Température corporelle d'un castor (une petite introduction aux séries temporelles)

GMMA 106

Cas d'étude

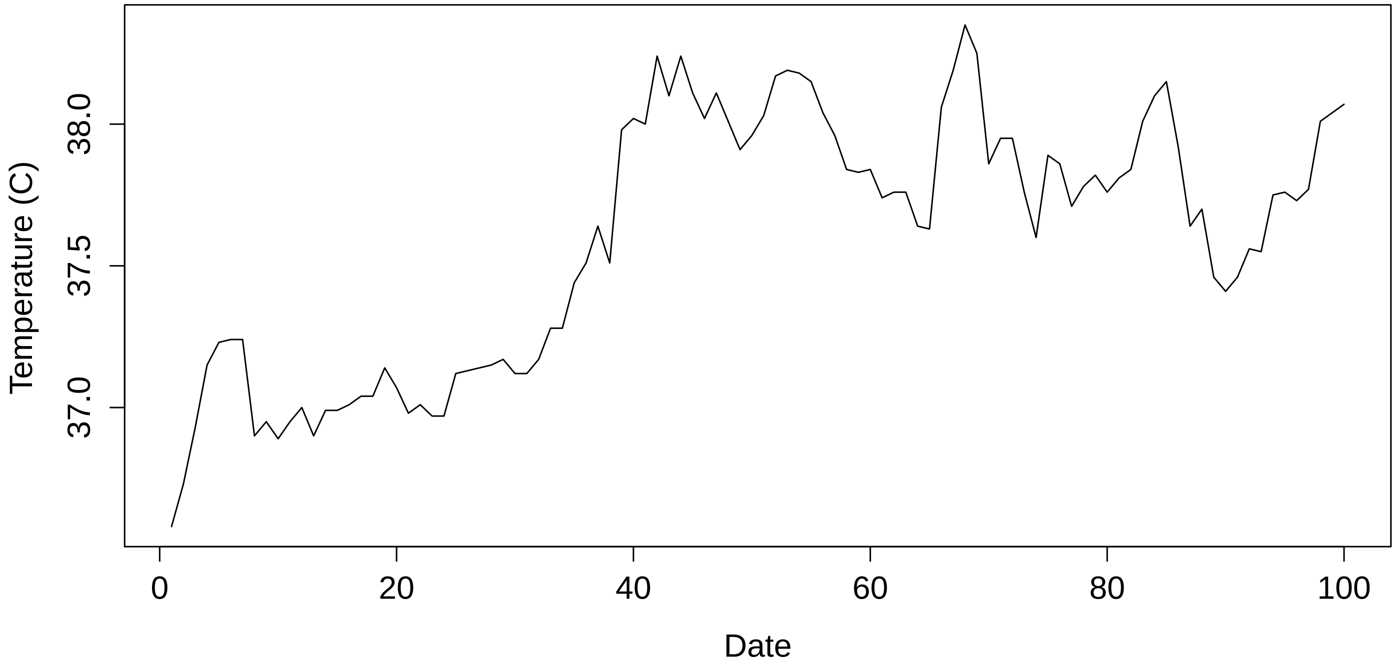

Figure 1: *Temperature corporelle d'un castor femelle (Castor canadensis) du Wisconsin. Une mesure toute les 10 minutes.*

- □ Ceci est une série temporelle.
- □ Quelles remarques/constats pouvez vous faire à la vue de la Figure 1 ?

Les données

```
> library(datasets)
> beaver2
    day time  temp activ
1   307  930 36.58     0
2   307  940 36.73     0
3   307  950 36.93     0
4   307 1000 37.15     0
.
.
37  307 1530 37.64     0
38  307 1540 37.51     0
39  307 1550 37.98     1
40  307 1600 38.02     1
.
.
98  308  140 38.01     1
99  308  150 38.04     1
100 308  200 38.07     1
```

day Jour de l'obervation

time Heure de l'observation 9h30 $\rightarrow$ 930

temp Température corporelle en $^\circ\mathrm{C}$

activ Variable binaire codant la présence dans son foyer (0) ou non (1).

Objectifs et éléments parcourus

- Notre objectif est de proposer un modèle statistique pour modéliser la température corporelle de notre femelle castor.
- Ceci nous permettra de croiser les objets suivants :
  - Processus stochastique
  - autocovariance, autocorrelation, autocorellation partielle
  - bruit blanc, processus auto regressifs (AR)
  - prédictions

Spécificités des séries temporelles

- La plupart des analyses statistiques supposent que les observations sont des réalisations indépendantes—parfois de même loi, i.e.,

$$Y_1, \ldots, Y_n \overset{\text{ind}}{\sim} F_1, \ldots, F_n, \qquad Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F.$$

- Les séries temporelles sont ordonnées dans le sens où les observations apparaissent successivement (notion de temps) introduisant ainsi de la dépendance
- Il existe une infinité de types de dépendance, tout comme il existe de nombreux types de séries temporelles. . .

Environnement

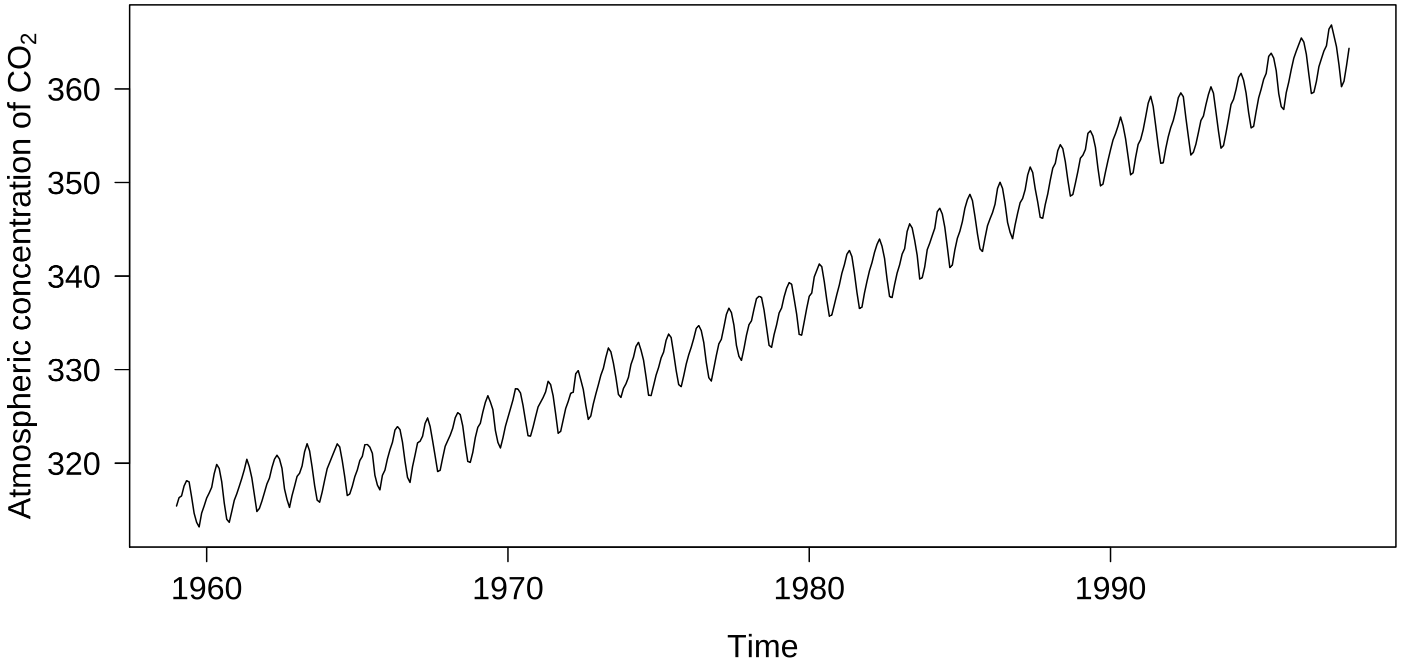

Figure 2: *Concentration de CO2 à Mauna Loa (Hawaï).*

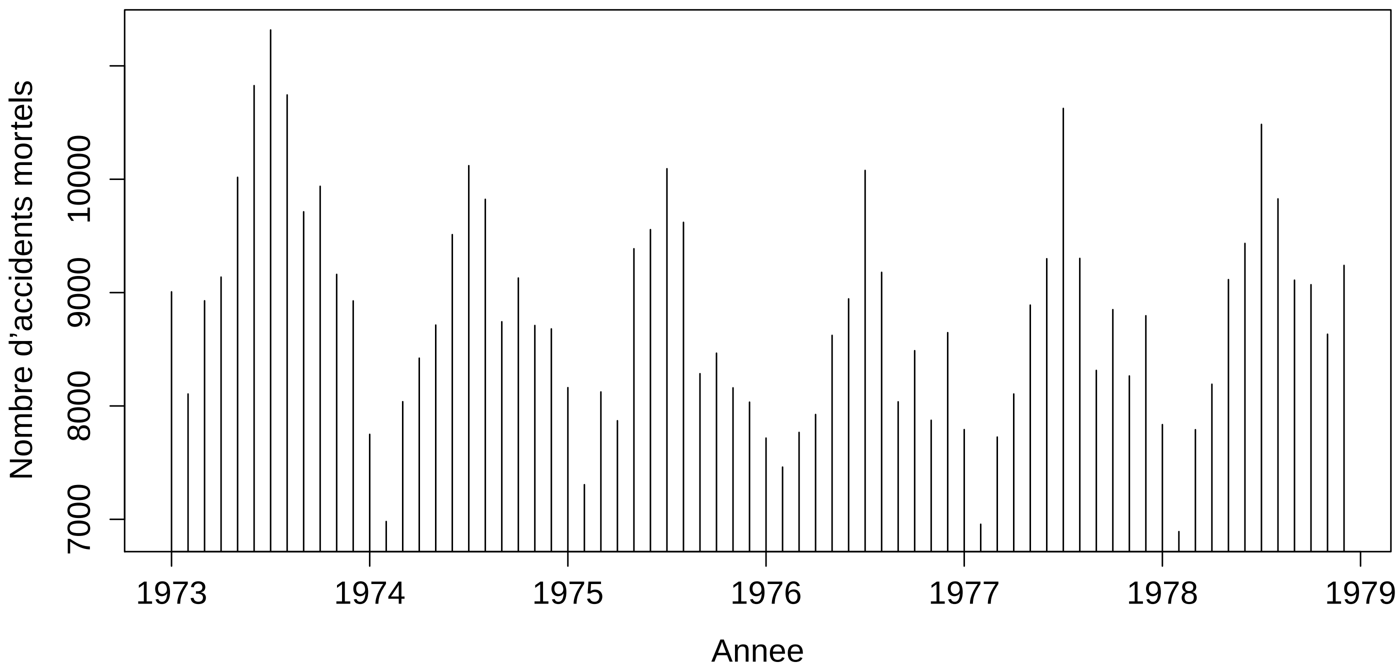

Figure 3: *Nombre d'accidents mortels par mois au USA.*

Finance

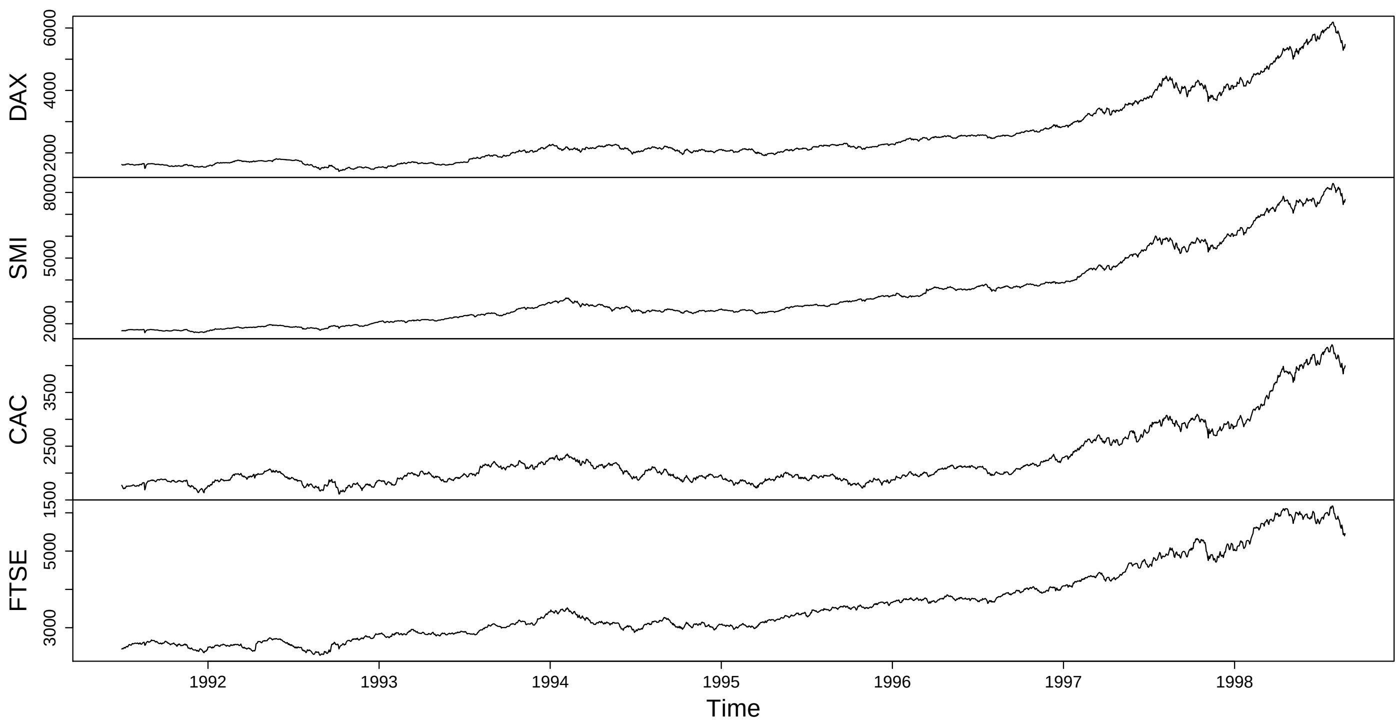

Figure 4: *Prix de fermeture journaliers de 4 indices européens: DAX (Allemagne), SMI (Suisse), CAC (France) et TFSE (UK).*

1. Un peu de théorie

Processus stochastiques

Définition 1. a) Un processus stochastique $\{Y_t\}_{t\in\mathcal{T}}$ défini sur un espace d'indice $\mathcal{T}$ est une famille de variables aléatoires définies sur un espace de probabilité commun $(\Omega, \mathcal{F}, P)$.

b) Une réalisation de $\{Y_t\}$ est le "résultat de l'expérience" $\{Y_t(\omega)\}_{t\in\mathcal{T}} = \{y_t\}_{t\in\mathcal{T}}$ pour un $\omega \in \Omega$.

- L'espace d'indice $\mathcal{T}$ sera la plupart du temps $\mathbb{R}, \mathbb{R}_+$ ou $\mathbb{Z}$.
- Cependant le processus sera souvent observé sur un sous-espace fini. . .
- En ce qui nous concerne on supposera que $\mathcal{T} = \mathbb{Z}$

Définition 2. Un processus stochastique $\{Y_t\}$ est appelé un bruit blanc si tous ses éléments sont décorrélés et tels que $\mathbb{E}[Y_t] = 0$ et $\text{Var}[Y_t] = \sigma^2$.

Exemple 1. Si $Y_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$, $t \in \mathbb{Z}$, on parle alors de bruit blanc Gaussien.

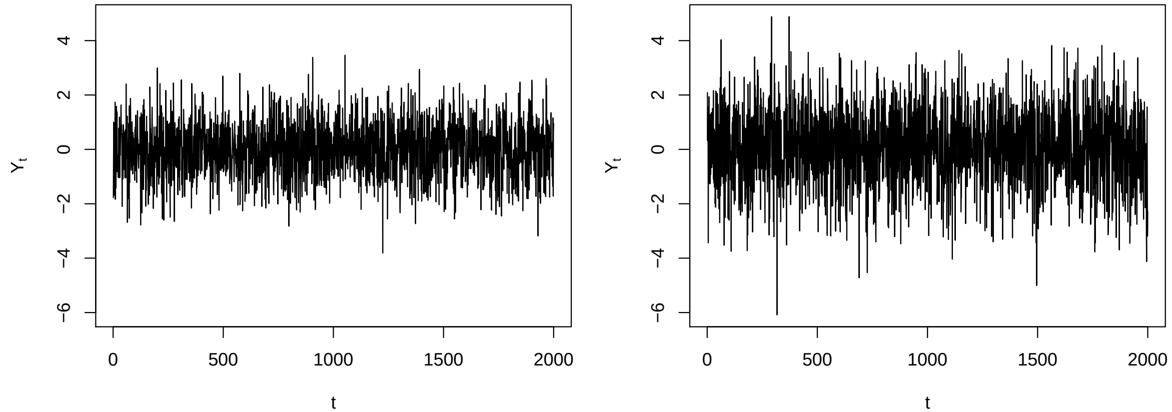

Figure 5: *Deux réalisations d'un bruit blanc gaussien avec $\sigma^2 = 1$ (gauche) et $\sigma^2 = 2$ (droite).*

Définition 3. Soit $\{Y_t\}_{t\in\mathcal{T}}$ un processus stochastique.

1. Si $\mathbb{E}[|Y_t|] < \infty$, alors on appelle moyenne du processus $\mu_t = \mathbb{E}[Y_y]$.
2. Si $\mathrm{Var}[Y_t] < \infty$, alors on appelle la fonction d'(auto)covariance du processus la fonction

$$\gamma(s,t) = \mathrm{Cov}(Y_s, Y_t) = \mathbb{E}\{(Y_s - \mu_s)(Y_t - \mu_t)\},$$

et la fonction d'(auto)correlation la fonction

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

Remarque. On a $\mathrm{Var}[Y_t] = \mathrm{Cov}(Y_t, Y_t) = \gamma(t,t)$ et $|\rho(s,t)| \leq 1$ (Cauchy–Schwartz).

Dépendance: ACF

- A partir d'observations (équidistantes) $y_1, \ldots, y_n$ de $\{Y_t\}$, γ_h est estimée par son estimateur empirique

$$\hat{\gamma}_h = \frac{1}{n-h-1} \sum_{i=1}^{n-h} (y_i - \overline{y})(y_{i+h} - \overline{y}), \qquad h = 0, 1, \ldots, n-2,$$

où $\overline{y}$ est la moyenne.

- Le corrélogramme (ACF) trace les points

$$\{(h, \hat{\rho}_h) \colon h = 1, \ldots h_{\max}\}, \qquad \hat{\rho}_h = \frac{\hat{\gamma}_h}{\hat{\gamma}_0}.$$

Exemple d'un ACF

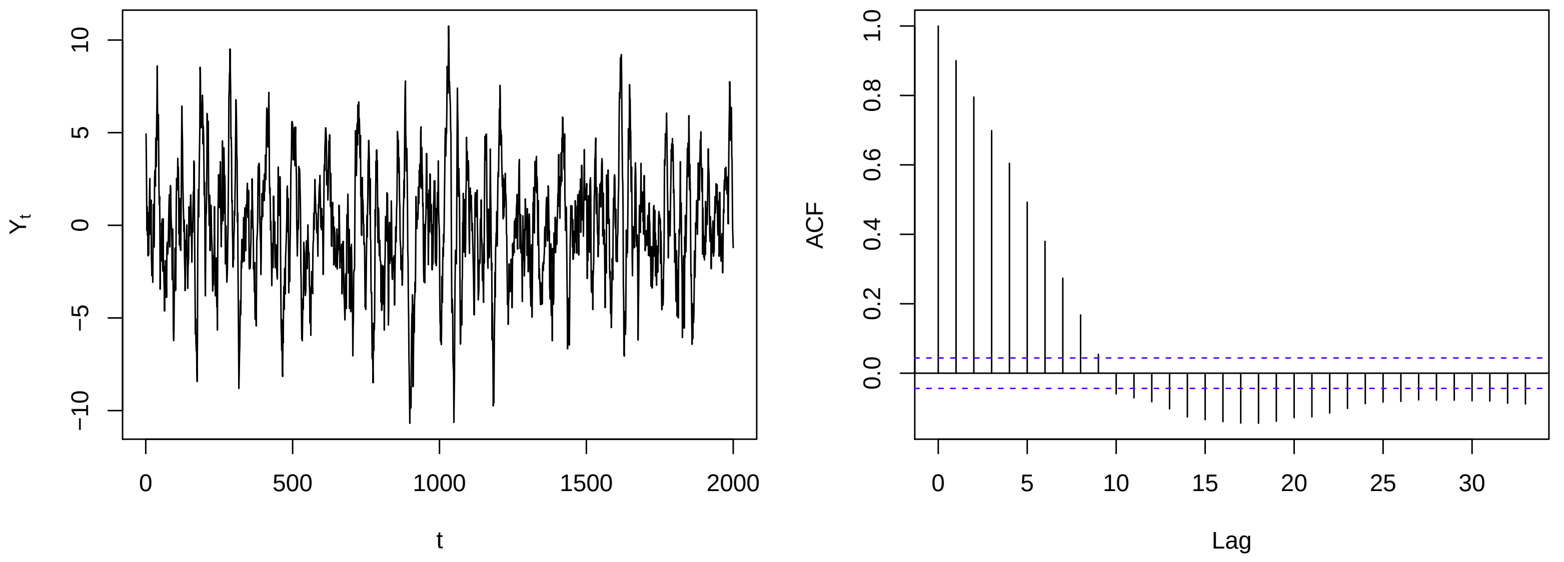

Figure 6: *Réalisation d'un MA(9) (gauche) et son ACF (droite)—que l'on ne verra pas :-(.*

Définition 4. Soient $\{Y_t\}$ un processus stationnaire et posons $\tilde{Y}_0$ et $\tilde{Y}_h$ deux combinaisons linéaires de $Y_1, \ldots, Y_{h-1}$ minimisant $\mathbb{E}[(Y_0 - \tilde{Y}_0)^2]$ et $\mathbb{E}[(Y_h - \tilde{Y}_h)^2]$ respectivement. Alors la fonction d'autocorrélation partielle est

$$\tilde{\rho}_1 = \text{Corr}(Y_0, Y_1), \qquad \tilde{\rho}_h = \text{Corr}(Y_0 - \tilde{Y}_0, Y_h - \tilde{Y}_h), \qquad h = 2, 3, \ldots$$

Le corrélogramme partiel (PACF) trace les points

$$\{(h, \tilde{\rho}_h) \colon h = 1, \ldots, h_{\max}\}.$$

- La logique derrière la corrélation partielle consiste à caractériser la corrélation entre Y_0 et Y_h tout en ignorant celle induite par $Y_1, \ldots, Y_{h-1}$.

Exemple d'un PACF

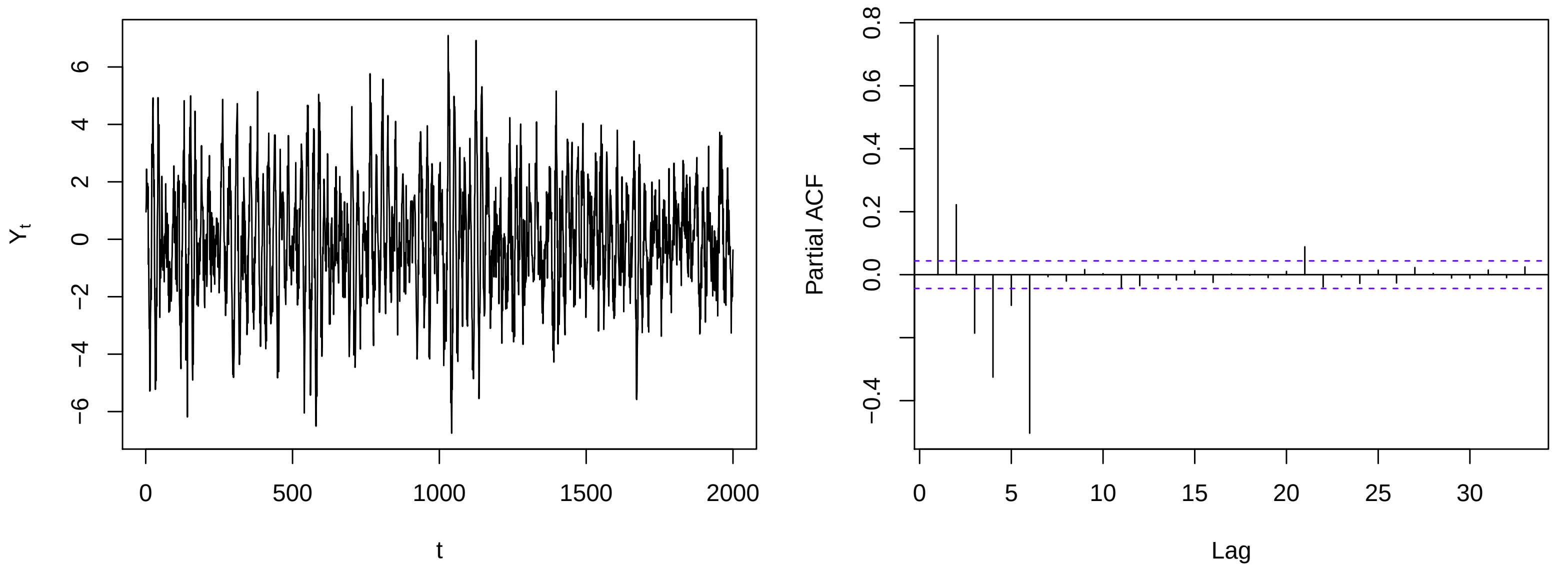

Figure 7: *Réalisation d'un AR(6) (gauche) et son PACF (droite)—que l'on ne verra pas vraiment.*

Notations: Pour un ensemble $\mathcal{S}$ et $u \in \mathcal{T}$, on écrira $u + \mathcal{S}$ pour l'ensemble $\{u + s \colon s \in \mathcal{S}\}$ et $Y_{\mathcal{S}}$ l'ensemble des variables aléatoires $\{Y_s \colon s \in \mathcal{S}\}$.

Définition 5. Un processus stochastique $\{Y_t\}_{t\in\mathcal{T}}$ est dit

1. strictement stationnaire si pour tout sous-ensemble fini $\mathcal{S} \subset \mathcal{T}$ et tout $u \in \mathcal{T}$ tel que $u + \mathcal{S} \subset \mathcal{T}$, on a $Y_{u+\mathcal{S}} \overset{\mathcal{L}}{=} Y_{\mathcal{S}}$.
2. stationnaire du deuxième ordre si la moyenne et la fonction de covariance de $\{Y_t\}$ ne dépendent ni de t ni de s.

- Si $\mathcal{T} \subset \mathbb{R}$ et $\{Y_t\}$ est stationnaire alors

$$\gamma(t, t+h) = \gamma(0, h) = \gamma(0, -h) \equiv \gamma_{|h|} = \gamma_h, \qquad h \in \mathbb{R}.$$

- Dans la suite on dira que $\{Y_t\}$ est stationnaire pour signifier la stationnarité du deuxième ordre

2. Le processus autorégressif AR(1)

Définition 6. Un processus $\{Y_t\}$ est appelé un processus autorégressif d'ordre 1, noté AR(1), si

$$Y_t = \alpha Y_{t-1} + \varepsilon_t, \qquad t \in \mathbb{Z},$$

où α est le paramètre contrôlant "l'autorégressivité" et $\{\varepsilon_t\}$ est un bruit blanc.
De manière plus générale, on peut définir un processus de moyenne $\mu \in \mathbb{R}$ par

$$Y_t - \mu = \alpha(Y_{t-1} - \mu) + \varepsilon_t, \qquad t \in \mathbb{Z}.$$

On utilisera toutefois la première expression pour la suite.

Exercice 1. Pourquoi est-il désirable d'avoir $|\alpha| < 1$? Que se passe-t-il lorsque $\alpha = 1$?

Plusieurs réalisations d'un AR(1)

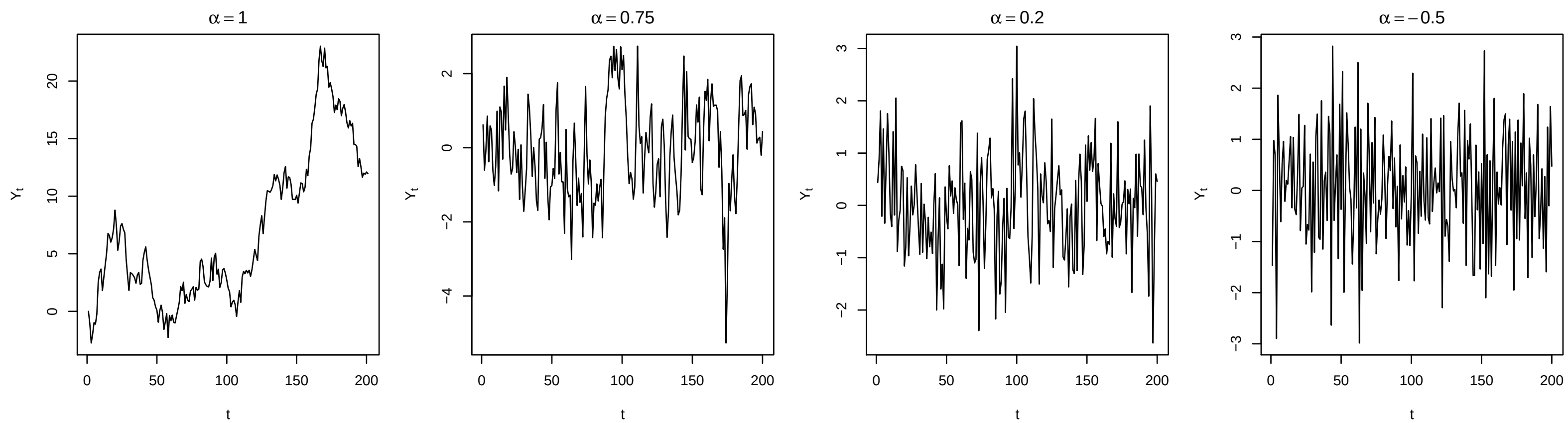

Figure 8: *Réalisation d'un AR(1) avec (de gauche à droite)* $\alpha = 1, 0.75, 0.2, -0.5.$

Lemme 1. *Soit* $\{Y_t\}$ *un AR(1) de paramètre d'autorégression* $|\alpha| < 1$. *Alors* $\{Y_t\}$ *admet la représentation linéaire suivante*

$$Y_t = \sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j},$$

et ainsi

$$\gamma_h = \alpha^{|h|} \frac{\sigma^2}{1-\alpha^2}.$$

ACF de l'AR(1)

Exercice 2. A quel type d'ACF allez vous vous attendre en présence d'un AR(1) ?

ACF de l'AR(1)

Exercice 2. A quel type d'ACF allez vous vous attendre en présence d'un AR(1) ?

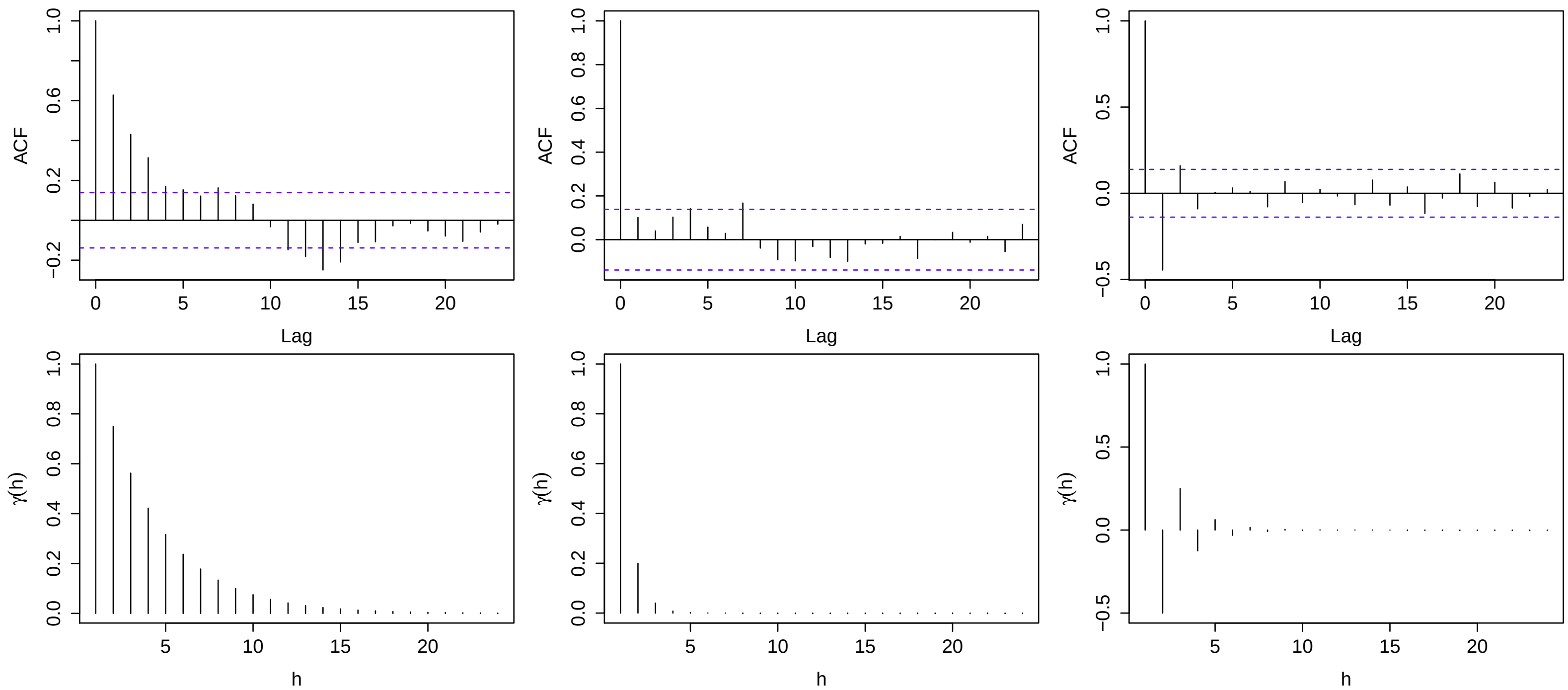

Figure 9: *ACF d'AR(1) avec $\alpha = 0.75, 0.2, -0.5$ (de gauche à droite). La ligne du haut est l'ACF estimé et celle du bas le théorique.*

Exercice 3. Montrez que pour $Y_1, \ldots, Y_n$ issues d'un processus stationnaire on a

$$\text{Var}[\overline{Y}] = \frac{\gamma_0}{n} \left\{ 1 + \frac{2}{n} \sum_{h=1}^{n-1} (n-h)\rho_h \right\},$$

et que lors d'un AR(1), $|\alpha| < 1$, on a pour n assez grand

$$\text{Var}[\overline{Y}] \sim \frac{\gamma_0}{n} \left(1 + \frac{2\alpha}{1-\alpha} \right).$$

Exercice 3. Montrez que pour $Y_1, \ldots, Y_n$ issues d'un processus stationnaire on a

$$\text{Var}[\overline{Y}] = \frac{\gamma_0}{n} \left\{ 1 + \frac{2}{n} \sum_{h=1}^{n-1} (n-h)\rho_h \right\},$$

et que lors d'un AR(1), $|\alpha| < 1$, on a pour n assez grand

$$\text{Var}[\overline{Y}] \sim \frac{\gamma_0}{n} \left(1 + \frac{2\alpha}{1-\alpha} \right).$$

- Pour un bruit blanc on a évidemment $\text{Var}[\overline{Y}] = n^{-1}\sigma^2$.
- Ici la variance est plus grande comme une conséquence de la dépendance entre les $Y_1, \ldots, Y_n$.

Définition 7. Un processus stochastique $\{Y_t\}$ est dit Markovien si pour tout ensembles $\mathcal{P} = \{t_1 < \cdots < t_K < t\}$ (le passé), la loi de Y_{t+1} dépend de celle de $Y_{\mathcal{P}}$ uniquement au travers de Y_t, i.e.,

$$\Pr[Y_{t+1} \leq y_{t+1} \mid Y_{\mathcal{P}} = y_{\mathcal{P}}] = \Pr[Y_{t+1} \leq y_{t+1} \mid Y_t = y_t].$$

On dit alors que $\{Y_t\}$ est une chaîne de Markov (d'ordre 1)—cela se généralise évidemment à l'ordre p.

Exercice 4. Montrez qu'un AR(1) est une chaîne de Markov.

Définition 7. Un processus stochastique $\{Y_t\}$ est dit Markovien si pour tout ensembles $\mathcal{P} = \{t_1 < \cdots < t_K < t\}$ (le passé), la loi de Y_{t+1} dépend de celle de $Y_{\mathcal{P}}$ uniquement au travers de Y_t, i.e.,

$$\Pr[Y_{t+1} \leq y_{t+1} \mid Y_{\mathcal{P}} = y_{\mathcal{P}}] = \Pr[Y_{t+1} \leq y_{t+1} \mid Y_t = y_t].$$

On dit alors que $\{Y_t\}$ est une chaîne de Markov (d'ordre 1)—cela se généralise évidemment à l'ordre p.

Exercice 4. Montrez qu'un AR(1) est une chaîne de Markov.

- ☐ La fonction d'autocorrélation partielle d'un AR(1) est donc toujours nulle sauf lorsque $h = 1$!

Exercice 5. Soit $Y_1, \ldots, Y_n$ les n premières observations d'un AR(1) Gaussien, i.e.,

$$Y_{t+1} = \alpha Y_t + \varepsilon_t, \qquad \varepsilon_t \overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2).$$

Donnez l'expression de la vraisemblance.

Prédictions

- □ A partir des observations $y_1, \ldots, y_n$, l'AR(1) (gaussien) ajusté nous donne $y_n = \hat{\alpha} y_{n-1} + r_t$
- □ En itérant on a $Y_{n+1} = \hat{\alpha} y_n + \varepsilon_{n+1}$
- □ Malheureusement ε_{n+1} est non observé mais une prédiction raisonnable est 0—car $\{\varepsilon_t\}$ est de moyenne nulle.
- □ Ainsi "la" prédiction et son intervalle de confiance de 95% associé est

$$\hat{y}_{n+1} = \hat{\alpha} y_n \pm 1.96 \hat{\sigma}.$$

Exercice 6. a) Trouvez "la" prédiction au temps $n+2$ et de manière plus générale au temps $n+k$ ainsi que leurs intervalles de confiance respectifs.

b) Que se passe-t-il lorsque $k \to \infty$?

3. Ce que nous ne verrons pas

Pour les personnes intéressées

Nous avons juste survolé les séries temporelles. Les autres processus à connaître à tout prix si les séries temporelles vous intéressent sont

- AR(p) : $Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t$
- MA(q) : $Y_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \cdots + \beta_q \varepsilon_{t-q}$
- ARMA(p, q) = AR(p) + MA(q)
- Différenciation, ARIMA(p, d, q)
- Densité spectrale
- Saisonnalité SAR(I)MA
- ARCH/GARCH

4. Les mains dans le cambouis

Nos données (rappel)

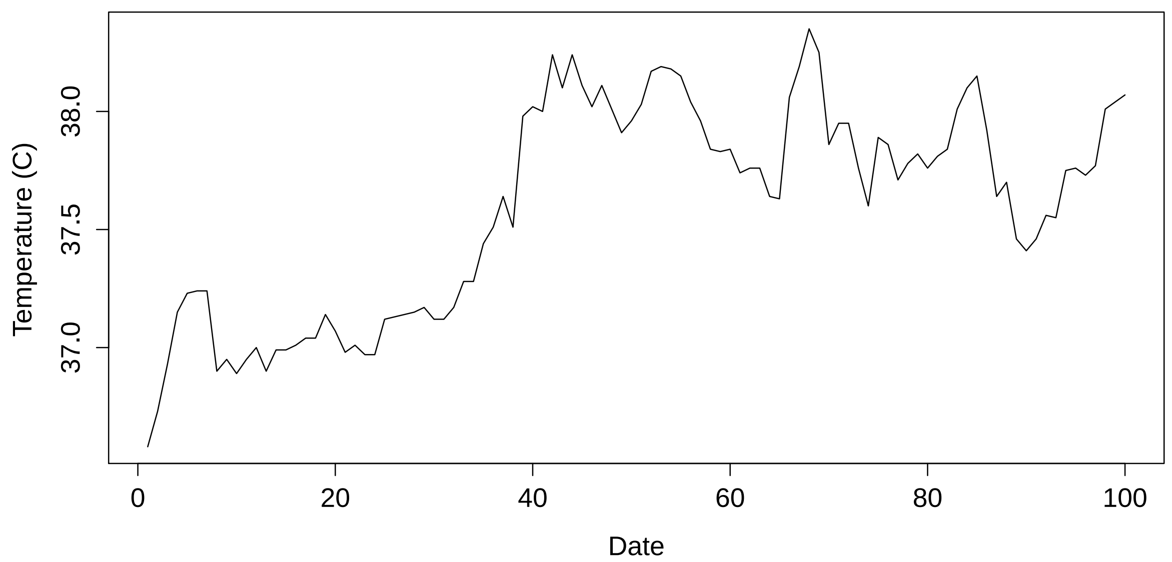

Figure 10: *Temperature corporelle d'un castor femelle (Castor canadensis) du Wisconsin. Une mesure toute les 10 minutes.*

- On dispose d'une variable binaire codant la présence dans le foyer
- Quels modèles statistiques (simples !!!) envisageriez–vous ?

On va considérer les 5 modèles suivants

- M1: $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$
- M2: $Y_1, \ldots, Y_\gamma \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ et $Y_{\gamma+1}, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu + \delta, \sigma^2)$, $\gamma = 38$.
- M3: $Y_1, \ldots, Y_n \sim \text{AR}(1)$
- M4: $Y_1, \ldots, Y_n \sim \text{AR}(1)$ avec changement de moyenne au temps $t = \gamma$, $\gamma = 38$.
- M5: Pareil que M4 mais avec γ inconnu.

Modélisation

On va considérer les 5 modèles suivants

- M1: $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$
- M2: $Y_1, \ldots, Y_\gamma \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ et $Y_{\gamma+1}, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu + \delta, \sigma^2)$, $\gamma = 38$.
- M3: $Y_1, \ldots, Y_n \sim \text{AR}(1)$
- M4: $Y_1, \ldots, Y_n \sim \text{AR}(1)$ avec changement de moyenne au temps $t = \gamma$, $\gamma = 38$.
- M5: Pareil que M4 mais avec γ inconnu.

Attention pour comparer les modèles il faut qu'ils soient tous ajustés sur les mêmes observations. Problème avec Y_0?

On va considérer les 5 modèles suivants

- M1: $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$
- M2: $Y_1, \ldots, Y_\gamma \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ et $Y_{\gamma+1}, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu + \delta, \sigma^2)$, $\gamma = 38$.
- M3: $Y_1, \ldots, Y_n \sim \text{AR}(1)$
- M4: $Y_1, \ldots, Y_n \sim \text{AR}(1)$ avec changement de moyenne au temps $t = \gamma$, $\gamma = 38$.
- M5: Pareil que M4 mais avec γ inconnu.

Attention pour comparer les modèles il faut qu'ils soient tous ajustés sur les mêmes observations. Problème avec Y_0? On écartera y_1 de tous nos modèles.

Ce que j'aimerai que vous fassiez

- ☐ Construire des fonctions `R` calculant la vraisemblance pour chaque modèle et trouver le MLE (avec les erreurs standards).
- ☐ Faire une sélection de modèles, adéquation du modèle
- ☐ Faire de joli graphiques (selon votre inspiration)