

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

Enseignement

- $2 \times 3h$ de cours théoriques (ici, faculté de médecine du KB).
- $2 \times 3h$ de TPs pratiques sous R (ici, KB-TP2 ? ou hôpital Paul-Brousse ?).
 - TPs dédoublés en demi-groupes ?
- Supports de cours, annales et (prochainement) TPs disponibles sur <https://fcheysson.github.io/serc.html>.

Evaluation

- Quiz / questions de cours au premier TP : ~ 10 minutes, ~ 3 points.
- Devoir sur table au dernier TP : ~ 30 minutes, ~ 7 points.
- Devoir maison : ~ 1 mois, ~ 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 :

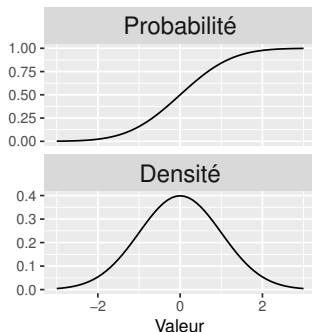
$$F_X(t) = \mathbb{P}(X \leq t);$$

- Fonction de densité f_X :

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

- Fonction quantile :

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



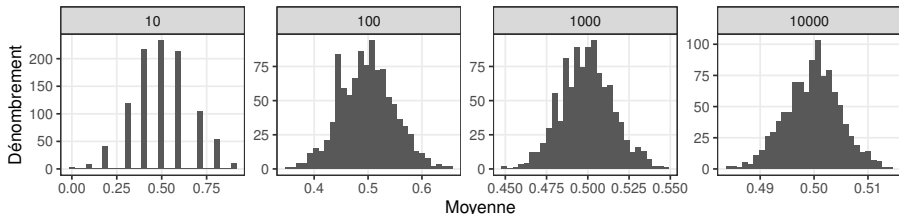
Rappels : loi normale $\mathcal{N}(\mu, \sigma^2)$

- 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 .
- Apparaît dans le **théorème central limite** : Soit une suite de variables aléatoires X_1, X_2, \dots indépendantes et identiquement distribuées (i.i.d.), d'espérance μ et de variance $\sigma^2 < \infty$. Alors

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



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 Kyndt², Mehdi Djennadi²

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

* mejdoubi-m@ch-valenciennes.fr



Abstract

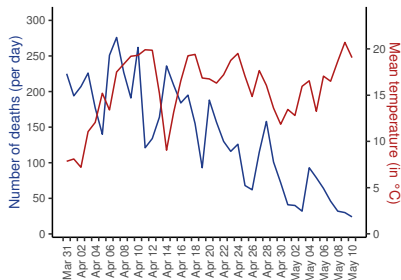
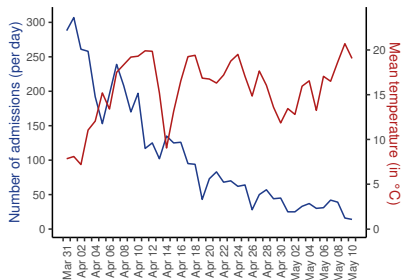
The purpose of this ecological study was to explore the association of weather with severity indicators of coronavirus disease 2019 (COVID-19). Daily COVID-19-related intensive care unit (ICU) admissions and in-hospital deaths in the Paris region and the daily weather characteristics of Paris midtown 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.

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 [3, 4], with an immunological mechanism known as “cytokine



OPEN ACCESS

Citation: Mejdoubi M, Kyndt X, Djennadi 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): e0242088. <https://doi.org/10.1371/journal.pone.0242088>

Editor: Christian den Uyl, Erasmus Medical Center, NETHERLANDS

Received: August 3, 2020

Accepted: October 22, 2020

Published: November 20, 2020

Copyright: © 2020 Mejdoubi et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data used in this study are public and freely accessible. Ambient temperature data are available from <https://www.météofrance.com> and daily numbers of COVID-19 deaths are available from <https://covid19.gouv.fr>.

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

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

Rappels : tests d'hypothèse

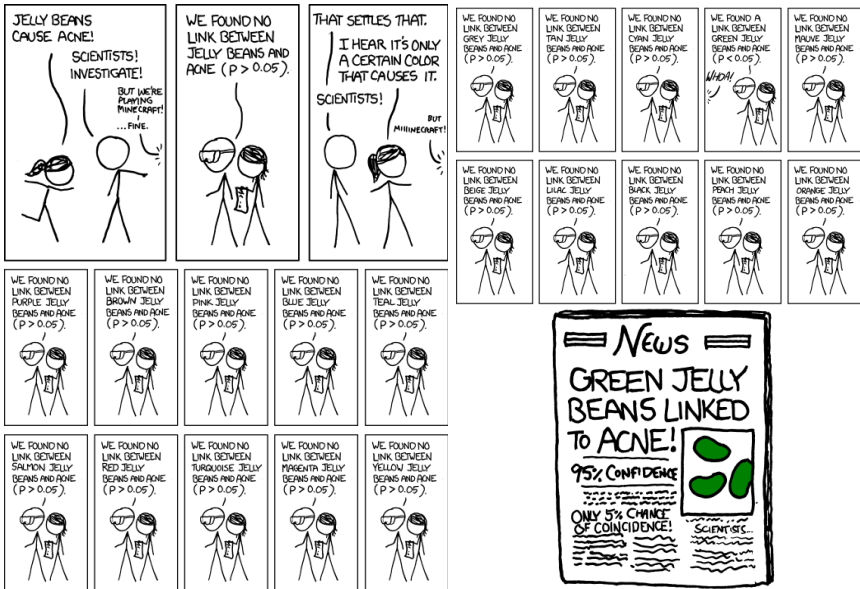
Procédure de décision permettant de rejeter une hypothèse sur la base de l'étude statistique d'un échantillon (X_1, \dots, X_n) .

- 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

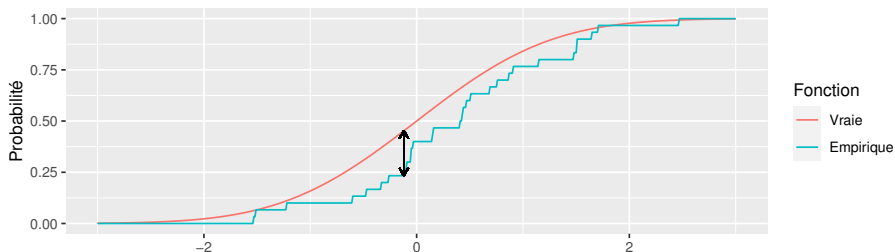
Permet de tester si l'échantillon provient d'une distribution F déterminée.

- 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 :

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

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 .



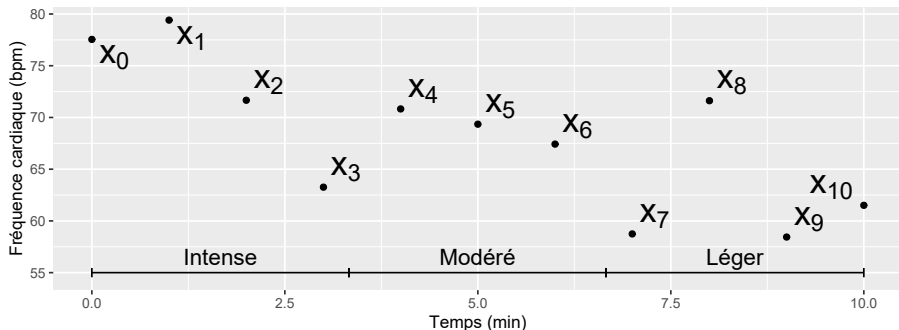
Définition : série temporelle

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

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

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

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



- On cherche à expliquer la fréquence cardiaque de l'individu en fonction de l'effort réalisé.
- **Naïvement.** On effectue une ANOVA en fonction de l'effort :

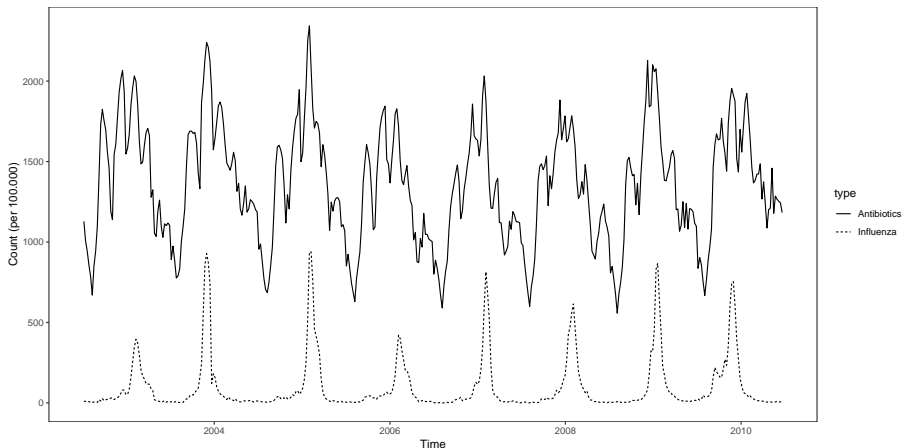
$$x_t = \mu + \alpha_{\text{effort}} + \varepsilon_t,$$

où $\varepsilon_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

Quel est le problème d'une telle modélisation ?

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.



Qualité de l'indicateur

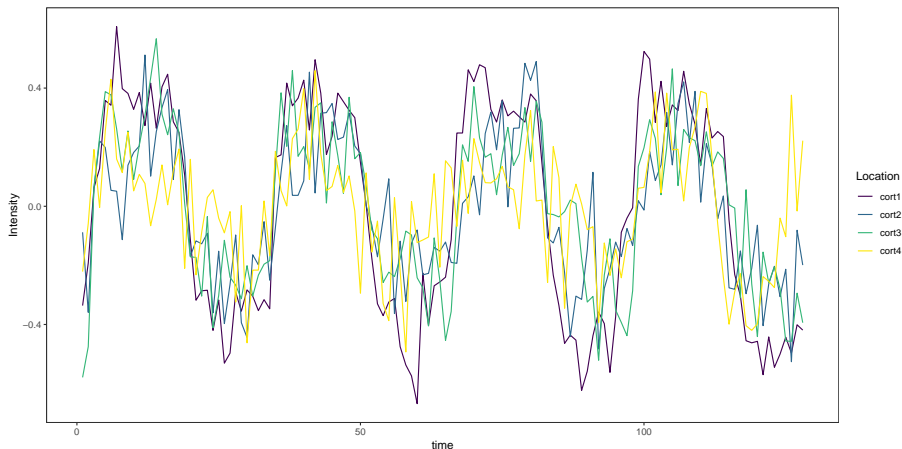
- **Représentatif ?**
 - Surveillance passive : sous-estimation.
 - Biais de recrutement.
- **Stabilité temporelle ?**
 - Evolution de l'indicateur dans le temps.

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

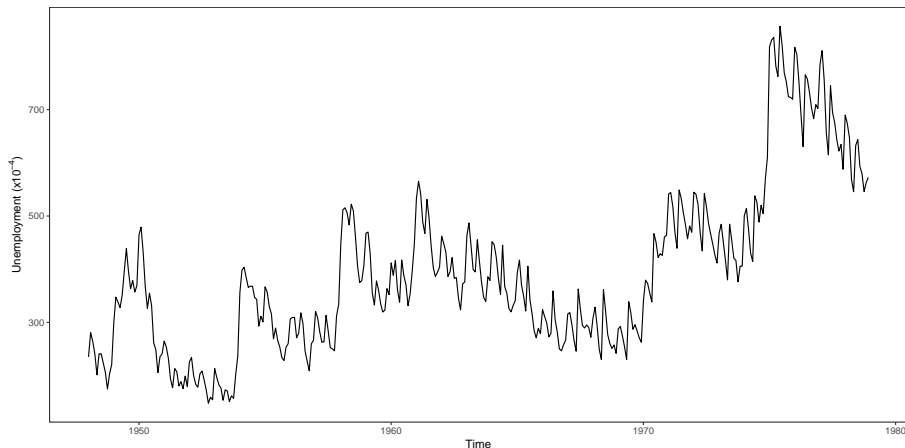
Intensité du signal BOLD (Blood Oxygenation-Level Dependent) dans le cortex d'un patient soumis à des stimuli, collectée par fMRI.



Exemple de données économiques

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

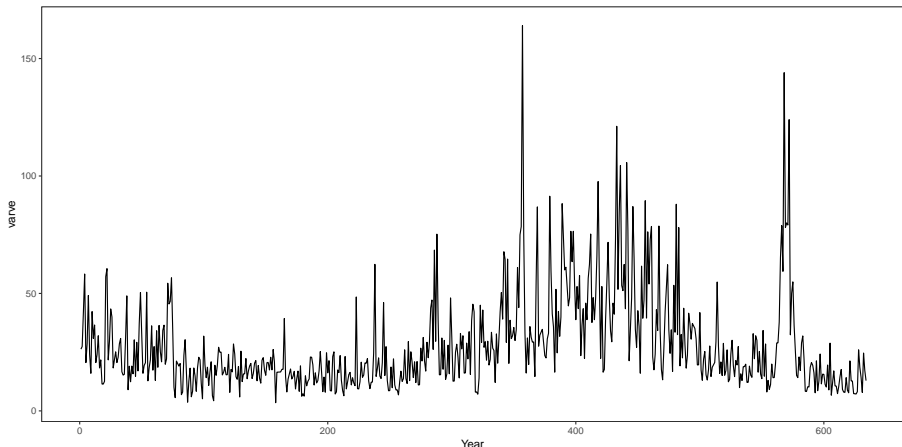
- Impact du choc pétrolier de 1973.



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 :

$$\begin{aligned}x_t &= T_t + S_t + \nu_t \\ y_t &= T'_t + S'_t + \beta x_t + v_t,\end{aligned}$$

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 v_t désignent des variables **résiduelles autodépendantes**, *i.e.*

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

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

Problématique

Prendre en compte la structure d'autodépendance des innovations.

Applications

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

Théoriques

- 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.

- 1 Exemples
- 2 Notions théoriques essentielles
- 3 Stationnarité
- 4 Estimateurs de la corrélation
- 5 Données non stationnaires
- 6 Modèles ARIMA
- 7 Modèles saisonniers SARIMA
- 8 Procédure de Box et Jenkins

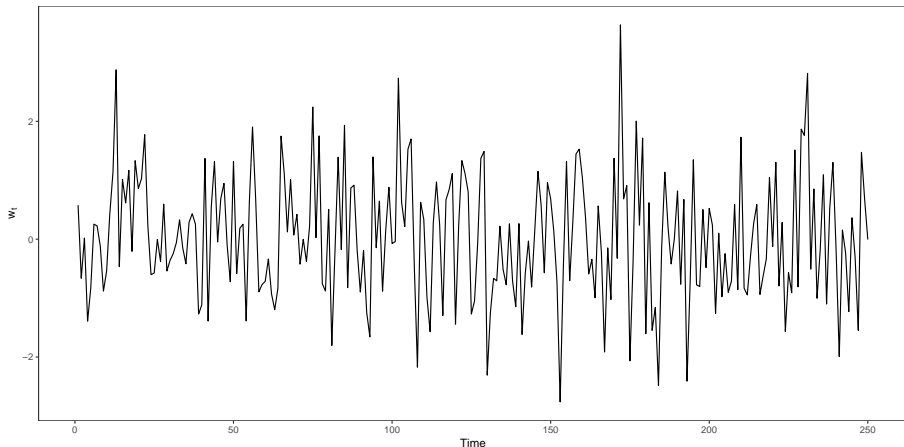
1 Exemples

- 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* :

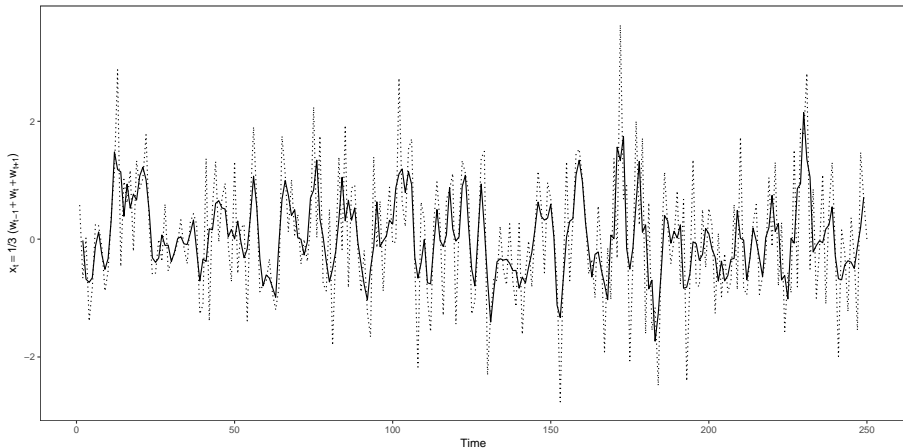
$$w_t \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_w^2)$$



Exemple 1.2 : Moyenne mobile

Combinaison linéaire d'un bruit blanc Gaussien :

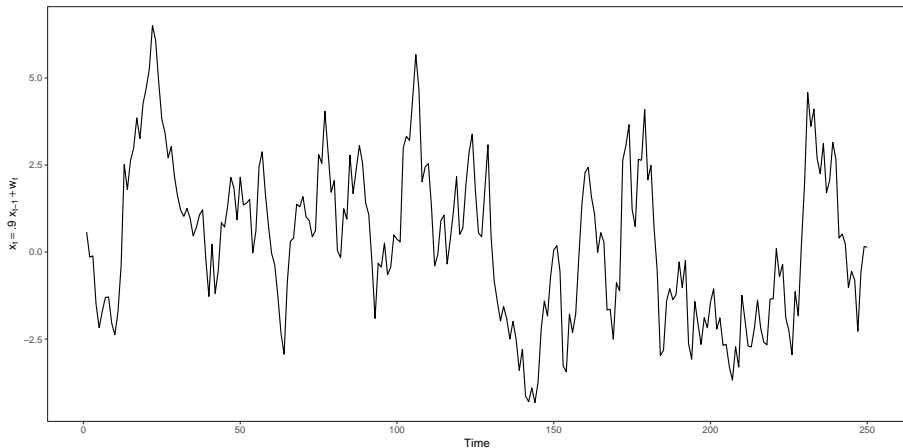
$$x_t = \frac{1}{3} (w_{t-1} + w_t + w_{t+1})$$



Exemple 1.3 : Processus autorégressif

Régression sur les valeurs passées de la série :

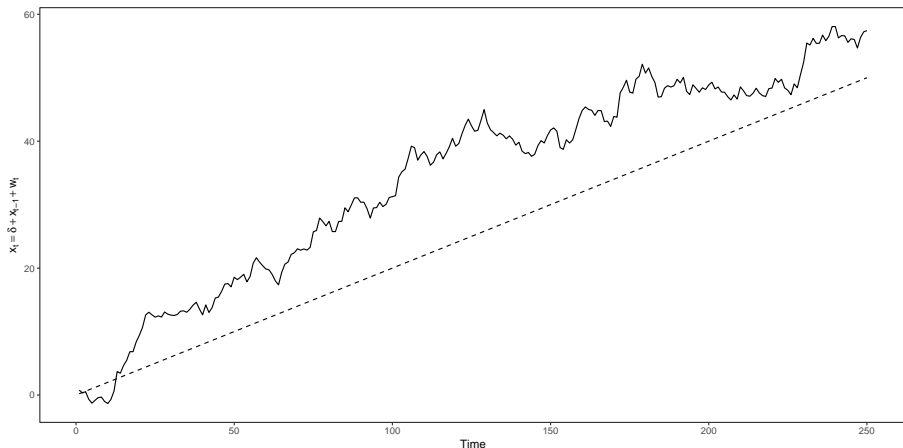
$$x_t = 0.9 x_{t-1} + w_t$$



Exemple 1.4 : Marche aléatoire avec dérive

Succession de pas aléatoire avec dérive δ :

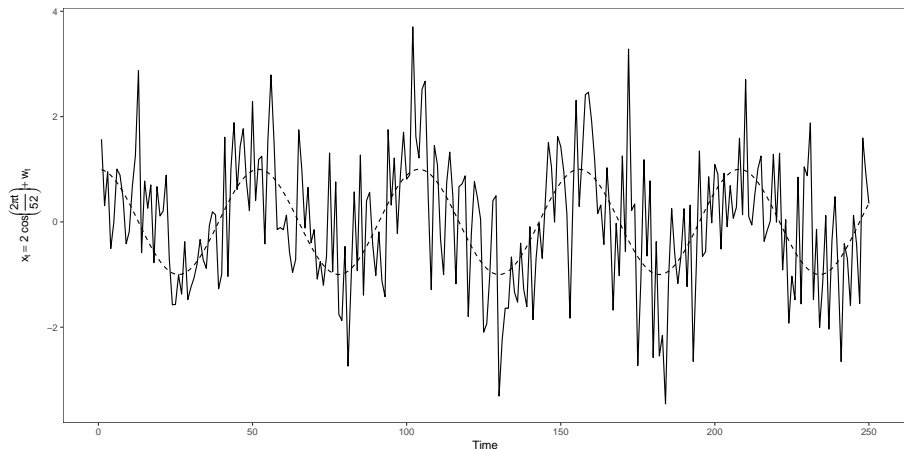
$$x_t = \delta + x_{t-1} + w_t$$



Exemple 1.5 : Signal bruité

Signal périodique bruité :

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



- 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)

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

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

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

$$\text{Cov}(A, B) = \text{Cov}(B, A)$$

$$\text{Cov}(A + B, C) = \text{Cov}(A, C) + \text{Cov}(B, C)$$

$$\text{Cov}(\alpha A, B) = \alpha \text{Cov}(A, B)$$

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

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

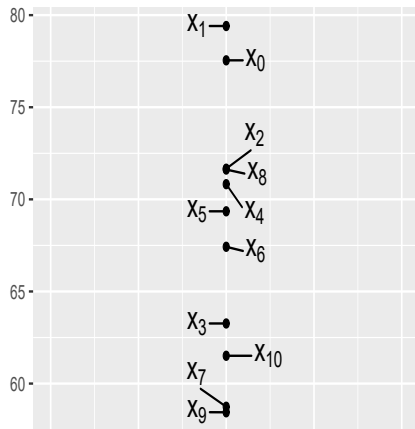
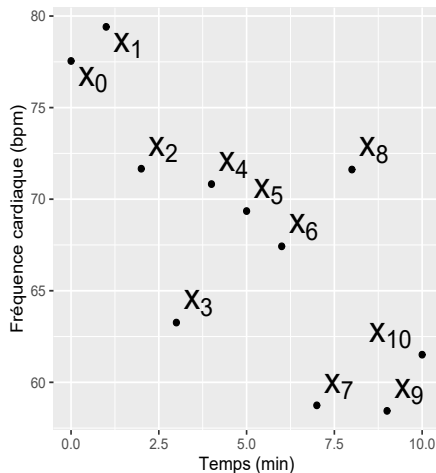
- *Remarque.* La fonction d'autocorrélation caractérise la prédictabilité linéaire de la série au temps t sachant la valeur de x_s .

3 Stationnarité

- Problématique
- Stationnarité au sens strict
- Stationnarité au sens faible

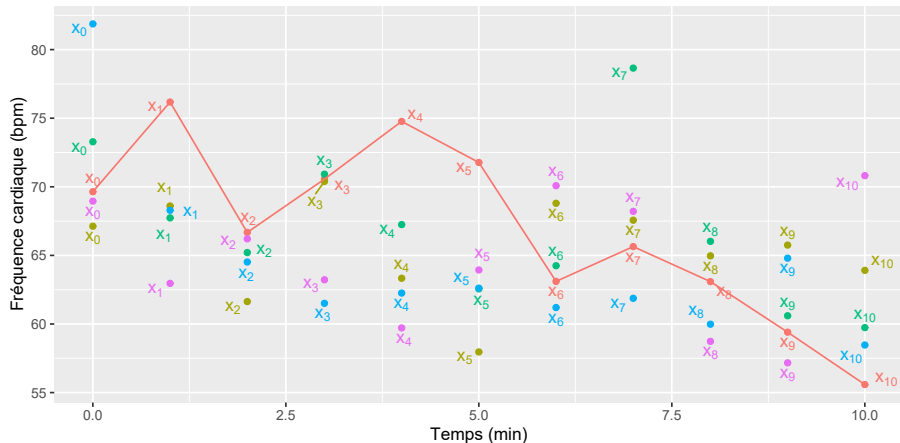
Stationnarité : problématique

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



Stationnarité : problématique (cont'd)

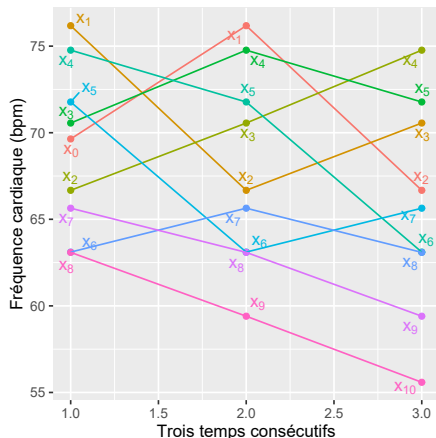
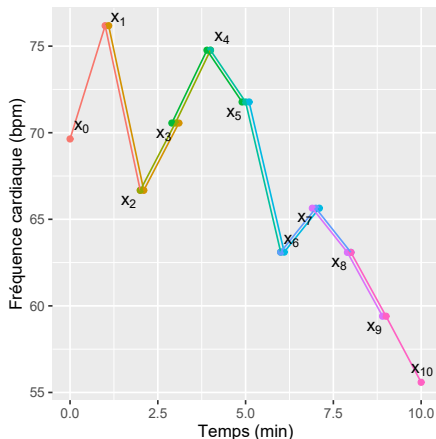
Pour la plupart des séries temporelles, les observations ne sont pas indépendantes.



Stationnarité au sens strict

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.

→ 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 2^{ème} 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 ;
- $\gamma(s, t) = \text{Cov}(x_s, x_t)$ ne dépend que de $|s - t|$.

- Stationnarité stricte \implies Stationnarité faible ;
- Stationnarité stricte \nleftarrow Stationnarité faible ;
- Stationnarité stricte $\xleftarrow{(x_t) \text{ Gaussien}}$ Stationnarité faible.

Stationnarité au sens faible (cont'd)

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

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

- **Variance.** $\text{Var}(x_t) = \gamma(t, t) = \gamma(0, 0) = \gamma(0)$;
- **Fonction d'autocorrélation.** $\rho(h) := \rho(t+h, t) = \dots = \gamma(h)/\gamma(0)$.
- *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.

4 Estimateurs de la corrélation

- Définitions
- Exemples

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

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

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

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

- **Propriété.** Pour un bruit blanc Gaussien (w_t), si n est “grand”, la loi de l'ACF empirique peut être approchée par :

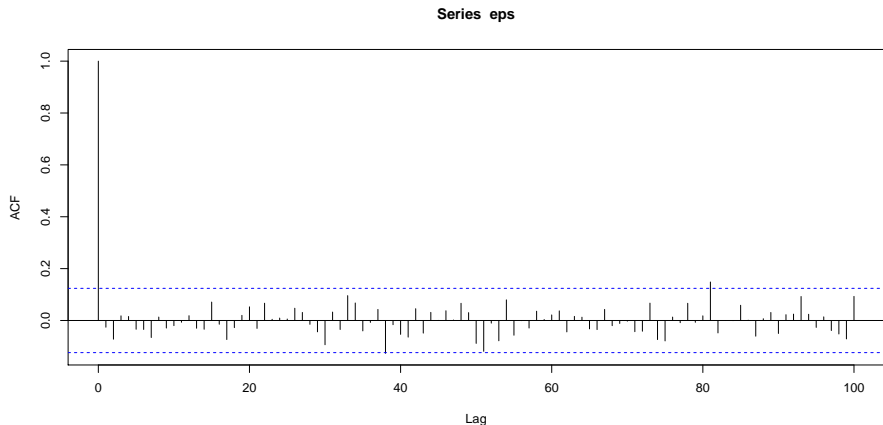
$$\hat{\rho}_w(h) \sim \mathcal{N}\left(0, \sigma_{\hat{\rho}_w(h)}^2\right),$$

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

ACF 4.1 : Bruit blanc Gaussien

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

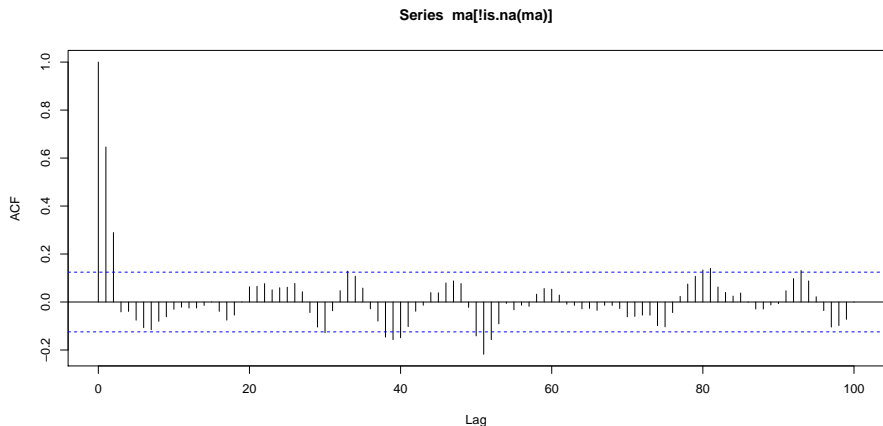
$$w_t \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_w^2)$$



ACF 4.2 : Moyenne mobile

Combinaison linéaire d'un bruit blanc Gaussien :

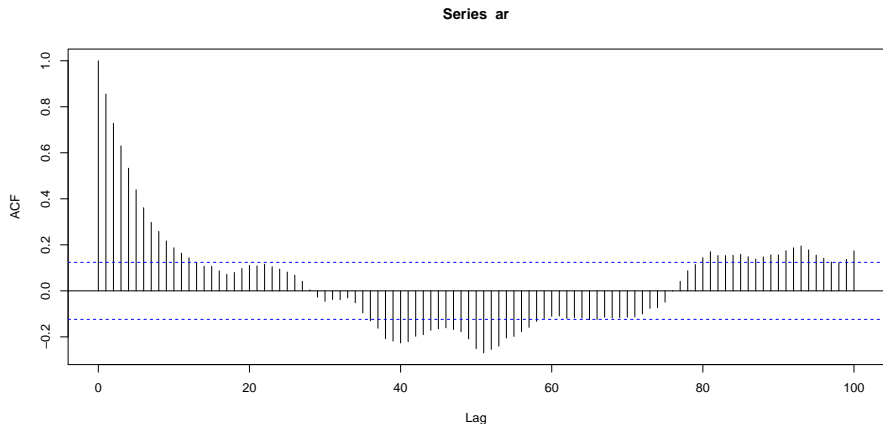
$$x_t = \frac{1}{3} (w_{t-1} + w_t + w_{t+1})$$



ACF 4.3 : Processus autorégressif

Régression sur les valeurs passées de la série :

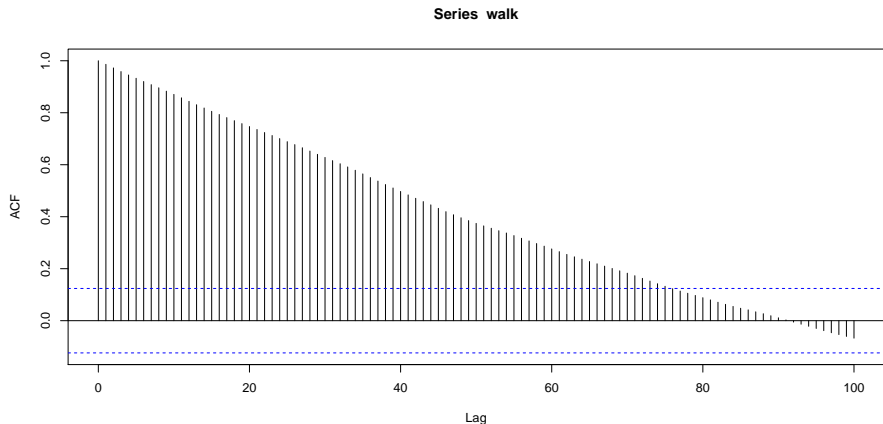
$$x_t = 0.9 x_{t-1} + w_t$$



ACF 4.4 : Marche aléatoire avec dérive

Succession de pas aléatoire avec dérive δ :

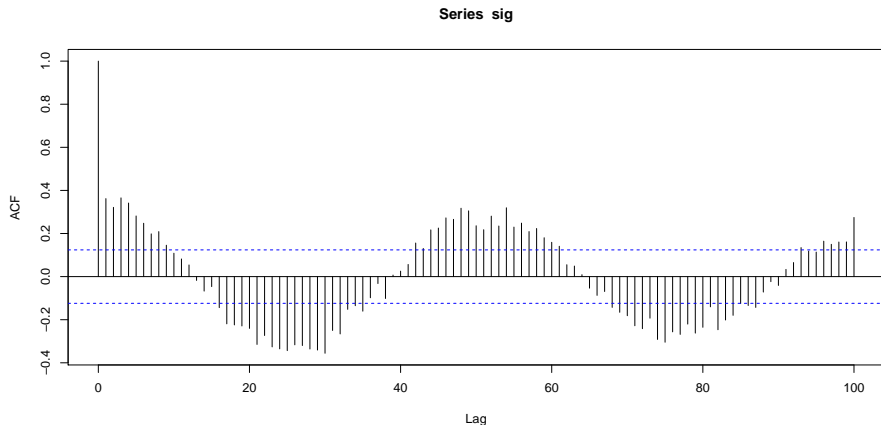
$$x_t = \delta + x_{t-1} + w_t$$



ACF 4.5 : Signal bruité

Signal périodique bruité :

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



5 Données non stationnaires

- Prise en compte d'une tendance par régression
- Prise en compte d'une tendance par différenciation
- Stabilisation de la variance

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 considère 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 $\hat{\nu}_t = x_t - \hat{\mu}_t$;
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 :

$$B x_t = x_{t-1},$$

$$B^2 x_t = B(Bx_t) = B(x_{t-1}) = x_{t-2},$$

$$B^k x_t = x_{t-k}.$$

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

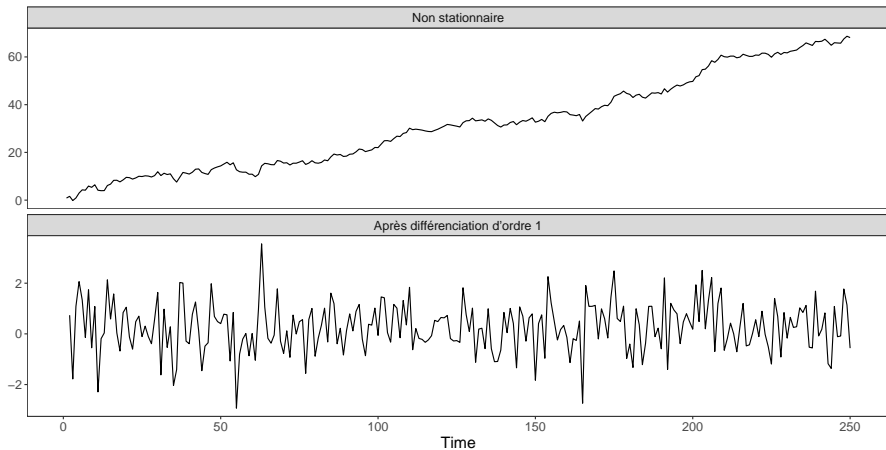
$$\nabla x_t = (1 - B) x_t = x_t - Bx_t = x_t - x_{t-1},$$

$$\nabla^2 x_t = (1 - B)^2 x_t = (1 - 2B + B^2)x_t = x_t - 2x_{t-1} + x_{t-2}.$$

- *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

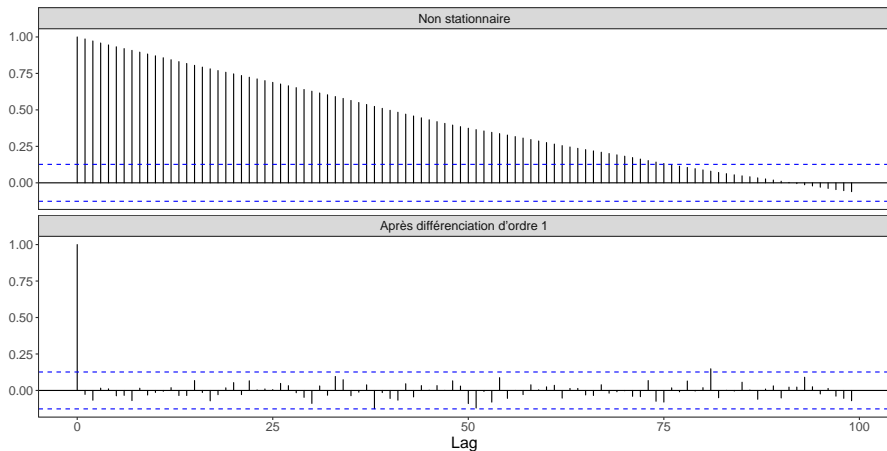
$$x_t = \delta + x_{t-1} + w_t$$



Stationnarité : Marche aléatoire avec dérive

Succession de pas aléatoire avec dérive δ :

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



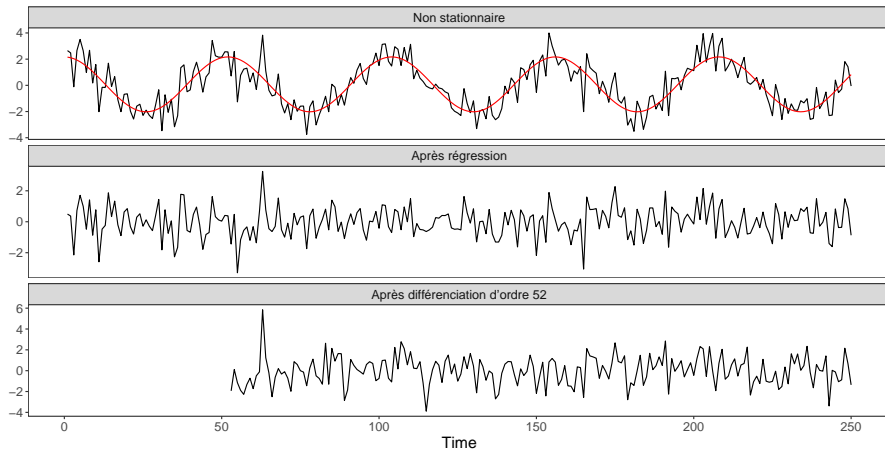
- Pour une série saisonnière, on pourra étudier une différenciation de plus grand ordre.
- On note $\nabla_S = 1 - B^S$ l'opérateur de différence d'ordre S :

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

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

Détermination ou régression du signal bruité

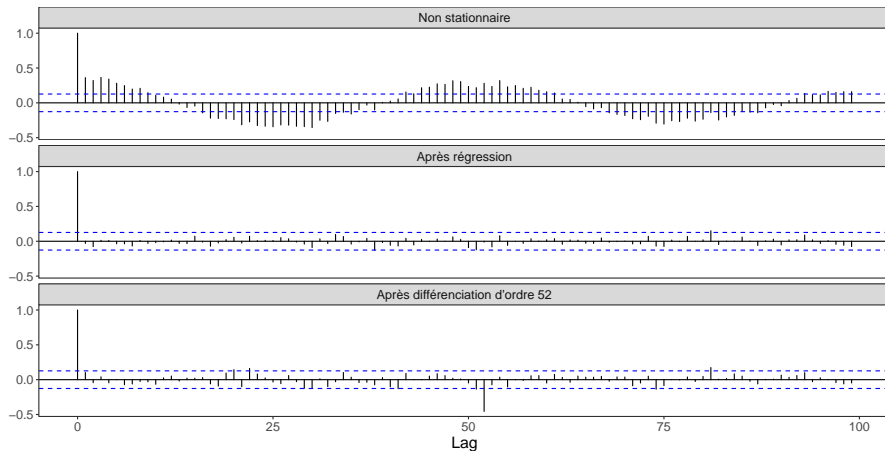
$$x_t = 2 \cos\left(\frac{2\pi t}{52}\right) + w_t$$



Stationnarité : Signal bruité

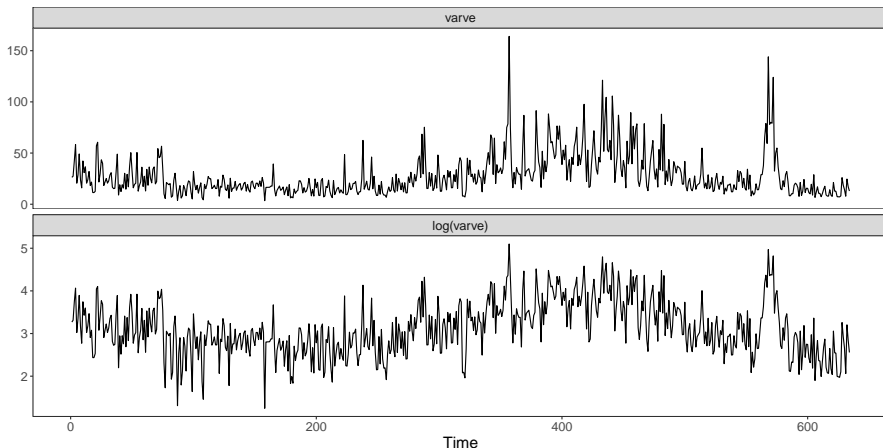
Signal périodique bruité :

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



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.



6 Modèles ARIMA

- Modèles MA
- Modèles AR
- Autocorrélation partielle
- Modèles ARMA
- Estimation
- Prédiction

Modèle moyenne mobile (*moving average*)

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

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

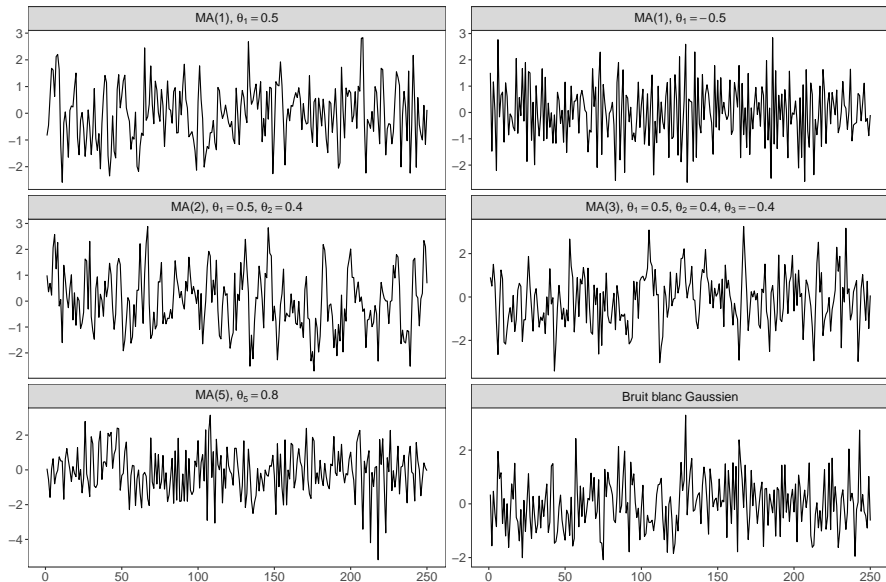
où $\theta_q \neq 0$ et $w_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2)$.

- *Remarque.*

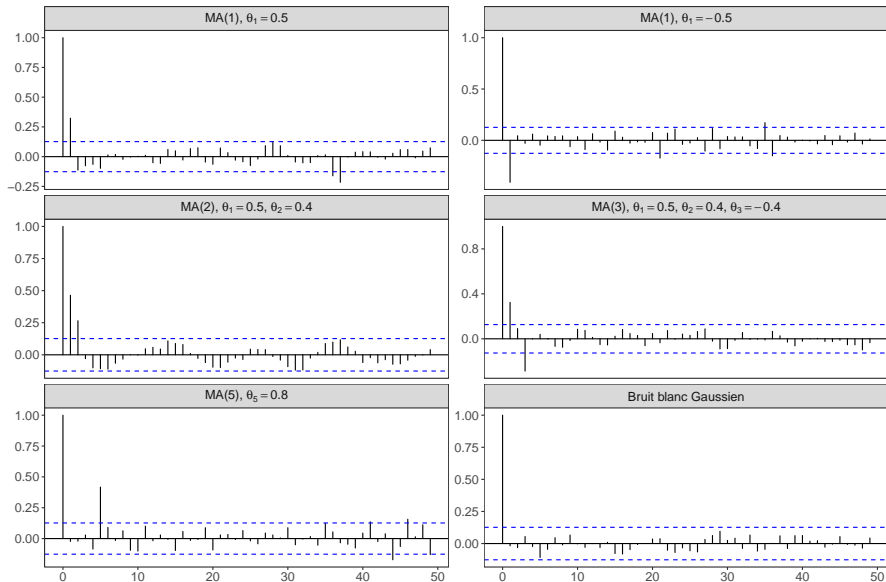
$$\begin{aligned}x_t &= (1 + \theta_1 B + \dots + \theta_q B^q) w_t \\ x_t &= \Theta(B) w_t\end{aligned}$$

- **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



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

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

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

- *Remarque.* Si $\mathbb{E}[x_t] = \mu$, on considérera $y_t = x_t - \mu$.
- *Remarque.*

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

- **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 $|\phi| < 1$ 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 :

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

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

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

- **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

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

- **Définition.** Lorsque, pour un $\text{MA}(q)$ 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.

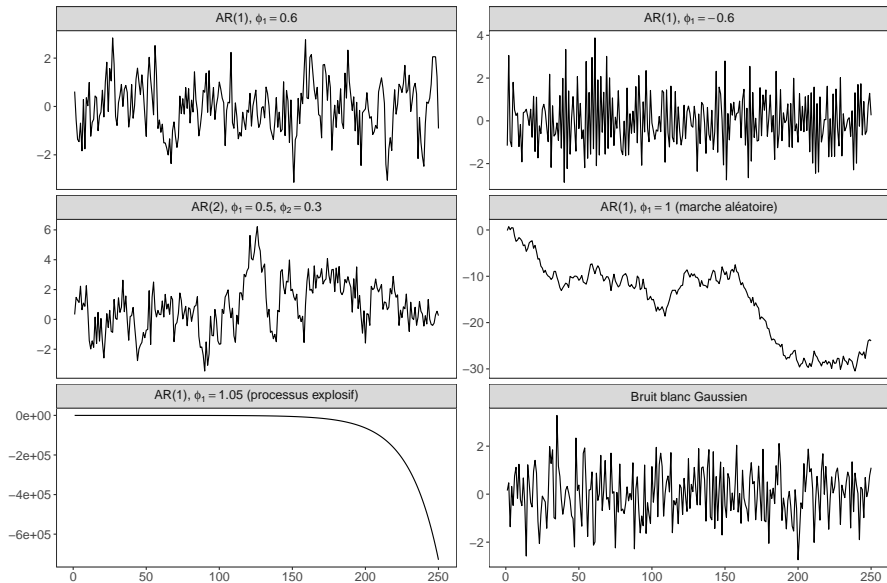
- *Remarque.* Dans ce cas, on a

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

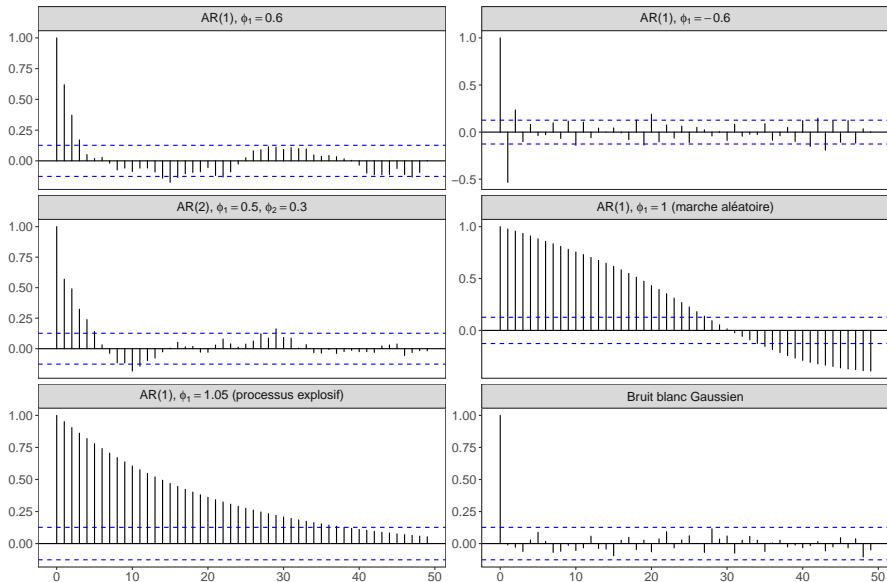
soit

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

Trajectoires de processus AR



ACF de processus AR



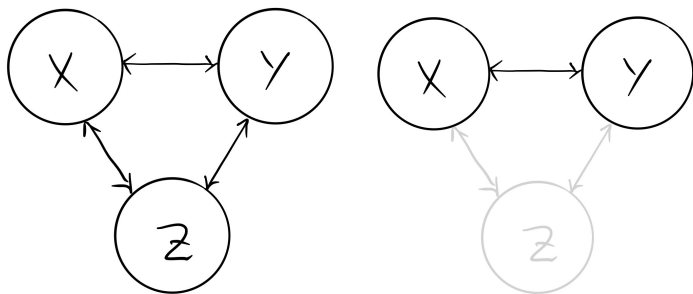
Corrélation partielle

- *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 :

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

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

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



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

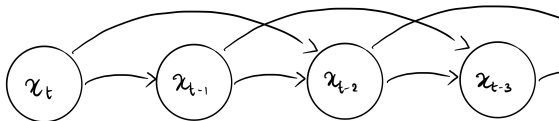
$$\phi_{11} = \text{Corr}(x_t, x_{t-1}) = \rho(1)$$

$$\phi_{hh} = \text{Corr}(x_t, x_{t-h} | x_{t-1}, \dots, x_{t-h+1}), \quad h \geq 2.$$

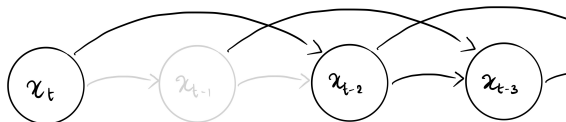
- *Idée.* On retire la dépendance linéaire des termes intermédiaires de la série.

Autocorrélation partielle : processus AR

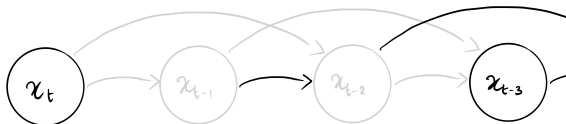
- On considère un processus AR(2) : $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$.



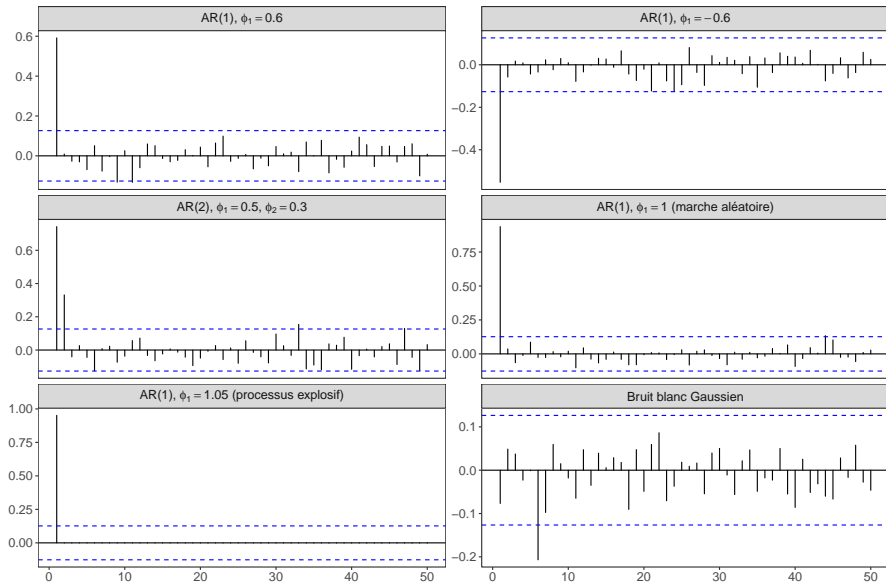
- PACF d'ordre 2 : $\phi_{22} = \text{Corr}(x_t, x_{t-2} | x_{t-1})$.



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

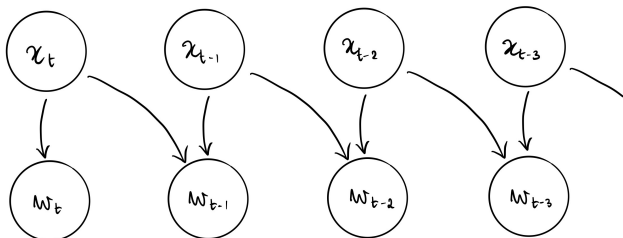


PACF de processus AR

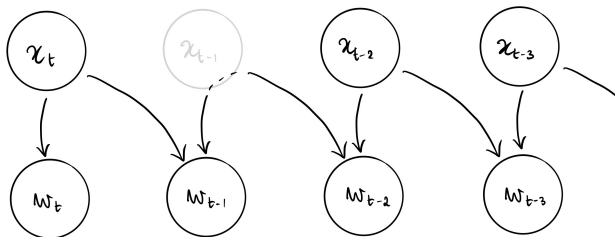


Autocorrélation partielle : processus MA

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

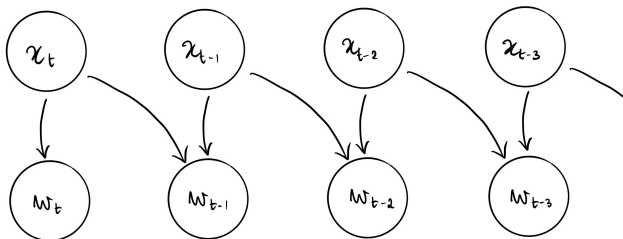


- PACF d'ordre 2 : $\phi_{22} = \text{Corr}(x_t, x_{t-2} | x_{t-1})$.

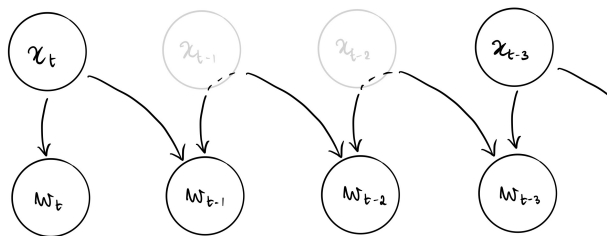


Autocorrélation partielle : processus MA (cont'd)

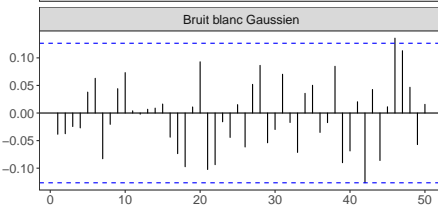
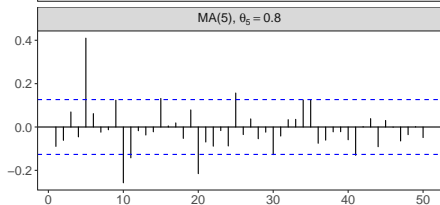
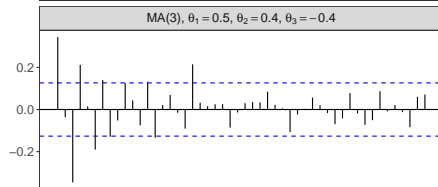
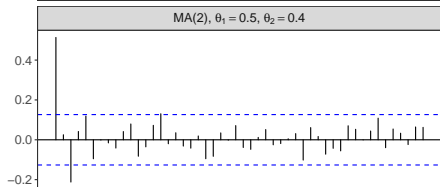
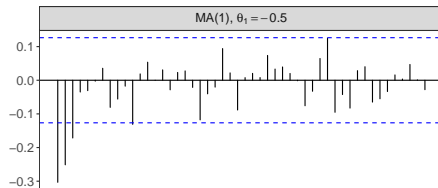
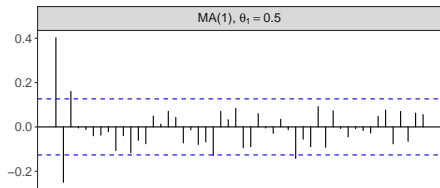
- On considère un processus MA(1) : $x_t = w_t + \theta_1 w_{t-1}$.



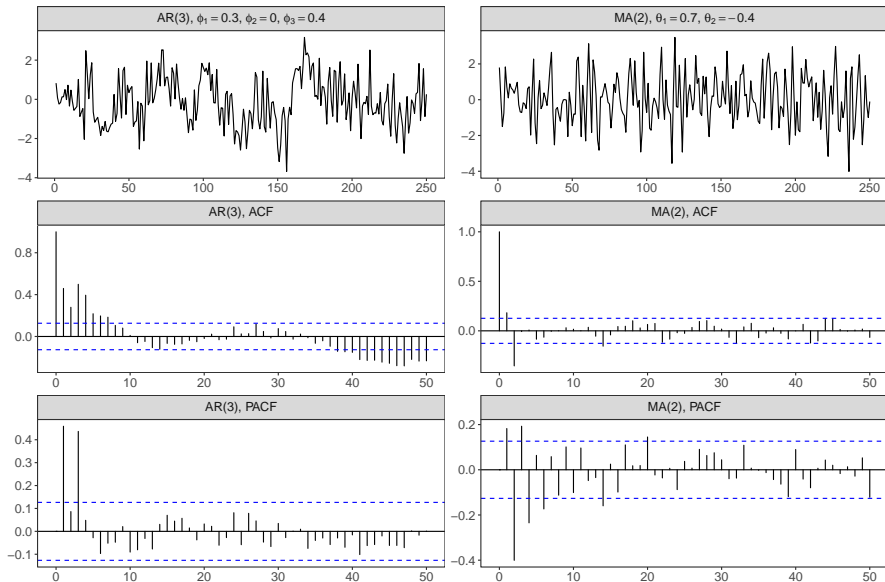
- PACF d'ordre 3 : $\phi_{33} = \text{Corr}(x_t, x_{t-3} | x_{t-1}, x_{t-2})$.



PACF de processus MA



Modélisation : identification des ordres du processus



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

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

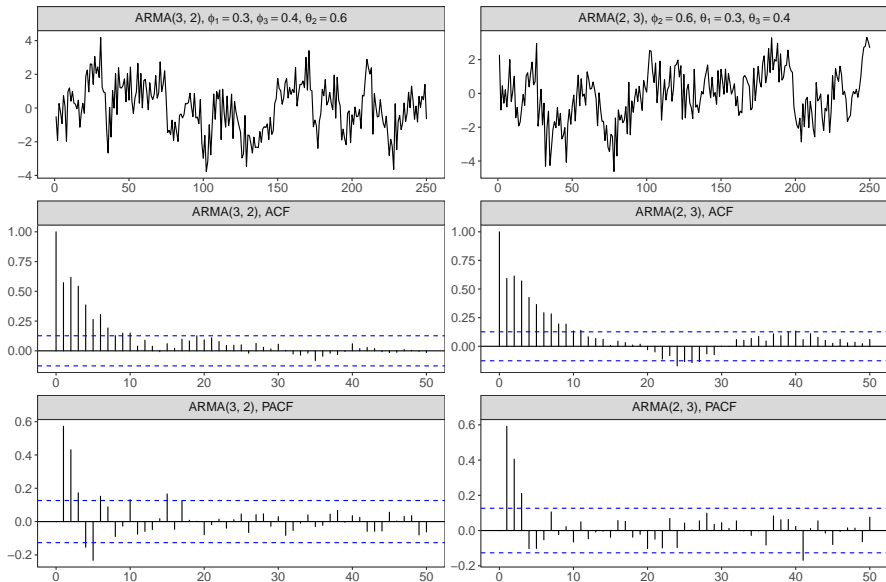
où $\phi_p \neq 0$, $\theta_q \neq 0$ et $w_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2)$.

- *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 $\text{ARMA}(p, q)$.

Identification des ordres du processus ?



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 :

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

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.

$$\text{AIC} = 2 \#\{\text{paramètres}\} - 2 \log(\text{vraisemblance}),$$

$$\text{BIC} = \log(n) \#\{\text{paramètres}\} - 2 \log(\text{vraisemblance}).$$

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

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

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

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

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

- Chaque fonction intermédiaire $f(x_j | x_{1:j})$ 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

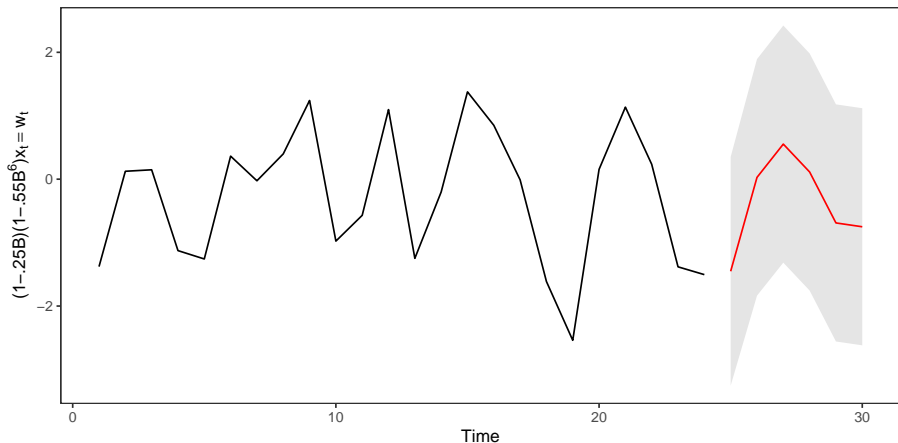
$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}(\beta)).$$

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

Prédiction

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

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



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

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

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

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

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

Alors, on a

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

7 Modèles saisonniers SARIMA

- ARIMA saisonnier pur
- Modèles SARIMA

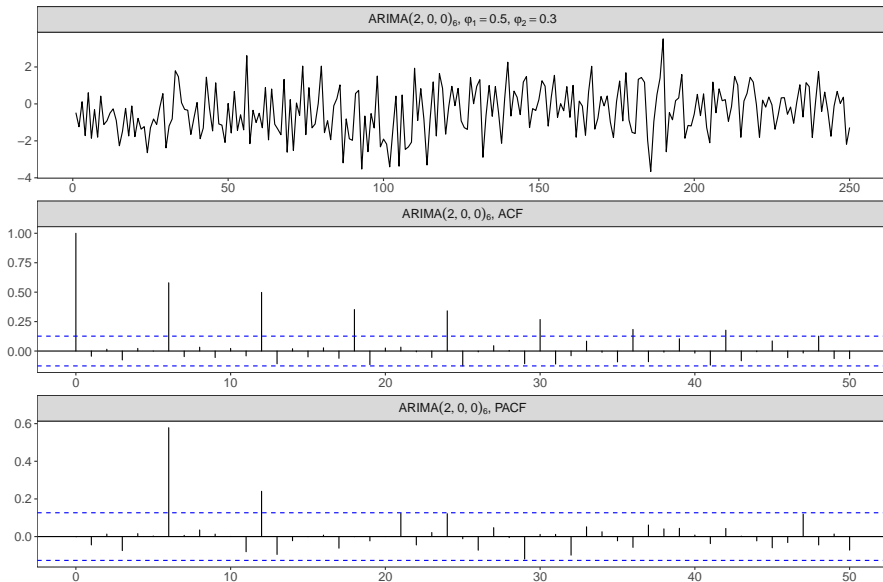
- *Idée.* On souhaite modéliser une structure de dépendance saisonnière.
- **Principe.** On va appliquer un modèle ARIMA, dont les ordres sont des multiples de la saisonnalité S .
- On note $\nabla_S := 1 - B^S$ l'opérateur de différence d'ordre S .
- **Définition.** On appelle processus ARIMA(P, D,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

$$\begin{aligned}\Phi_P(B^S) &= 1 - \varphi_1 B^S - \dots - \varphi_P B^{PS}, \\ \Theta_Q(B^S) &= 1 + \vartheta_1 B^S + \dots + \vartheta_Q B^{QS}.\end{aligned}$$

Trajectoire et (P)ACF d'ARIMA saisonnier

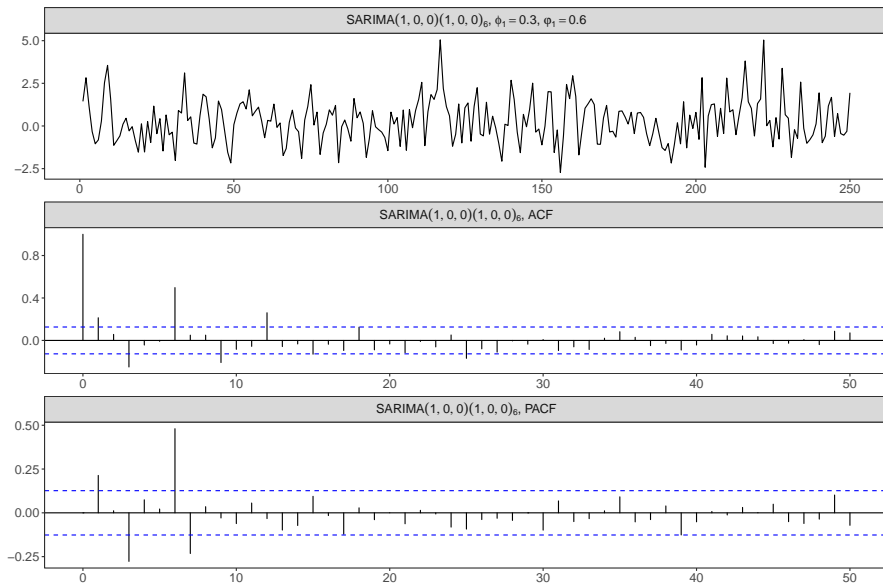


- *En pratique*, il existe généralement deux structures de dépendances :
 - Une à court terme, généralement ≤ 3 pas de temps ;
 - Une à long terme, due à la saisonnalité S de la série.
- **Idée.** On va combiner deux modèles ARIMA, le premier pour expliquer la dépendance à court terme, le deuxième pour 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



8 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.
 - `arma(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

