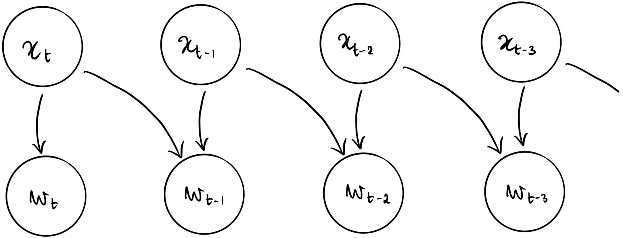


PACF de processus AR

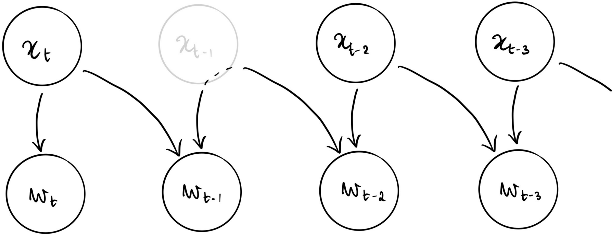


Autocorrélation partielle : processus MA

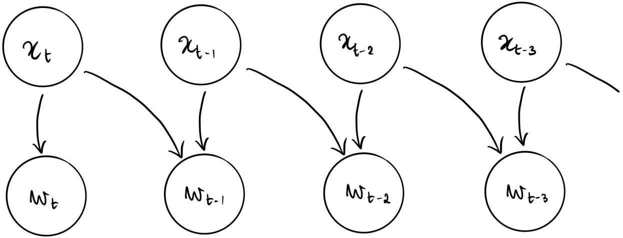
- On considère un processus MA(1) : $x_t = w_t + \theta_1 w_{t-1}$.



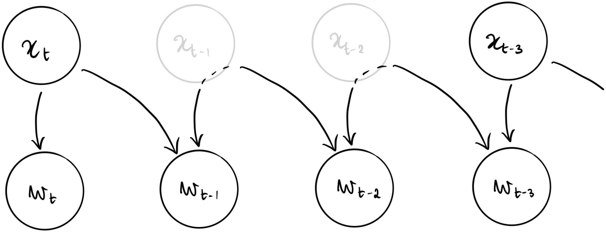
- PACF d'ordre 2 : $\phi_{22} = \mathrm{Corr}(x_t, x_{t-2} | x_{t-1})$.



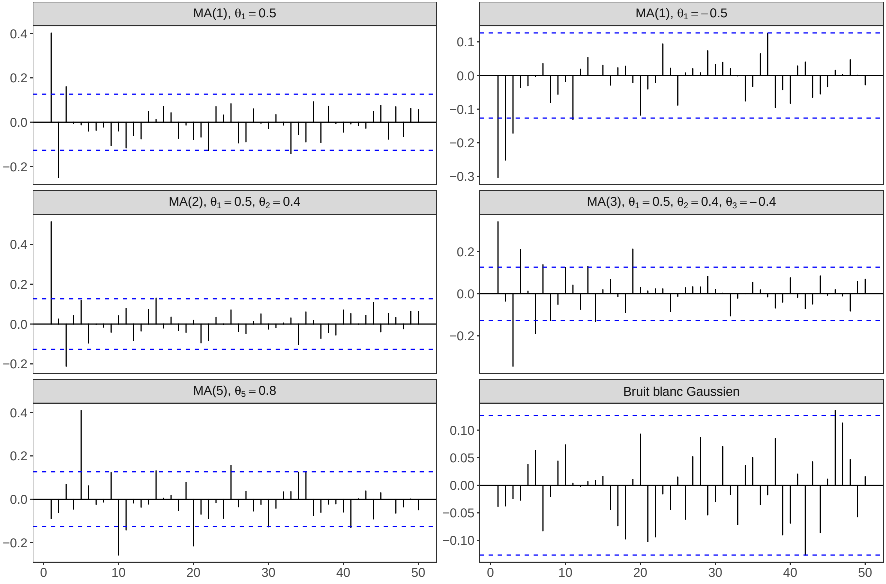
- On considère un processus MA(1) : $x_t = w_t + \theta_1 w_{t-1}$.



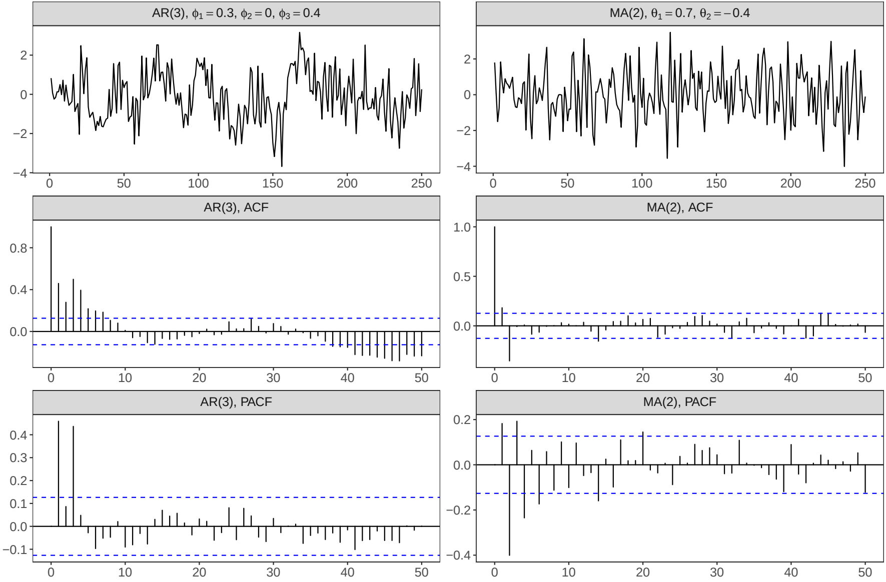
- PACF d'ordre 3 : $\phi_{33} = \mathrm{Corr}(x_t, x_{t-3} | x_{t-1}, x_{t-2})$.



PACF de processus MA



Modélisation : identification des ordres du processus



Modèle autorégressif moyenne mobile

- **Définition**. On appelle processus autorégressif moyenne mobile d'ordre (p, q), noté ARMA(p, q), le processus stationnaire x_t vérifiant :

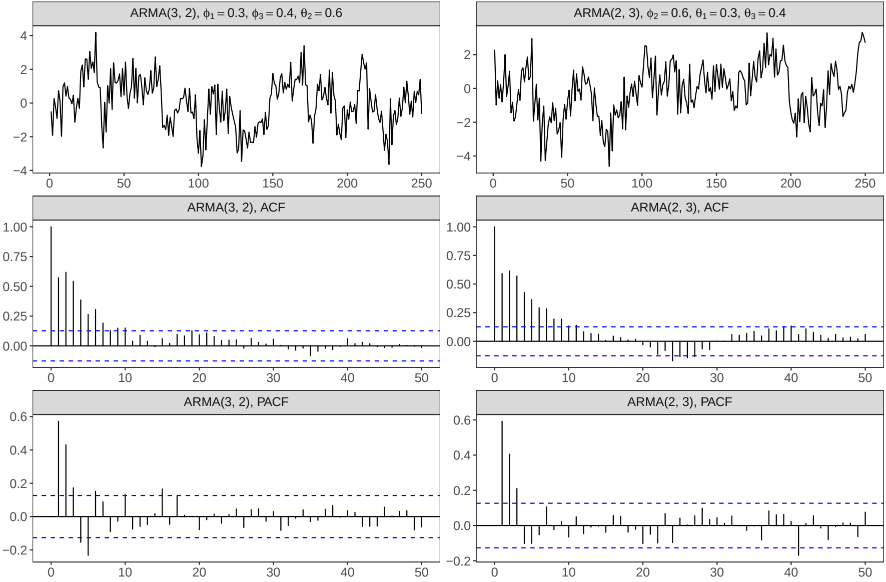
$$x_t = w_t + \sum_{i=1}^{p} \phi_i x_{t-i} + \sum_{j=1}^{q} \theta_j w_{t-j},$$

  où $\phi_p \neq 0$, $\theta_q \neq 0$ et $w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2)$.
- *Remarque*. En utilisant les polynômes AR Φ et MA Θ, on a :

$$\Phi(B)x_t = \Theta(B)w_t.$$

- **Théorème de représentation de Wold**. Toute série stationnaire au second ordre peut être représentée par un processus ARMA(p, q).

Identification des ordres du processus ?



Redondance paramétrique

Un sophisme

- Soit $x_t = w_t$ un bruit blanc Gaussien.
- On a donc $0.5\,x_{t-1} = 0.5\,w_{t-1}$.
- Puis $x_t - 0.5\,x_{t-1} = w_t - 0.5\,w_{t-1}$.
- Donc x_t est un processus ARMA$(1,\,1)$.

Redondance paramétrique. La surparamétrisation cache le fait que le processus est un bruit blanc Gaussien :

$$(1 - 0.5B)x_t = (1 - 0.5B)w_t.$$

Critères pénalisés d'information : critères de parcimonie

Pour choisir entre deux modèles, on utilise l'**AIC** ou le **BIC** (plus parcimonieux).

- Basés sur la fonction de vraisemblance ;
- Pénalisent le nombre de paramètres du modèle ;
- Compromis entre la qualité de l'ajustement et la complexité du modèle.

$$\begin{aligned} \mathrm{AIC} &= \quad 2\,\#\{\text{paramètres}\} - 2\log(\text{vraisemblance}), \\ \mathrm{BIC} &= \log(n)\,\#\{\text{paramètres}\} - 2\log(\text{vraisemblance}). \end{aligned}$$

Modèle autorégressif intégré moyenne mobile

- **Définition**. On appelle processus autorégressif intégré moyenne mobile d'ordre (p, d, q), noté ARIMA(p, d, q), un processus x_t tel que $\nabla^d x_t = (1 - B)^d x_t$ est un modèle ARMA(p, q) :

$$\Phi(B)\nabla^d x_t = \Theta(B) w_t.$$

- Aucune nouveauté ici : il s'agit principalement d'un jeu d'écriture.
- *Remarque*. L'écriture ARIMA est surtout utile pour R pour effectuer des prédictions.

Estimation des modèles ARMA

- On procède par maximum de vraisemblance : on cherche à maximiser $\mathcal{L}(\mu, \Phi, \Theta, \sigma_w^2; x_{1:n}) = f_{\mu,\Phi,\Theta,\sigma_w^2}(x_1, \ldots, x_n)$.
- Cette fonction de densité jointe se décompose en conditionnant par rapport aux points du passé :

$$f_{\mu,\Phi,\Theta,\sigma_w^2}(x_1, \ldots, x_n) = f(x_1)f(x_2 \mid x_1) \cdots f(x_n \mid x_{1:(n-1)}).$$

- Chaque fonction intermédiaire $f(x_j \mid x_{1:j})$ est la fonction de densité d'une loi normale univariée.

- **Propriété**. Sous de "bonnes conditions", pour un processus ARMA **causal** et **inversible**, le maximum de vraisemblance donne des estimateurs optimaux de σ_w^2 et $\beta = (\mu, \Phi, \Theta)$, et

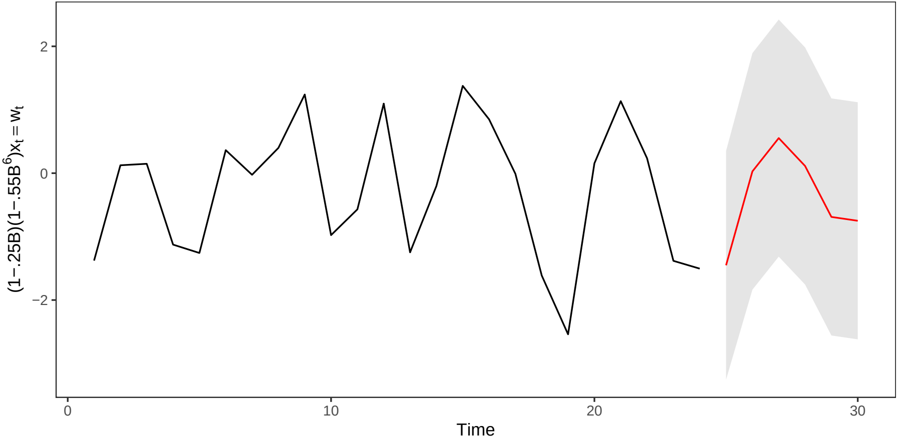
$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\big(0, \mathcal{I}^{-1}(\beta)\big).$$

- *Remarque*. Dans le cas d'un modèle surparamétrisé, la variance des estimateurs augmente : $\mathrm{Var}(\widehat{\beta}) \nearrow$.

Prédiction

- On cherche à prédire x_{n+m} à partir des observations $(x_1, \ldots, x_n)$.
- On appelle **prédicteur au pas de temps $\mathbf{m}$** la variable aléatoire définie par :

$$x_{n+m}^n = \mathbb{E}[x_{n+m} \mid x_{1:n}].$$



Prédiction (cont'd)

- *Exemple*. Pour un AR(p) :

$$\text{À } n+1 : x_{n+1}^n = \phi_1 x_n + \phi_2 x_{n-1} + \ldots + \phi_p x_{n-p+1},$$
$$\text{À } n+2 : x_{n+2}^n = \phi_1 x_{n+1}^n + \phi_2 x_n + \ldots + \phi_p x_{n-p+2}.$$

- *Exemple*. Pour un ARMA(p, q), il faut d'abord l'écrire sous forme de MA(∞) :

$$x_{n+m} = \sum_{j=0}^{\infty} \psi_j w_{n+m-j}.$$

Alors, on a

$$x_{n+m}^n = \sum_{j=m}^{m+n} \psi_j w_{n+m-j}.$$

Outline

7 Modèles saisonniers SARIMA
- ARIMA saisonnier pur
- Modèles SARIMA

ARIMA saisonnier pur

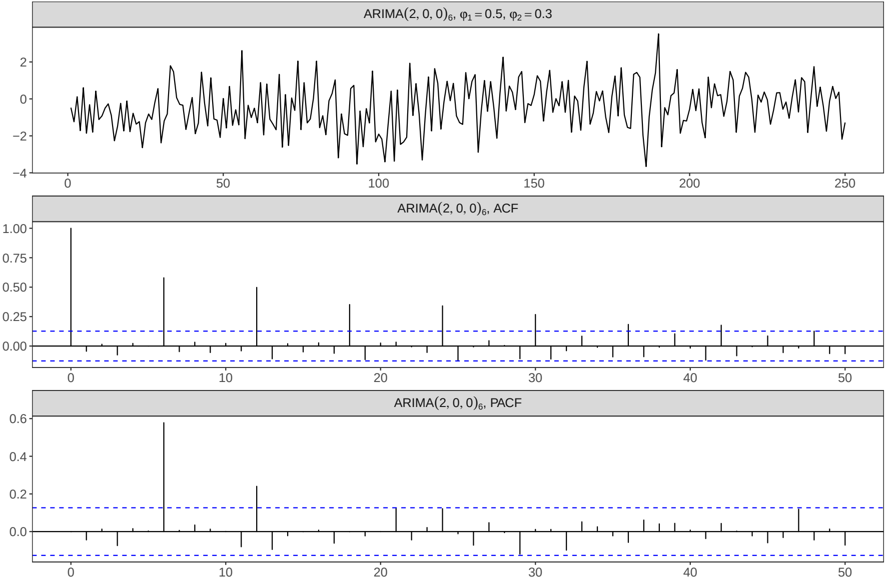
- *Idée*. On souhaite modéliser une structure de dépendance saisonnière.
- **Principe**. On va appliquer un modèle ARIMA, dont les ordres sont des multiples de la saisonnalité S.
- On note $\nabla_S := 1 - B^S$ l'opérateur de différence d'ordre S.
- **Définition**. On appelle processus ARIMA(P, D, Q)$_S$ saisonnier le processus stationnaire x_t vérifiant :

$$\Phi_P(B^S)\nabla_S^D x_t = \Theta_Q(B^S)w_t,$$

où

$$\Phi_P(B^S) = 1 - \varphi_1 B^S - \ldots - \varphi_P B^{PS},$$
$$\Theta_Q(B^S) = 1 + \vartheta_1 B^S + \ldots + \vartheta_Q B^{QS}.$$

Trajectoire et (P)ACF d'ARIMA saisonnier



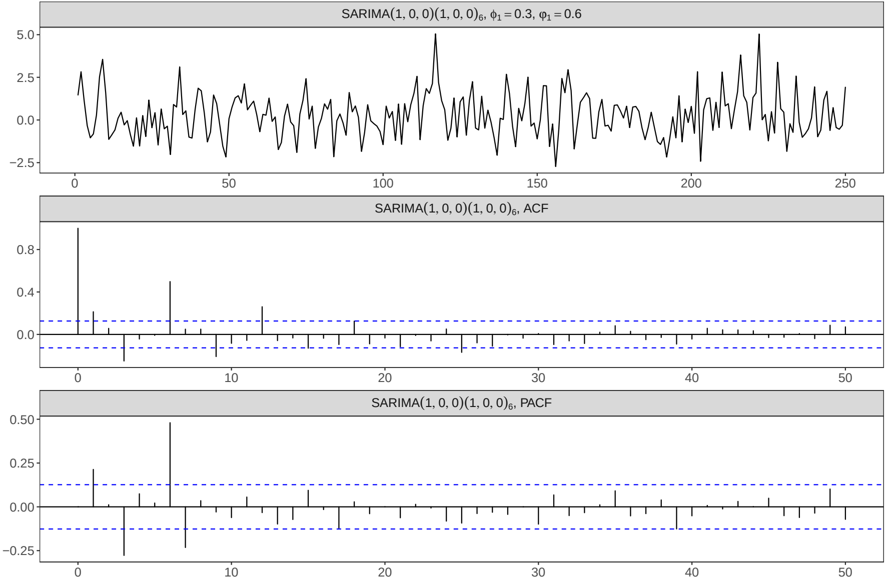
Modèles SARIMA

- *En pratique*, il existe généralement deux structures de dépendances :
    - Une à court terme, généralement ≤ 3 pas de temps ;
    - Une à long terme, due à la saisonnalité S de la série.
- **Idée**. On va combiner deux modèles ARIMA, le premier pour expliquer la dépendance à court terme, le deuxième pour la dépendance à long terme.
- **Définition**. On appelle processus SARIMA(p, d, q)$\times$(P, D, Q)$_S$ le processus stationnaire x_t vérifiant :

$$\Phi_P(B^S)\Phi(B)\nabla_S^D\nabla^d x_t = \Theta_Q(B^S)\Theta(B)w_t,$$

  où Φ et Θ sont les polynômes AR et MA de court terme,
  et Φ_P et Θ_Q sont ceux modélisant la saisonnalité.

Trajectoire et (P)ACF d'ARIMA saisonnier



Outline

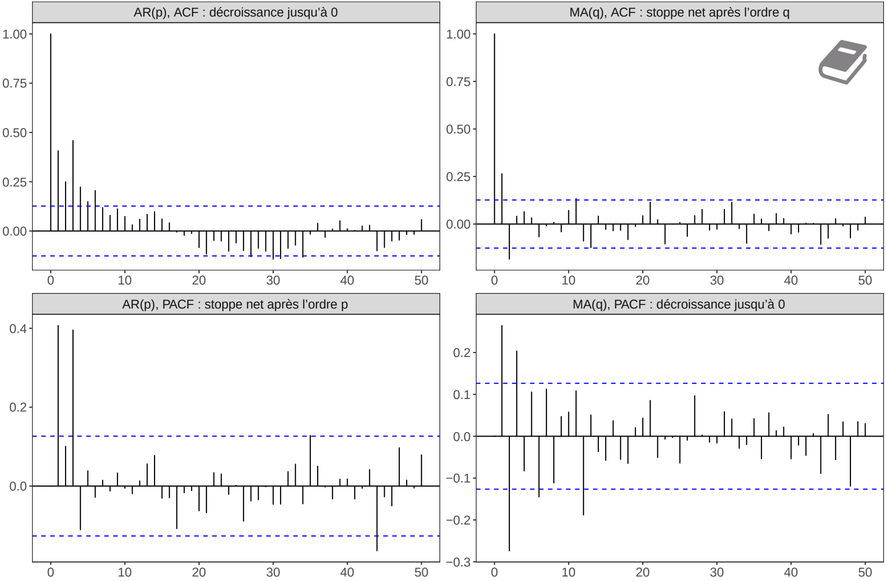
Procédure de Box et Jenkins

0a. Représenter les données.

0b. Transformer les données **si nécessaire**.
- Régression linéaire ou différentiation ;
- Stabilisation de la variance par logarithme.

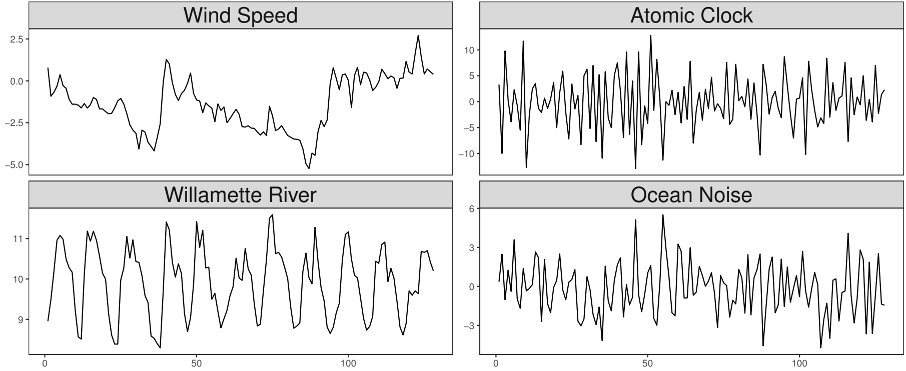
1. **Identification** les ordres p et q du modèle ARMA.
   - ACF (`acf`) et PACF (`pacf`) empiriques.
   - Identifier les structures saisonnières avant celles de court terme.
2. **Estimation** des paramètres par maximum de vraisemblance.
   - ```
     arima(series, order=c(p,d,q),
             seasonal=list(order=c(P,D,Q), period=S)).
     ```
3. **Diagnostic** des résidus.
   - ACF des résidus : existe-t-il des pics de covariance non pris en compte ?
   - Test de Ljung-Box : l'ensemble des pics de covariance est-il conforme avec un bruit blanc ?
   - Graphiques et tests usuels : la normalité des résidus est-elle vérifiée ?
4. Choix du modèle.
   - Utilisation des critères d'information parcimonieux.
     ```

Identification des ordres du processus



Introduction to spectral analysis

We consider four examples of time series $x_1, x_2, \ldots, x_{128}$. How would you describe them?

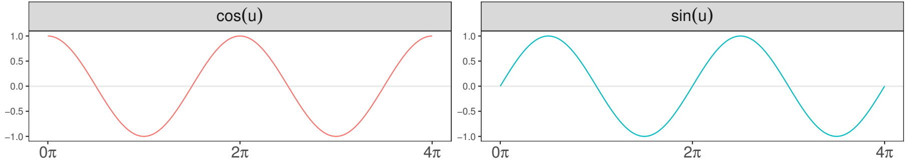


Spectral analysis describes x_t's by comparing them to sines and cosines.

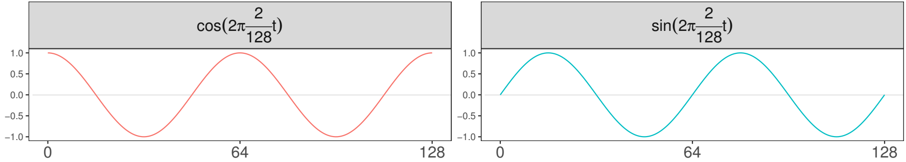
Sines and cosines?

The functions $\sin$ and $\cos$ are 2π-periodic: for $u \in \mathbb{R}$ and $k \in \mathbb{Z}$,

$$\cos(u + 2k\pi) = \cos(u)$$



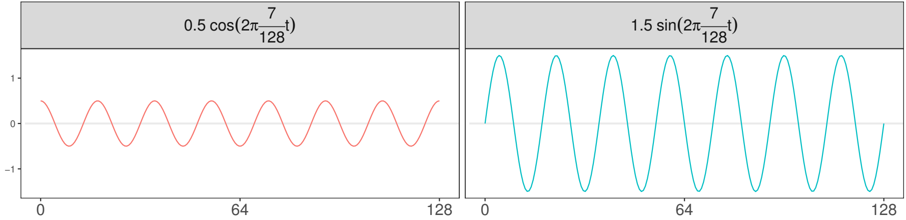
Let $u = 2\pi \frac{2}{128} t$ for $t = 1, 2, \ldots, 128$.



$\frac{2}{128}$ can be interpreted as 2 cycles over the time span of 128.

Sines and cosines?

Similarly, with $u = 2\pi \frac{7}{128} t$ for $t = 1, 2, \ldots, 128$.



- $\cos(2\pi \frac{k}{n} t)$ and $\sin(2\pi \frac{k}{n} t)$ have k cycles per n time steps.
- The quantity $f = \frac{k}{n}$ is called the *frequency* of the sine or cosine.
  - It is the amount (or rather fraction) of cycles per time step.
  - If f is small (large), the sine is said to have low (high) frequency.
- The *period* $T = \frac{1}{f}$ is the time steps needed for a full cycle.
- The *amplitude* is the maximum range of variation and is equal to 1 for the functions $\sin$ and $\cos$.

Sines and cosines?

Summing up sines and cosines of different amplitudes and frequencies create time series that resemble actual data.

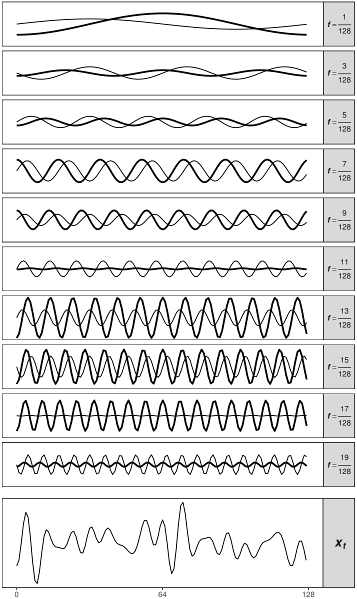
Goal of spectral analysis

Given a time series x_t, figure out its *Fourier representation*, *i.e.* its decomposition into sines and cosines:

$$x_t = \sum_k a_k \cos\left(2\pi \frac{k}{n} t\right) + b_k \sin\left(2\pi \frac{k}{n} t\right)$$

Actually easy to compute a_k and b_k:

$$a_k \propto \mathrm{Cov}\left\{x_t, \cos\left(2\pi \frac{k}{n} t\right)\right\}$$



The spectrum and periodogram

$\sigma_k^2 = a_k^2 + b_k^2$, the squared amplitude of the sine-cosine pair, highlights the importance of the frequency $\frac{k}{n}$ in the decomposition of x_t.

- $(\sigma_k^2)_{k=1,\ldots,n}$ is called the *spectrum* of the time series x_t.
- If σ_k^2 is large, there are strong patterns of frequency $\frac{k}{n}$.
- The sample spectrum, noted I_k, is called the *periodogram* of x_t.

Decomposition of variance

Since the sines and cosines of different frequencies are uncorrelated, then

$$\mathrm{Var}\ x_t = \sum_k (a_k^2 + b_k^2) = \sum_k \sigma_k^2$$

The spectrum (σ_k^2) is the decomposition of the variance of the time series x_t into its different frequencies $\frac{k}{n}$.

Examples

Recall the four examples: here are their periodograms.