STAT2—Introduction to Time Series

Mathieu Ribatet—Full Professor of Statistics

Some bibliographic references

[1] P.J. Brockwell and R.A. Davis. *Time Series: Theory and Methods*. Springer Series in Statistics. Springer, 2009.

[2] P.J. Brockwell and R.A. Davis. *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer International Publishing, 2016.

[3] Robert Shumway and David Stoffer. *Time Series Analysis and Its Applications With R Examples*, volume 9. 01 2011.

Motivation

- Usually in statistics we often suppose that we have independent (or even identically distributed) realizations, i.e.,

$$X_1, \ldots, X_n \overset{\text{ind}}{\sim} F_1, \ldots, F_n, \qquad X_1, \ldots, X_n \overset{\text{iid}}{\sim} F.$$

- Time series are about analysis of ordered observations, most often a time ordering.
- As so, observation will show serial dependence.
- Many type of dependence exists and in this course we will cover only a few.

Definition 1. A stochastic process $\{X_t : t \in T\}$ defined on a index set T is a collection of random variable defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 1. A stochastic process $\{X_t : t \in T\}$ defined on a index set T is a collection of random variable defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 2. A time series is a stochastic process whose index set T is one of $\mathbb{N}, \mathbb{Z}, [0, \infty)$ or $\mathbb{R}$.

Definition 3. We call sample path of a stochastic process $\{X_t : t \in T\}$ the mapping $t \mapsto X_t(\omega)$, $\omega \in \Omega$.

Stochastic processes and time series

Definition 1. A stochastic process $\{X_t \colon t \in T\}$ defined on a index set T is a collection of random variable defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 2. A time series is a stochastic process whose index set T is one of $\mathbb{N}, \mathbb{Z}, [0, \infty)$ or $\mathbb{R}$.

Definition 3. We call sample path of a stochastic process $\{X_t \colon t \in T\}$ the mapping $t \mapsto X_t(\omega)$, $\omega \in \Omega$.

Remark. Most often, time series will have as index set $T = \mathbb{Z}$. In this course, we will assume so!

Some time serires

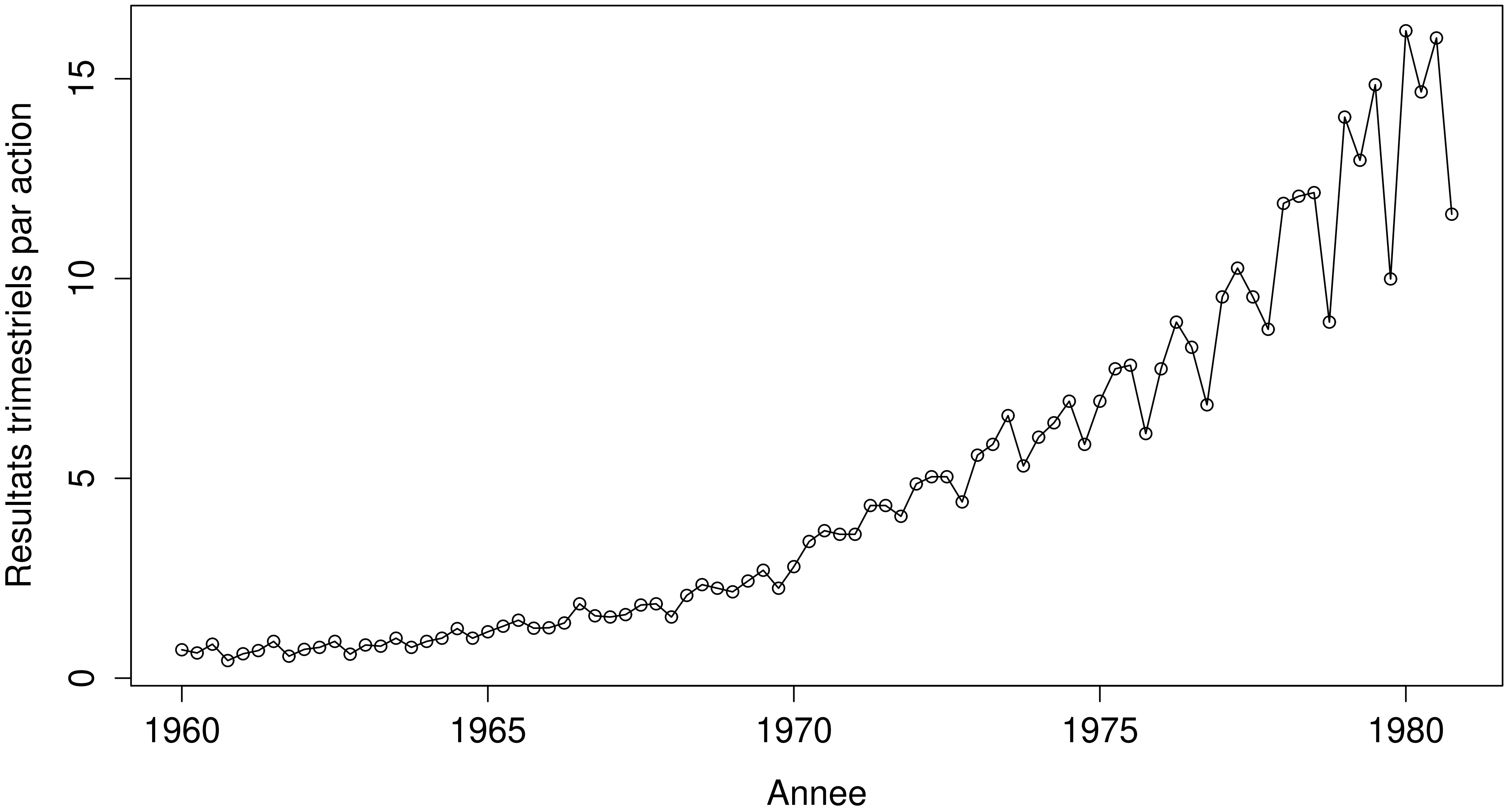


Figure 1: *Johnson and Johnson quarterly earnings per share from 1960 to 1980.*

Some time serires

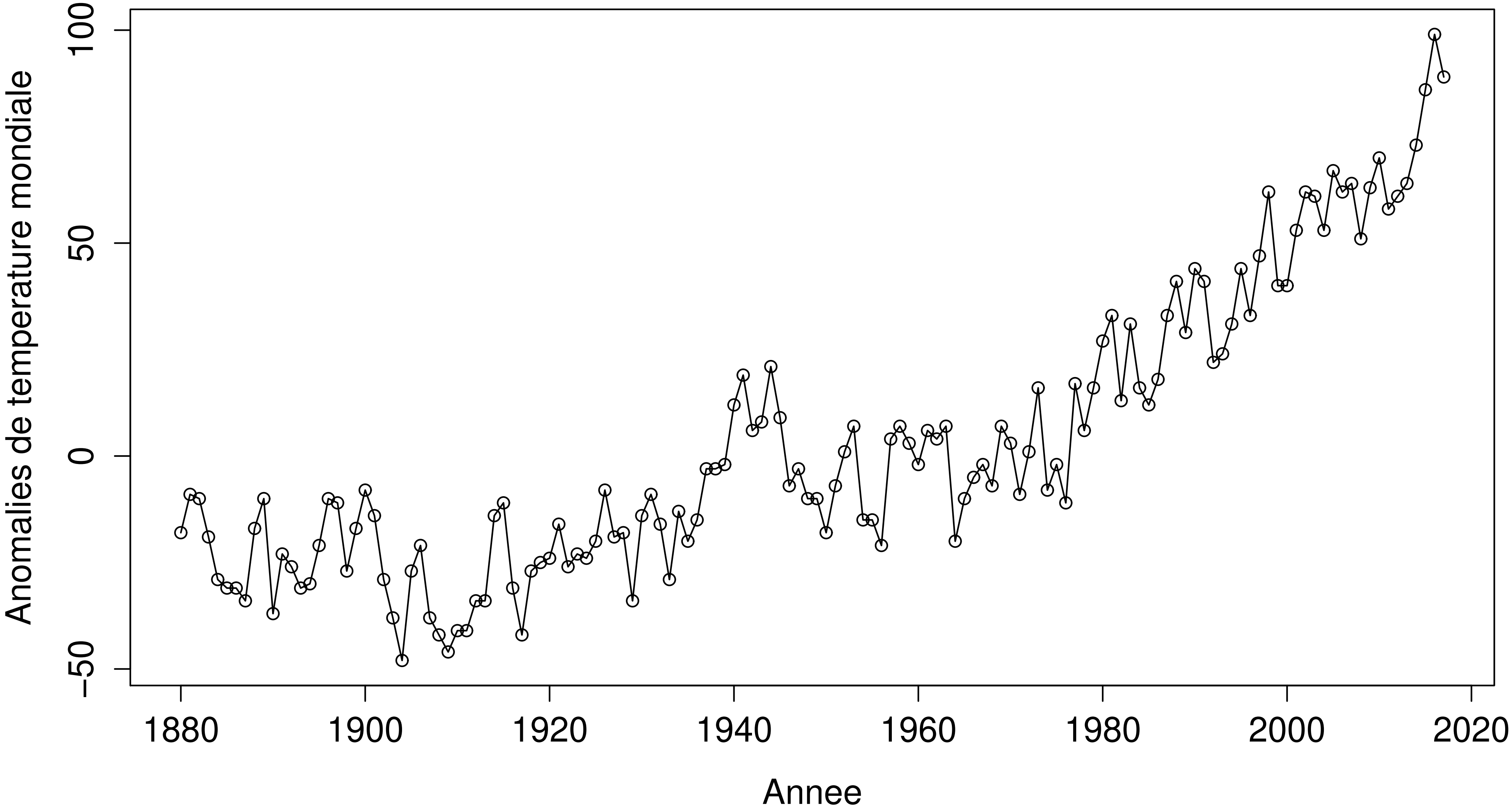


Figure 1: *Global temperature anomalies since 1880—reference period: 1951–1980.*

Some time serires

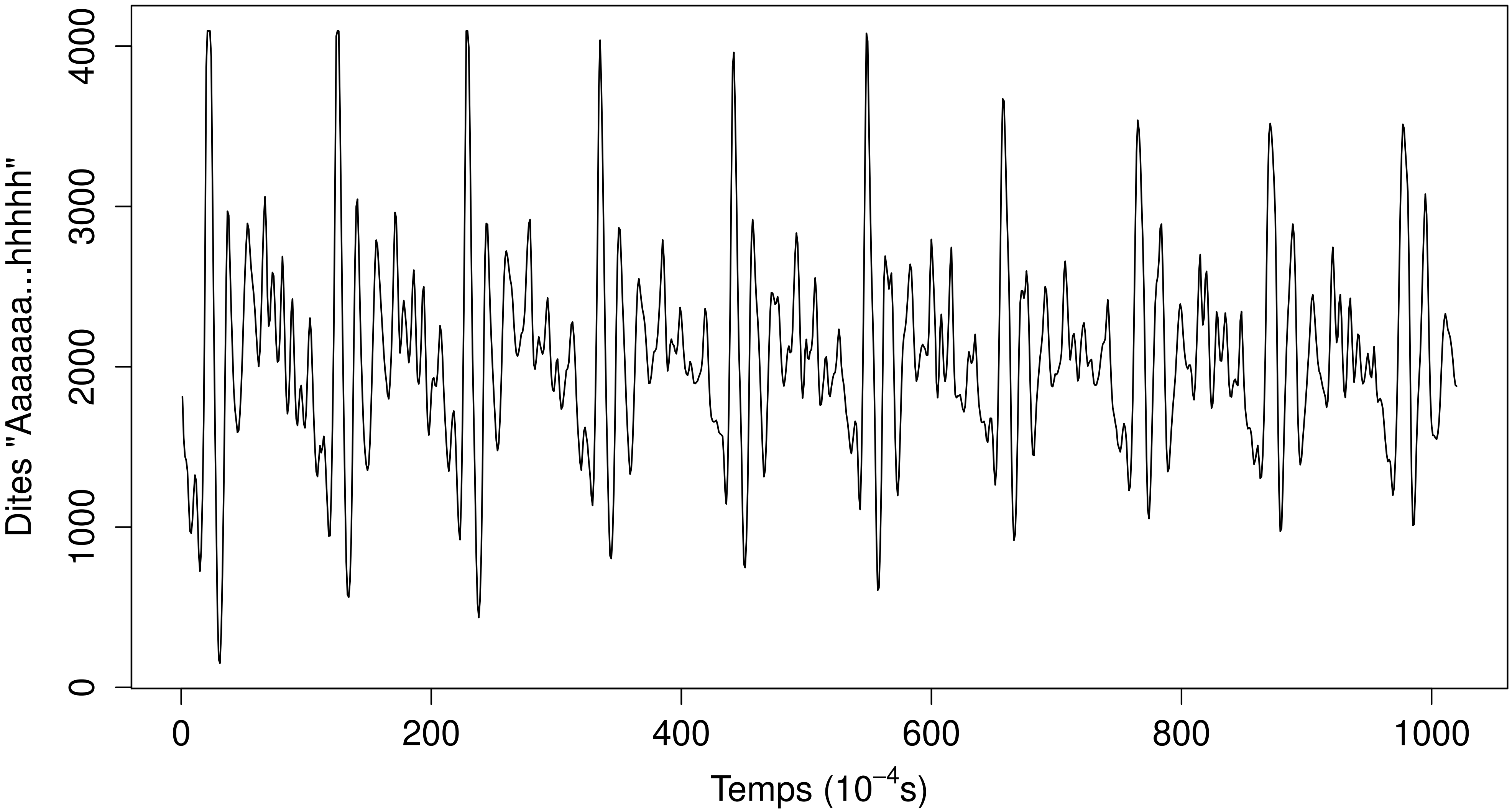


Figure 1: *Sample of recorded speech for the phrase "aaahhh".*

Some time serires

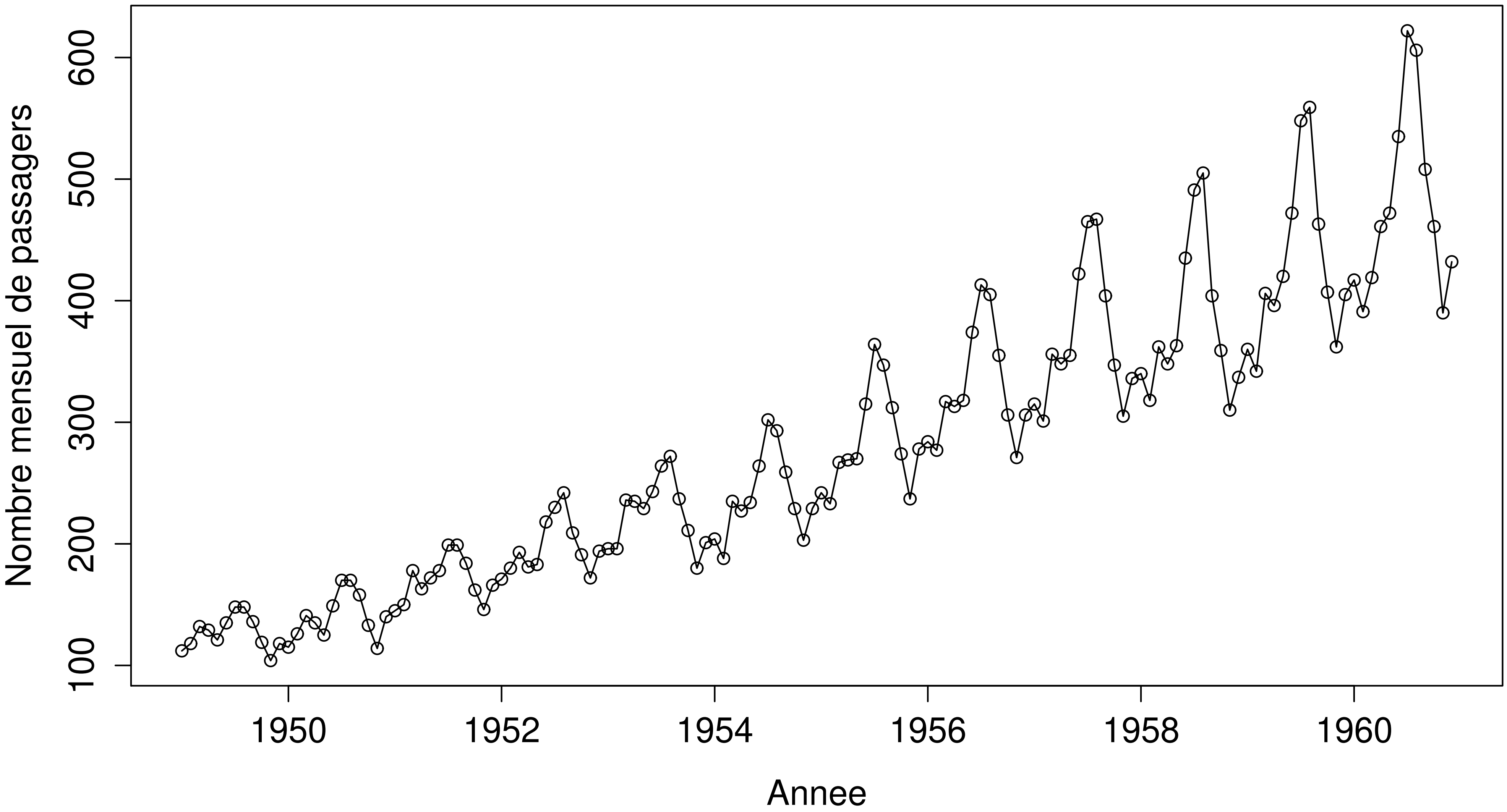


Figure 1: *Monthly totals of international airline passengers from 1949 to 1960.*

1. Basic quantities

Definition 4. Consider the set
$\mathcal{T} = \{\mathbf{t} = (t_1, \ldots, t_n) \in \mathbb{Z}^n \colon t_1 < t_2 < \cdots < t_n, n = 1, 2, \ldots\}$. The finite dimensional distributions of $\{X_t \colon t \in \mathbb{Z}\}$ are the functions $\{\mathbf{x} \mapsto F_{\mathbf{t}}(\mathbf{x}), \mathbf{t} \in \mathcal{T}\}$ where

$$F_{\mathbf{t}}(\mathbf{x}) = \Pr\left(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n\right), \qquad \mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n.$$

Definition 5. A time series $\{X_t \colon t \in \mathbb{Z}\}$ is said strictly stationary if the finite dimensional distributions of $\{X_{t+h} \colon t \in \mathbb{Z}\}$, $h \in \mathbb{Z}$, and that of $\{X_t \colon t \in \mathbb{Z}\}$ are identical. identiques.

Definition 4. Consider the set
$\mathcal{T} = \{\mathbf{t} = (t_1, \ldots, t_n) \in \mathbb{Z}^n \colon t_1 < t_2 < \cdots < t_n, n = 1, 2, \ldots\}$. The finite dimensional distributions of $\{X_t \colon t \in \mathbb{Z}\}$ are the functions $\{\mathbf{x} \mapsto F_{\mathbf{t}}(\mathbf{x}), \mathbf{t} \in \mathcal{T}\}$ where

$$F_{\mathbf{t}}(\mathbf{x}) = \Pr\left(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n\right), \qquad \mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n.$$

Definition 5. A time series $\{X_t \colon t \in \mathbb{Z}\}$ is said strictly stationary if the finite dimensional distributions of $\{X_{t+h} \colon t \in \mathbb{Z}\}$, $h \in \mathbb{Z}$, and that of $\{X_t \colon t \in \mathbb{Z}\}$ are identical. identiques.

☞ Often it is a much too strong hypothesis (that we cannot check for in practice). Hence a “relaxed definition” is typically assumed.

Definition 6. A time series $\{X_t \colon t \in \mathbb{Z}\}$ is second order if, for all $t \in \mathbb{Z}$, $\text{Var}(X_t) < \infty$.

☞ Ceci nous permet de considérer les notions suivantes

Definition 6. A time series $\{X_t \colon t \in \mathbb{Z}\}$ is second order if, for all $t \in \mathbb{Z}$, $\text{Var}(X_t) < \infty$.

☞ Ceci nous permet de considérer les notions suivantes

Definition 7. Let $\{X_t \colon t \in \mathbb{Z}\}$ a second order time series. The trend of the above time series is defined by

$$\begin{aligned} \mu \colon \mathbb{Z} &\longrightarrow \mathbb{R} \\ t &\longmapsto \mu(t) = \mathbb{E}(X_t). \end{aligned}$$

Further the autocovariance function is

$$\begin{aligned} \gamma \colon \mathbb{Z}^2 &\longrightarrow \mathbb{R} \\ (s,t) &\longmapsto \gamma(s,t) = \text{Cov}\left(X_s, X_t\right) = \mathbb{E}\left[\{X_s - \mu(s)\}\{X_t - \mu(t)\}\right]. \end{aligned}$$

Definition 8. Let $\{X_t : t \in \mathbb{Z}\}$ a second order time series. The autocorrelation function is given by

$$\begin{aligned} \rho \colon \mathbb{Z}^2 &\longrightarrow [-1, 1] \\ (s, t) &\longmapsto \rho(s, t) = \frac{\gamma(s, t)}{\sqrt{\gamma(s, s)\gamma(t, t)}}. \end{aligned}$$

Autocorrelation function

Definition 8. Let $\{X_t : t \in \mathbb{Z}\}$ a second order time series. The autocorrelation function is given by

$$\rho\colon \mathbb{Z}^2 \longrightarrow [-1, 1]$$
$$(s,t) \longmapsto \rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}.$$

☞ $|\rho(s,t)| \leq 1$ (Cauchy–Schwartz).

Weak stationarity

Definition 9. A time second order time series $\{X_t : t \in \mathbb{Z}\}$ is said weakly statitionary (or second order stationary) if

1. the trend $\mu(t)$ is constant, i.e., doesn't depend on t;
2. the autocovariance function $\gamma(t, t+h)$ doesn't depend on t for all $h \in \mathbb{Z}$.

Weak stationarity

Definition 9. A time second order time series $\{X_t : t \in \mathbb{Z}\}$ is said weakly statitionary (or second order stationary) if

1. the trend $\mu(t)$ is constant, i.e., doesn't depend on t;
2. the autocovariance function $\gamma(t, t+h)$ doesn't depend on t for all $h \in \mathbb{Z}$.

☞ We often use the abuse of phrasing saying "stationary" rather than "weakly stationary".

Proposition 1. *If $\{X_t : t \in \mathbb{Z}\}$ is stationary then*

$$\gamma(t, t+h) = \gamma(0, h) = \gamma(0, -h) := \gamma(h), \qquad \rho(t, t+h) := \rho(h),$$

i.e., the autocovariance function, and therefore the autocorrelation function, is a function of a single variate and is symmetric about 0. We will refer to h as the lag.

Consider a stationary time series $\{X_t : t \in \mathbb{Z}\}$ observed at $X_1, \ldots, X_n$

Definition 10. The empirical autocovariance function is

$$h \mapsto \hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (X_{t+h} - \bar{X})(X_t - \bar{X}), \qquad \bar{X} = \frac{1}{n} \sum_{t=1}^{n} X_t.$$

Similarly we defined the empirical autocorrelation function (ACF) as

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

Consider a stationary time series $\{X_t : t \in \mathbb{Z}\}$ observed at $X_1, \ldots, X_n$

Definition 10. The empirical autocovariance function is

$$h \mapsto \hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (X_{t+h} - \bar{X})(X_t - \bar{X}), \qquad \bar{X} = \frac{1}{n} \sum_{t=1}^{n} X_t.$$

Similarly we defined the empirical autocorrelation function (ACF) as

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

☞ Note that we divide by n and not $n - h - 1$ to ensure that $h \mapsto \hat{\gamma}(h)$ is positive definite.

ACF of 'aaaaaahhhh'

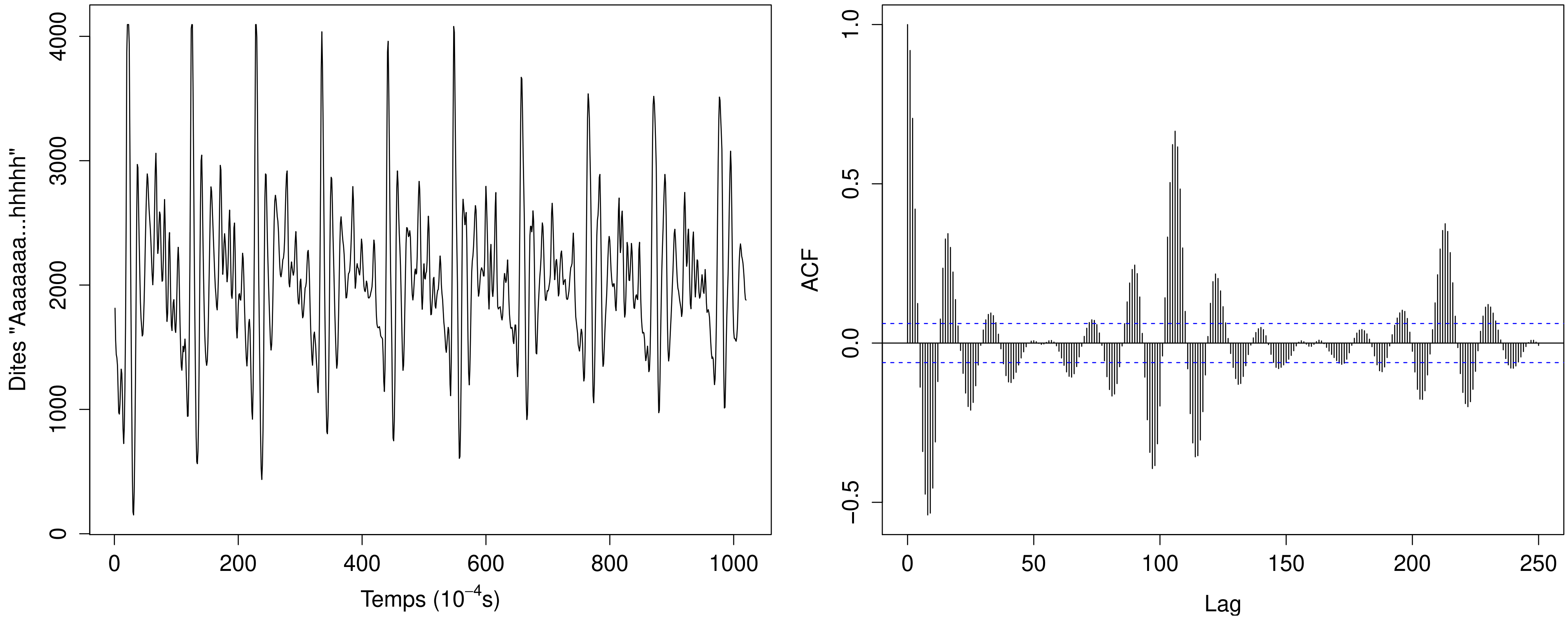


Figure 2: *Empirical autocorrelation function of 'aaaaahhhhh'.*

ACF of 'aaaaaahhhh'

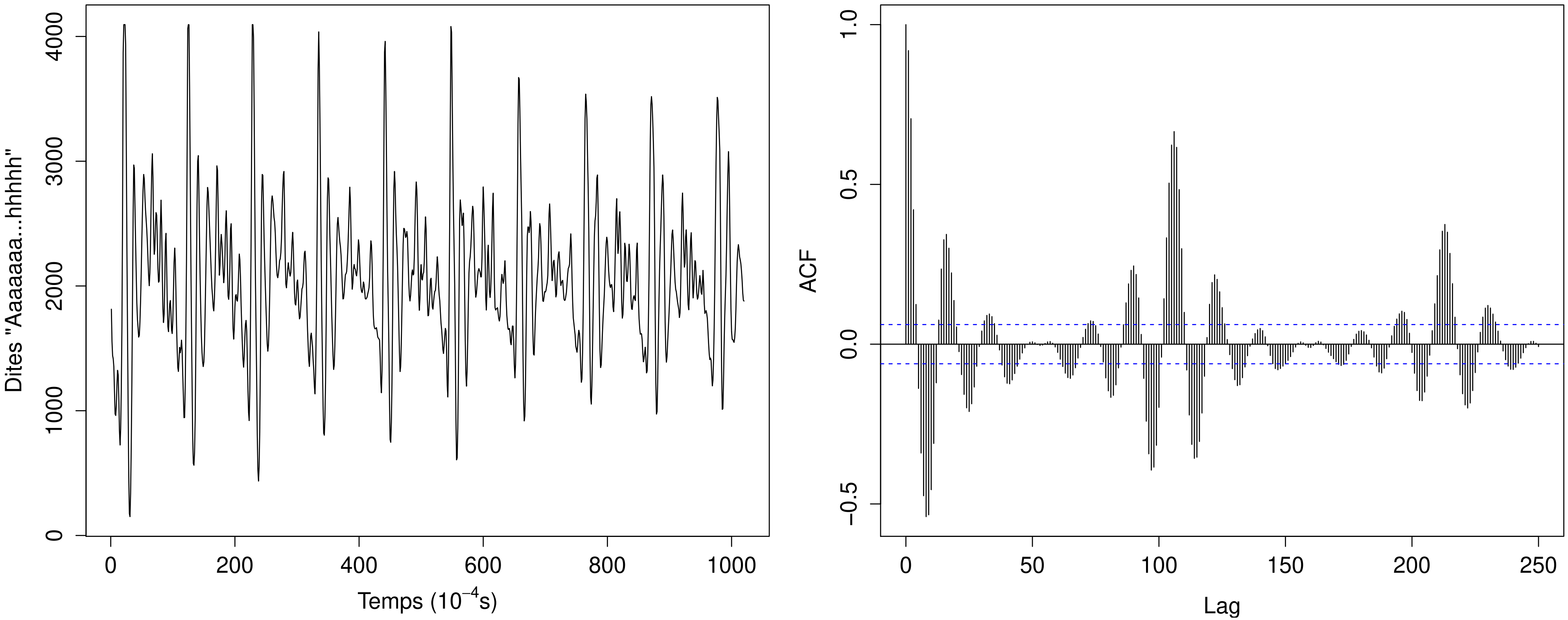


Figure 2: *Empirical autocorrelation function of 'aaaaahhhhh'.*

□ The time series shows a periodicity that is induced on the ACF

ACF of 'aaaaaahhhh'

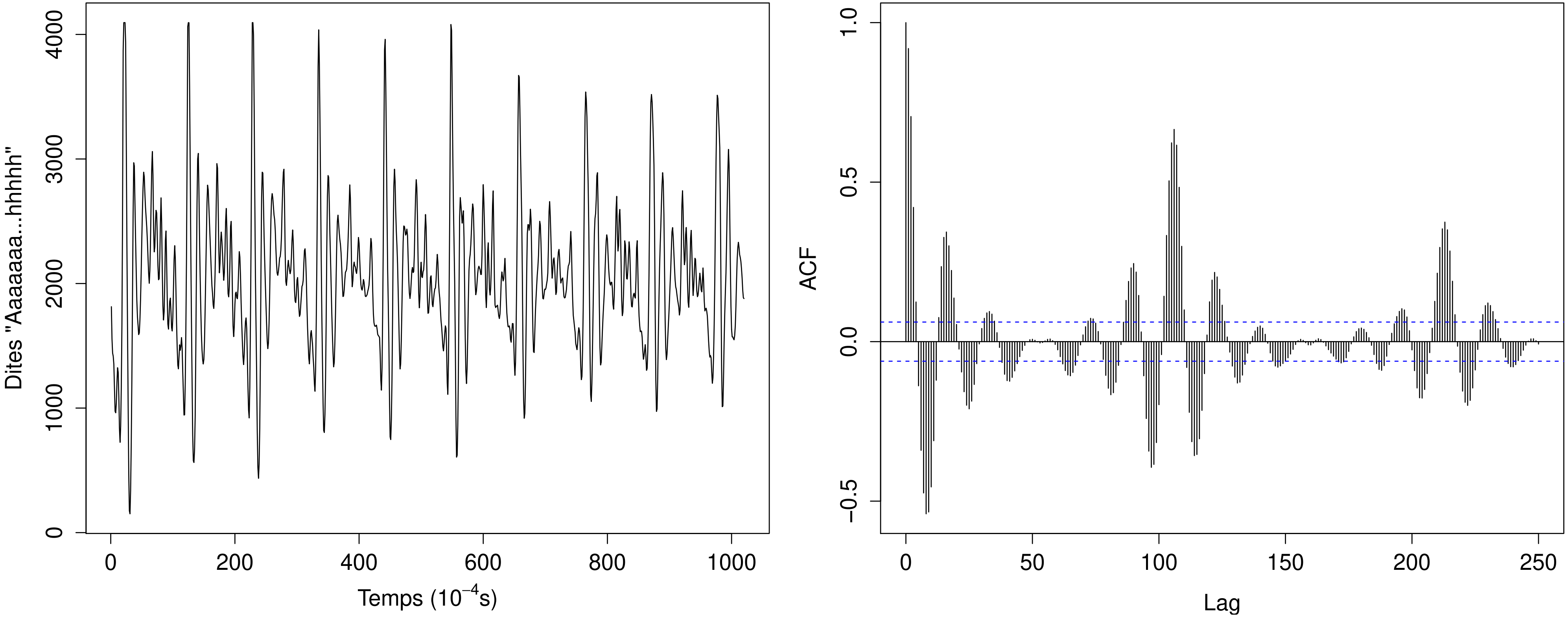


Figure 2: *Empirical autocorrelation function of 'aaaaahhhhh'.*

- The time series shows a periodicity that is induced on the ACF
- As the ACF should be analyzed on stationary time series, the analysis should be made only on the first period!

Remark about ACF

- For now, you should remember that

Time series	ACF behavior
White noise	Null
Trend	(very) Slow decreasing towards 0
Periodic	Periodic

Definition 11. Let $X_0, \ldots, X_h$ serial observation for a stationary time series and $\tilde{X}_0$ and $\tilde{X}_h$ linear forms of $X_1, \ldots, X_{h-1}$ minimizing $\mathbb{E}\{(X_0 - \tilde{X}_0)^2\}$ et $\mathbb{E}\{(X_h - \tilde{X}_h)^2\}$ respectively.
The partial autocorrelation function (PACF) is given by

$$\tilde{\rho}(1) = \text{Cor}(X_0, X_1), \qquad \tilde{\rho}(h) = \text{Cor}(X_0 - \tilde{X}_0, X_h - \tilde{X}_h), \quad h \geq 2.$$

In practice we will use its empirical version.

Definition 11. Let $X_0, \ldots, X_h$ serial observation for a stationary time series and $\tilde{X}_0$ and $\tilde{X}_h$ linear forms of $X_1, \ldots, X_{h-1}$ minimizing $\mathbb{E}\{(X_0 - \tilde{X}_0)^2\}$ et $\mathbb{E}\{(X_h - \tilde{X}_h)^2\}$ respectively.
The partial autocorrelation function (PACF) is given by

$$\tilde{\rho}(1) = \text{Cor}(X_0, X_1), \qquad \tilde{\rho}(h) = \text{Cor}(X_0 - \tilde{X}_0, X_h - \tilde{X}_h), \quad h \geq 2.$$

In practice we will use its empirical version.

- If the time series is Gaussian then

$$\tilde{\rho}(h) = \text{Cor}(X_0, X_h \mid X_1, \ldots, X_{h-1}),$$

  since, in this case, the conditional expectation is linear.
- PACF is useful to identify Markovian behavior.

PACF of “aaaaaahhhh”

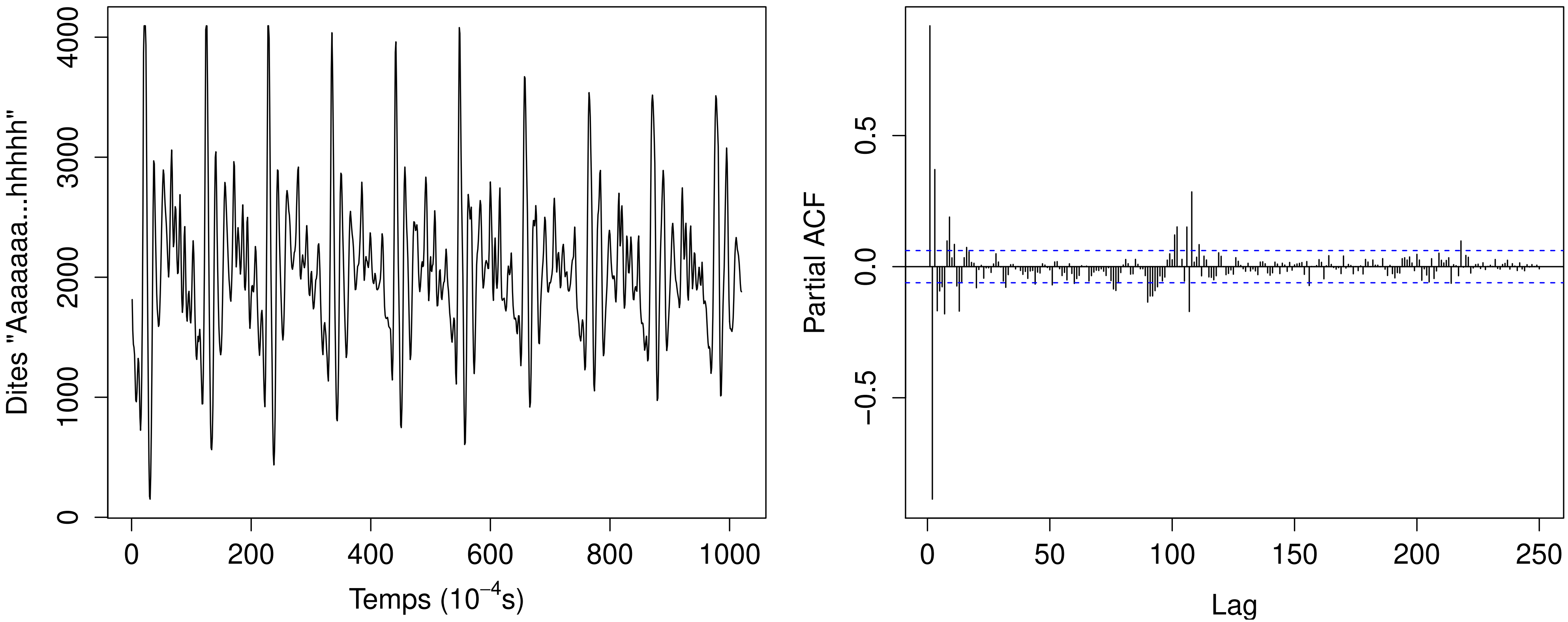


Figure 3: *Empirical partial autocorrelation function of “aaaaahhhhh”.*

Backshift operator and differentiated series

Definition 12. Let $\{X_t \colon t \in \mathbb{Z}\}$ be a time series. We define the backshift operator B as follows

$$BX_t = X_{t-1},$$

and we will say that we differentiate (at order 1) the times series $\{X_t \colon t \in \mathbb{Z}\}$ by considering the new time series

$$Y_t = X_t - X_{t-1} = (1 - B)X_t := DX_t.$$

Definition 12. Let $\{X_t : t \in \mathbb{Z}\}$ be a time series. We define the backshift operator B as follows

$$BX_t = X_{t-1},$$

and we will say that we differentiate (at order 1) the times series $\{X_t : t \in \mathbb{Z}\}$ by considering the new time series

$$Y_t = X_t - X_{t-1} = (1 - B)X_t := DX_t.$$

Remark. We can extend this operation to higher orders, i.e.,

$$\begin{aligned} B^2 X_t &= B(BX_t) = X_{t-2}, & B^3 X_t &= \dots \\ D^2 X_t &= D(DX_t) = D(X_t - X_{t-1}) = X_t - 2X_{t-1} + X_{t-2}, & D^3 X_t &= \dots \end{aligned}$$

Why differentiate a time series?

- Consider a time series with a linear trend, i.e.,

$$X_t = \beta_0 + \beta_1 t + \varepsilon_t,$$

we thus have

$$DX_t = \beta_0 + \beta_1 t + \varepsilon_t - \beta_0 - \beta_1(t-1) - \varepsilon_{t-1} = \beta_1 + D\varepsilon_t.$$

- We can generalize this result to polynomial trends, i.e.,

$$X_t = \sum_{i=0}^{k} \beta_i t_i + \varepsilon_t,$$

and we can show that $D^k X_t = k!\beta_k + D^k \varepsilon_t$.

Why differentiate a time series?

- Consider a time series with a linear trend, i.e.,

$$X_t = \beta_0 + \beta_1 t + \varepsilon_t,$$

we thus have

$$DX_t = \beta_0 + \beta_1 t + \varepsilon_t - \beta_0 - \beta_1(t-1) - \varepsilon_{t-1} = \beta_1 + D\varepsilon_t.$$

- We can generalize this result to polynomial trends, i.e.,

$$X_t = \sum_{i=0}^{k} \beta_i t_i + \varepsilon_t,$$

and we can show that $D^k X_t = k!\beta_k + D^k \varepsilon_t$.

☞ Differenciation is thus useful to remove any polynomial trends.

Illustration

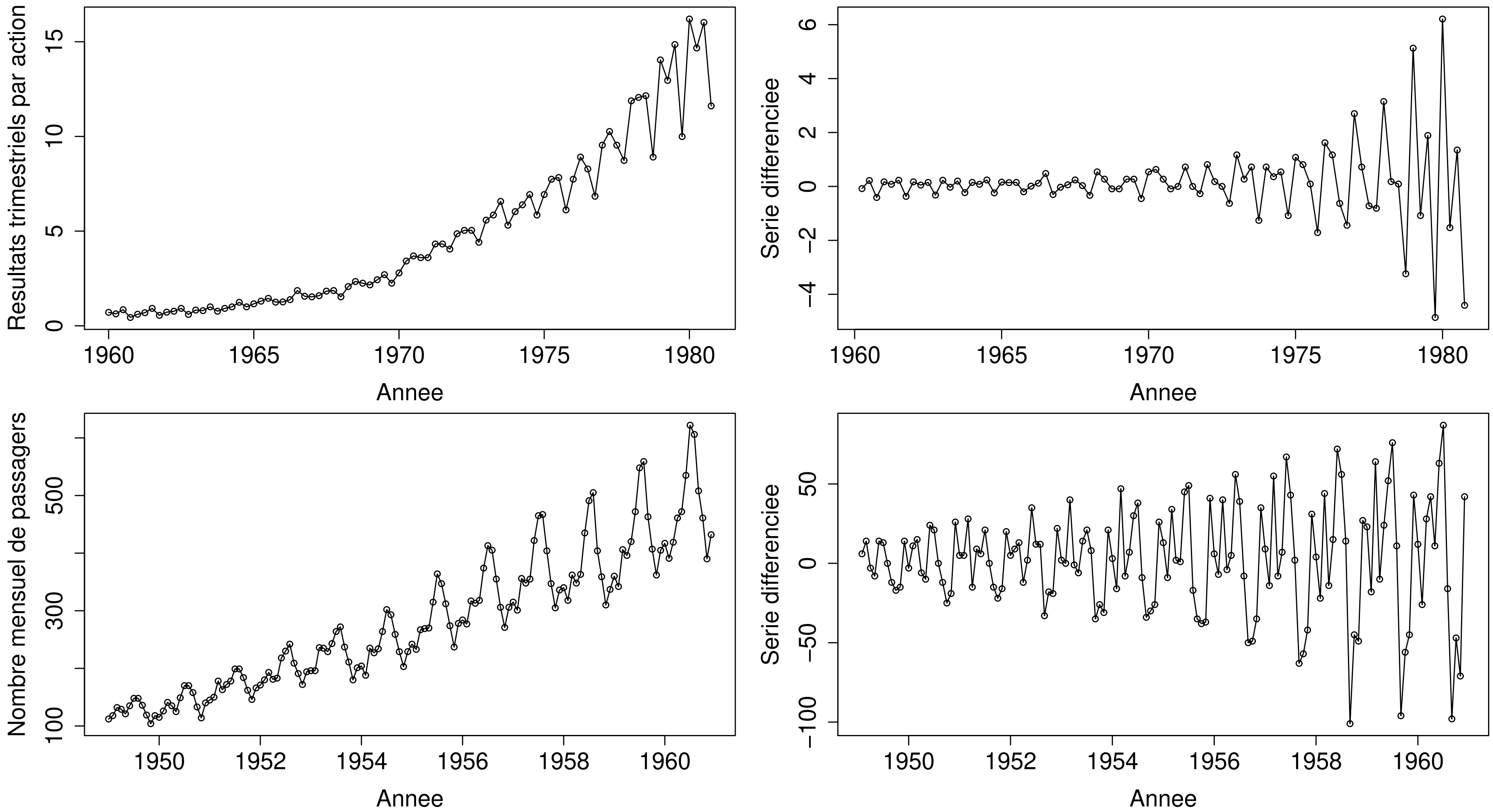


Figure 4: *Differentiation (of order 1) of the Johnson & Johnson times series and that of the international airline passengers.*

Why using the differentiation $1 - B^s$?

- Consider a periodic time series with period s, i.e.,

$$X_t = S_t + \varepsilon_t, \qquad S_{t+s} = S_t, \quad t \in \mathbb{N}.$$

  We have

$$\begin{aligned}(1 - B^s)X_t &= X_t - X_{t-s}\\ &= S_t + \varepsilon_t - S_{t-s} - \varepsilon_{t-s}\\ &= (1 - B^s)\varepsilon_t\end{aligned}$$

Why using the differentiation $1 - B^s$?

□ Consider a periodic time series with period s, i.e.,

$$X_t = S_t + \varepsilon_t, \qquad S_{t+s} = S_t, \quad t \in \mathbb{N}.$$

We have

$$\begin{aligned}(1 - B^s)X_t &= X_t - X_{t-s} \\ &= S_t + \varepsilon_t - S_{t-s} - \varepsilon_{t-s} \\ &= (1 - B^s)\varepsilon_t\end{aligned}$$

☞ The differentiation $1 - B^s$ remove any seasonal pattern.

Illustration

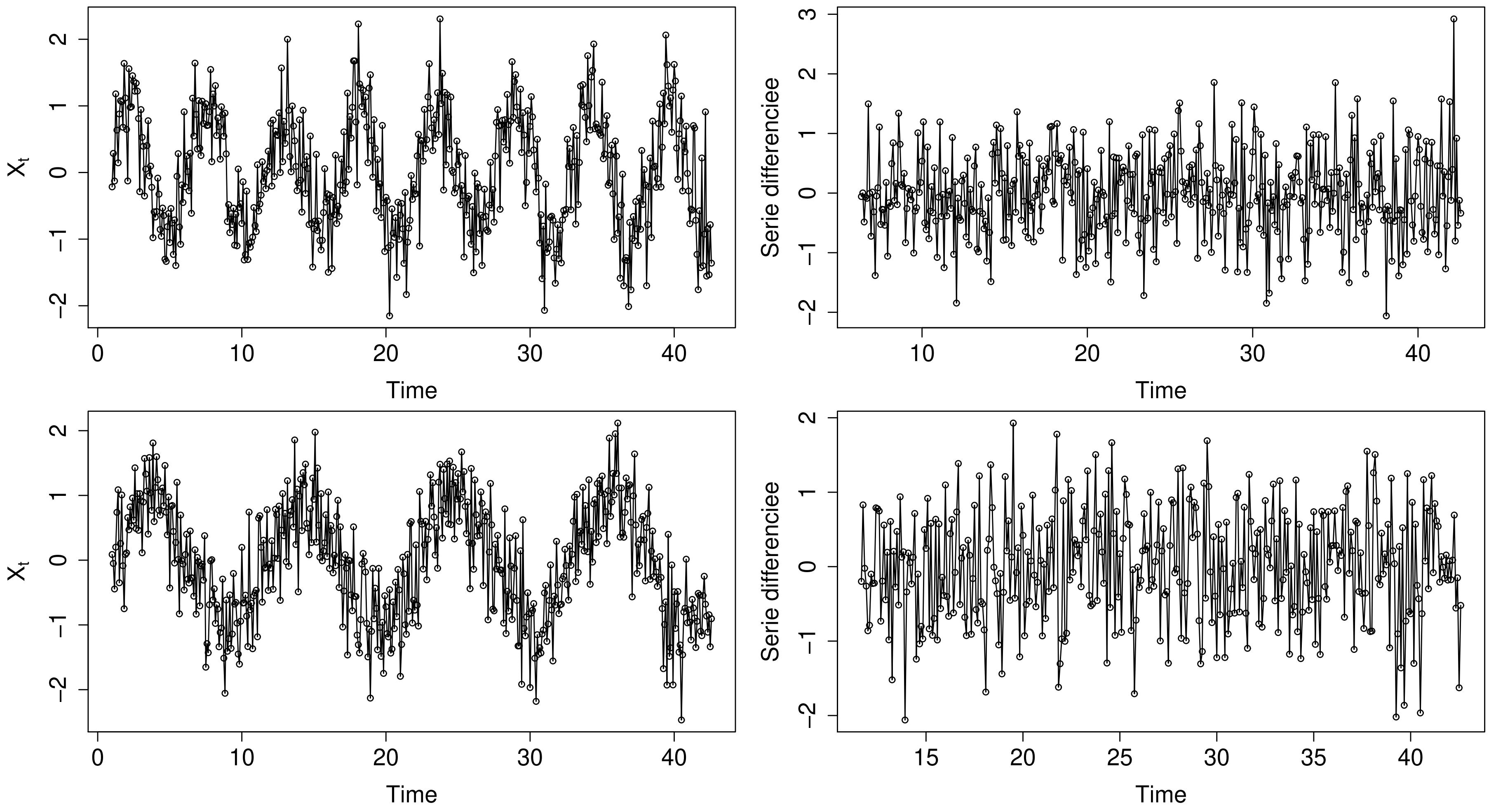


Figure 5: *Use of* $(1 - B^k)$ *for the periodic time series* $X_t = \sin(2\pi t/\omega) + \varepsilon_t$.

Variance stabilization

- Many distributions show a relationship between $\mu = \mathbb{E}(X)$ and its variance.
- Such connection is defined through the variance function $\text{Var}(X) \propto V(\mu)$:

Normal $\text{Var}(X) = \sigma^2$ so that $V(\mu) = 1$;
Poisson $\text{Var}(X) = \mu$ so that $V(\mu) = \mu$;
Gamma $\text{Var}(X) = \kappa\mu^2$ so that $V(\mu) = \mu^2$.

Variance stabilization

- Many distributions show a relationship between $\mu = \mathbb{E}(X)$ and its variance.
- Such connection is defined through the variance function $\text{Var}(X) \propto V(\mu)$:

  Normal $\text{Var}(X) = \sigma^2$ so that $V(\mu) = 1$;
  Poisson $\text{Var}(X) = \mu$ so that $V(\mu) = \mu$;
  Gamma $\text{Var}(X) = \kappa\mu^2$ so that $V(\mu) = \mu^2$.

Proposition 2. *If a random variable X has variance function $V(\mu)$, then*

$$Y = h(X), \quad h(x) = \int_{x_-}^{x} V(u)^{-1/2} du, \quad x_- = \inf\{x \in \mathbb{R} \colon \Pr(X > x_-) > 0\},$$

has an (approximately) constant variance.
In particular, when $V(\mu) = \mu^\lambda$, the function $h(x) = x^{(2-\lambda)/2}$ stabilize the variance.

Exercise 1. Proof it!

- Data are counts so the Poisson distribution may be sensible...
- Since for such distribution $\mathbb{E}(X) = \mathsf{Var}(X)$, we have $V(\mu) = \mu$.
- One may thus expect that the transformed time series $\{Y_t = \sqrt{X_t} : t \geq 0\}$ has constant variance, i.e., homoscedasticity.

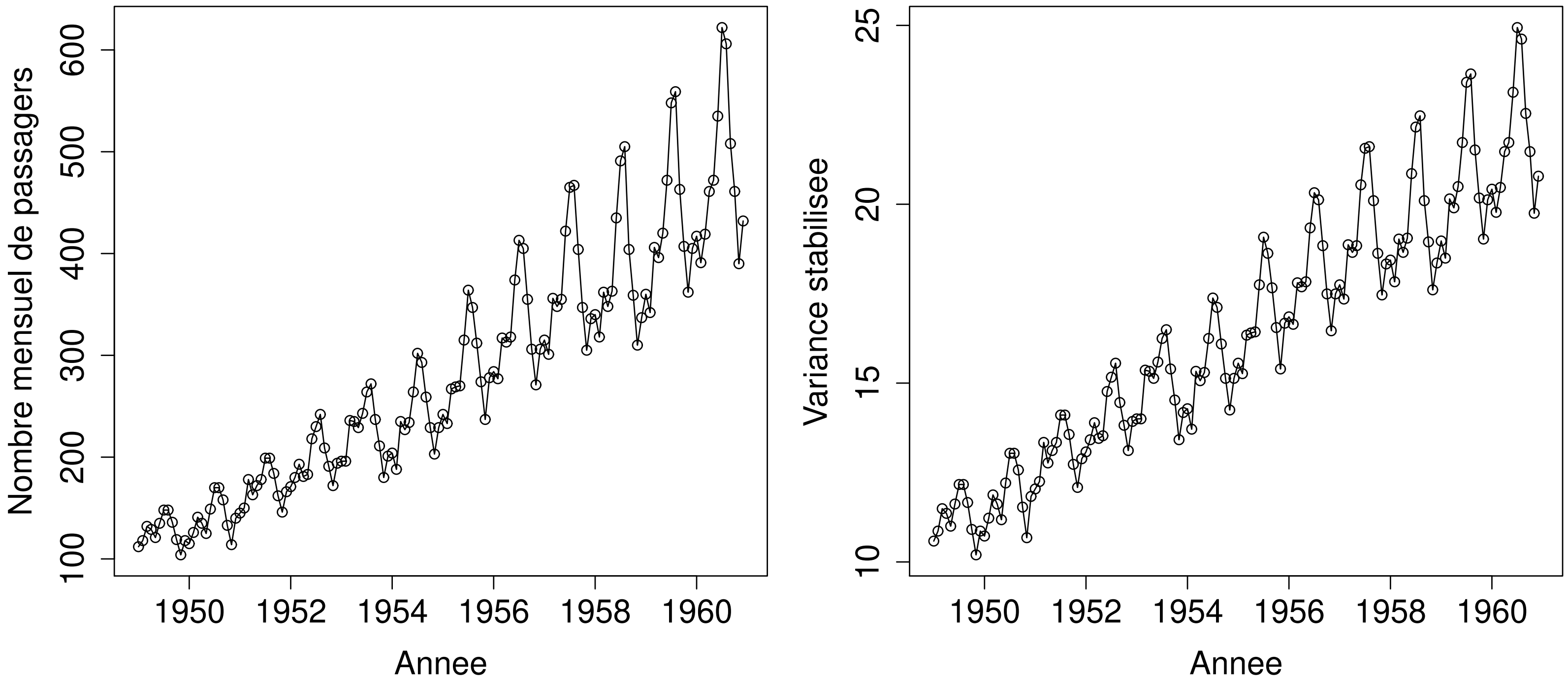


Figure 6: *Attempt to stabilize the variance for the international airline passengers time series. Actually it is not 100% relevant and a log transformation is more relevant—no model are perfect ;-).*

- So far we were able to stabilize the variance (more or less)
- The linear trend is still present but hopefully we know how to remove it using differentiation (on the transformed series)

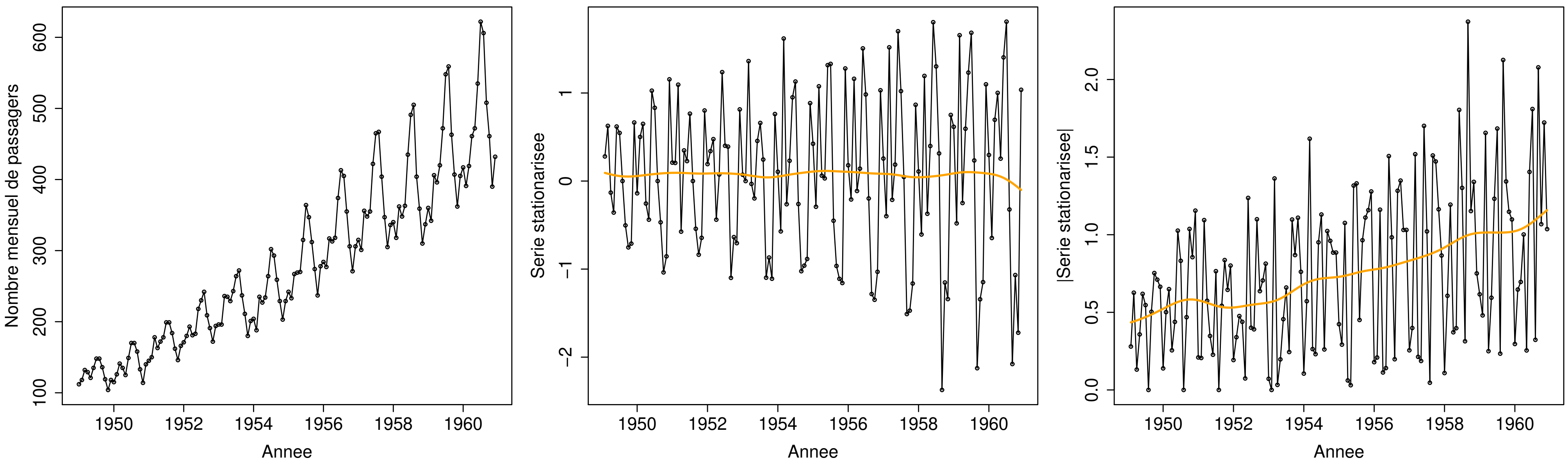


Figure 7: *Attempt to stationnarize the international airline passengers time series. Orange curve: scatterplot smoothing (Nadaraya–Watson to be taught in another course).*

No free lunch...

☞ Beware often differentiated / transformed a time series yields to more complex dependence structures.

- Whenever possible, we will try to work on the original time series rather than on its transformed version so that
  - simpler models will be used;
  - forecasting and interpretation will be easier.

2. Classical models

White noise a.k.a. nothing to model

Definition 13. A second order time series $\{X_t : t \in \mathbb{Z}\}$ is a white noise if it satisfies

$$\mu(t) = 0, \quad t \in \mathbb{Z}, \qquad \gamma(h) = \begin{cases} \sigma^2, & h = 0, \\ 0, & h \neq 0. \end{cases}$$

We will refer to Gaussian white noise if we further have $X_t \sim N(0, \sigma^2)$.

Definition 13. A second order time series $\{X_t \colon t \in \mathbb{Z}\}$ is a white noise if it satisfies

$$\mu(t) = 0, \quad t \in \mathbb{Z}, \qquad \gamma(h) = \begin{cases} \sigma^2, & h = 0, \\ 0, & h \neq 0. \end{cases}$$

We will refer to Gaussian white noise if we further have $X_t \sim N(0, \sigma^2)$.

☞ As stated in the title, from a modelling point of view such time series is irrelevant since

$$\Pr(X_{t+1} \leq x_{t+1} \mid X_t, \ldots, X_1) = \Pr(X_{t+1} \leq x_{t+1}).$$

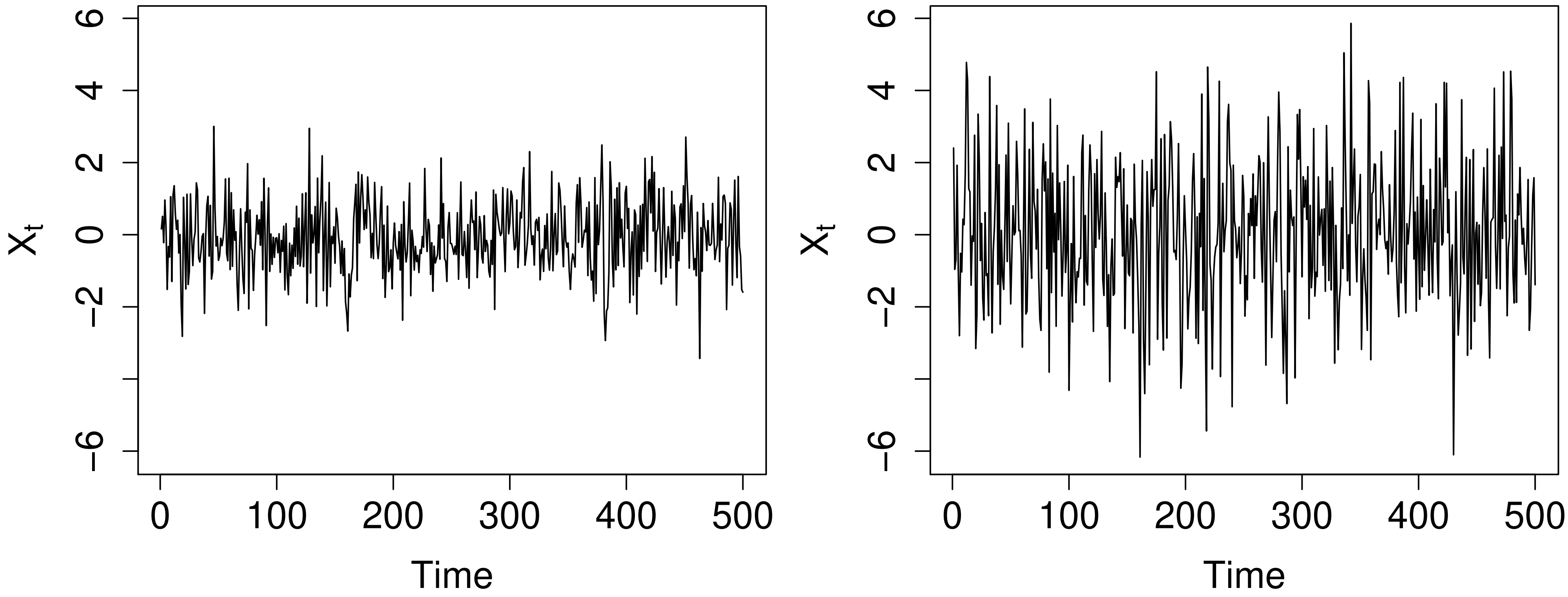


Figure 8: *Two Gaussian white noise time series with* $\sigma^2 = 1$ *and* 4.

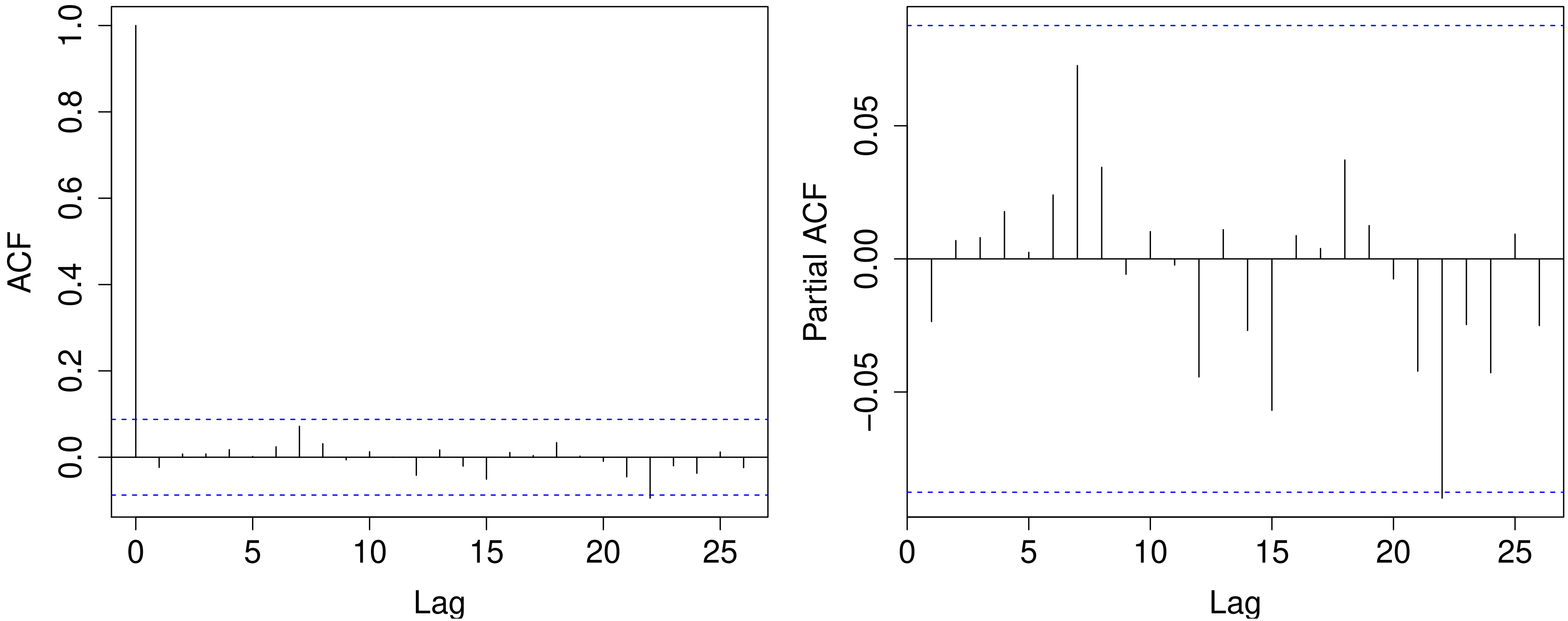


Figure 9: *ACF and PACF of a white noise. Dashed lines correspond to pointwise 95% confidence intervals for a white noise, i.e.,* $\pm 1.96/sqrtn$. *It is useful to detect any departure from white noise.*

Hypothesis testing for white noise

$$H_0\colon \{X_t\colon t \geq 0\} \text{ is white noise} \quad \text{vs} \quad \{X_t\colon t \geq 0\} \text{ is not}$$

- The original test was proposed by Box and Pierce (JASA, 1970) and was later refined by Ljung and Box (Biometrika, 1978) and is a consequence of the following statement.
- Under the null, and for n large enough and $m \ll n$,

$$Q_m = n(n+2) \sum_{h=1}^{m} \frac{\hat{\rho}(h)^2}{n-h} \dot{\sim} \chi^2_m.$$

Sketch. Asymptotic normality of $\hat{\rho}$, normalization, sum of χ^2_1 □

Remark. The test lacks of power whenever m is too large or too small!
In practice we plot the evolution of the p–values for various m values and look for a pattern above or below a threshold, i.e., 5%.

Illustration

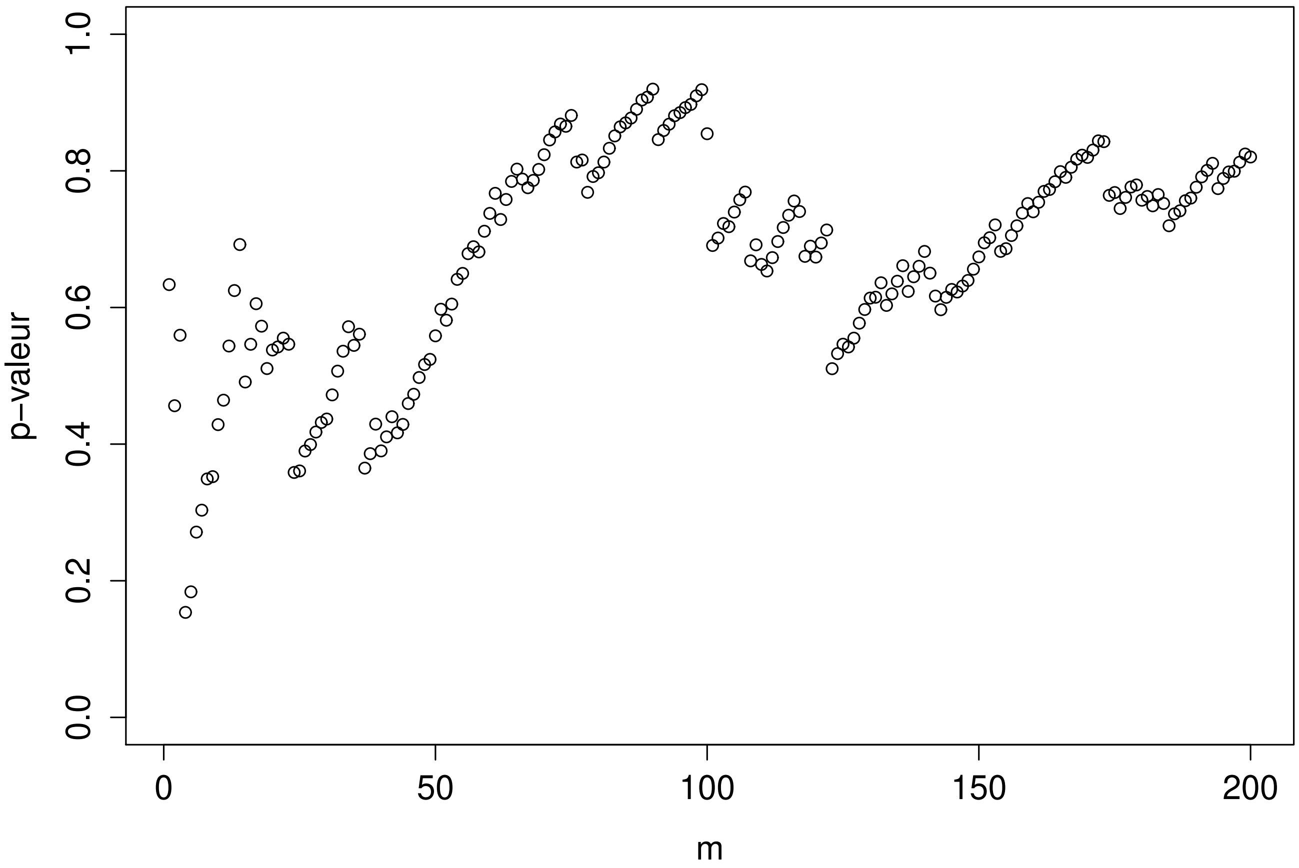


Figure 10: *Evolution of the p-values for the Ljung–Box test. What can we say?*

Keep in mind

- Throughout this lecture, we will assume that the times series are centered, i.e., $\mu(t) = 0$.
- In practice we will have $\mathbb{E}(X_t) = \mu$ for some unknown μ.
- For such cases we just have to use the to be presented models on the centered time series $\{Y_t = X_t - \mu \colon t \geq 0\}$.

Definition 14. The auto-regressive model of order p is given by

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t,$$

where $\{\varepsilon_t : t \in T\}$ is a white noise and $\phi_1, \ldots, \phi_p$, $\phi_p \neq 0$, are parameters of the model to be estimated from data.

Definition 15. The auto-regressive operator of an $AR(p)$ is

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p.$$

Definition 14. The auto-regressive model of order p is given by

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t,$$

where $\{\varepsilon_t : t \in T\}$ is a white noise and $\phi_1, \ldots, \phi_p$, $\phi_p \neq 0$, are parameters of the model to be estimated from data.

Definition 15. The auto-regressive operator of an $AR(p)$ is

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p.$$

☞ We can thus write an $AR(p)$ in a more compact way

$$\phi(B) X_t = \varepsilon_t.$$

ACF of an $AR(p)$

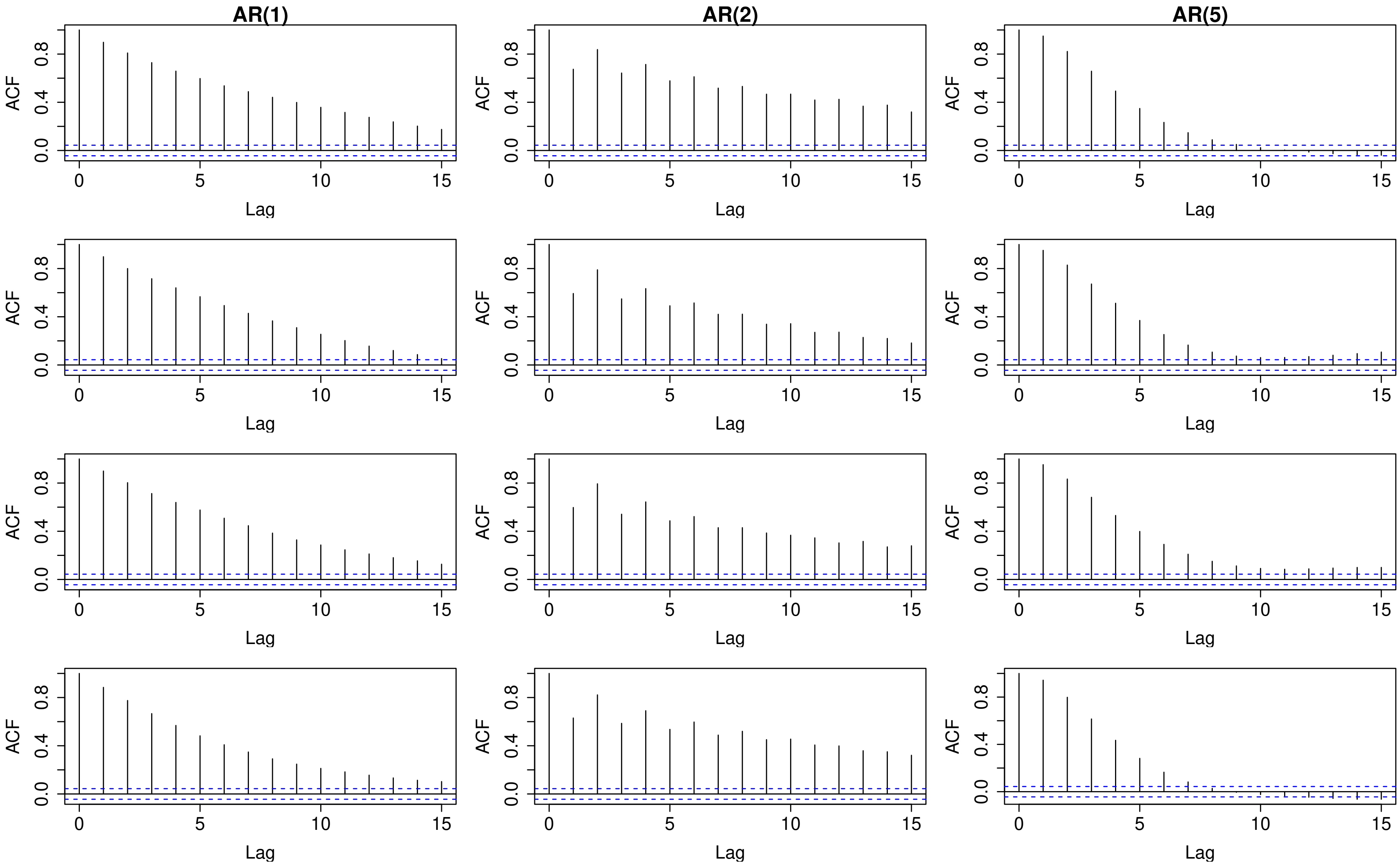


Figure 11: *ACF of 4 independent realizations (each row) of an $AR(p)$ with, from left to right, $p = 1, 2, 5$.*

PACF of an $AR(p)$

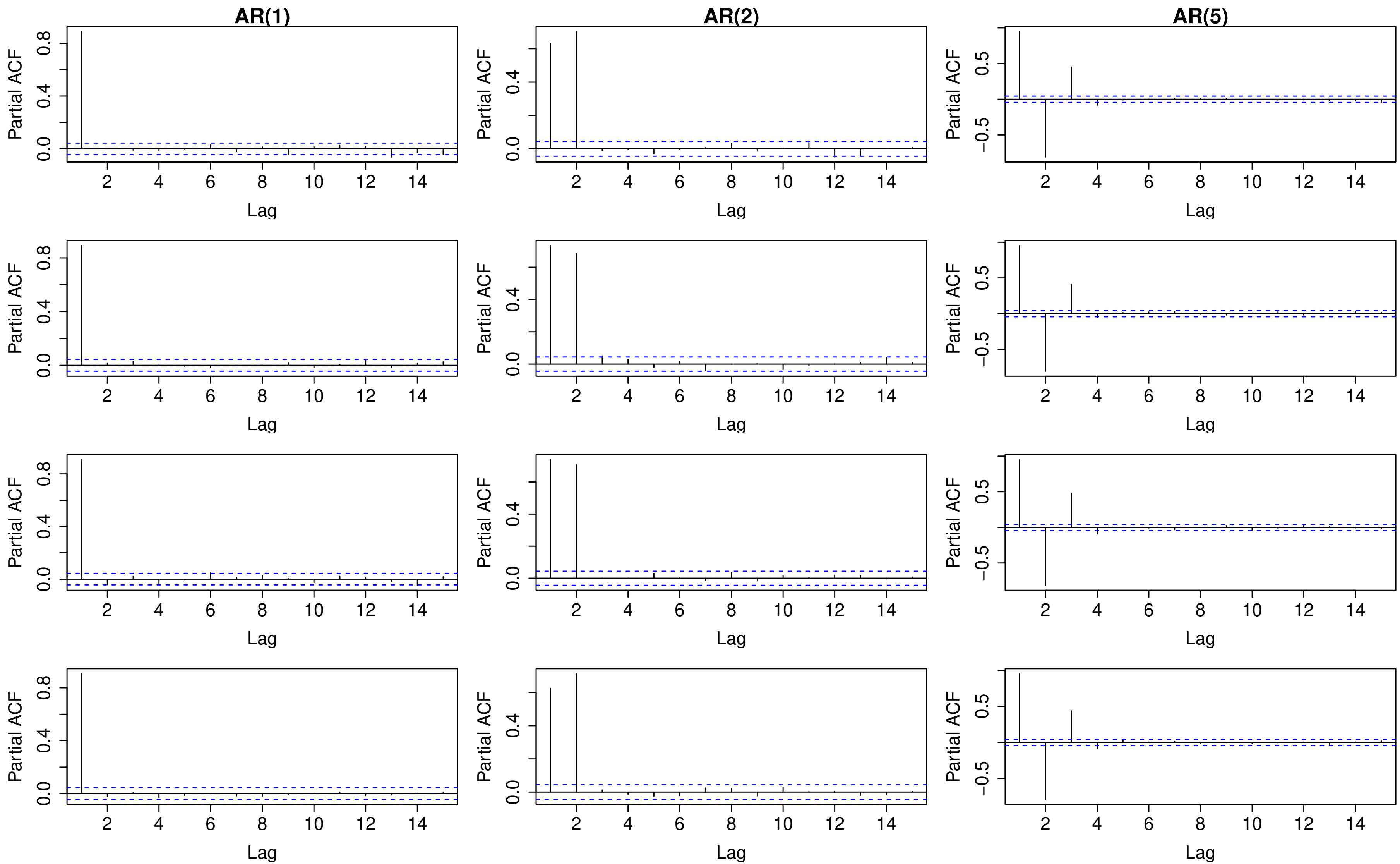


Figure 12: *PACF of 4 independent realizations (each row) of an $AR(p)$ with, from left to right, $p = 1, 2, 5$.*

Definition 16. The moving average model of order q is given by

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

where $\{\varepsilon_t \colon t \in T\}$ is a white noise and $\theta_1, \ldots, \theta_q$, $\theta_q \neq 0$, are unknown parameters to be estimated from data.

Definition 17. The moving average operator of an $MA(q)$ is

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q.$$

Definition 16. The moving average model of order q is given by

$$X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q},$$

where $\{\varepsilon_t \colon t \in T\}$ is a white noise and $\theta_1, \ldots, \theta_q$, $\theta_q \neq 0$, are unknown parameters to be estimated from data.

Definition 17. The moving average operator of an $MA(q)$ is

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q.$$

☞ We can thus write in a compact way any $MA(q)$ as

$$X_t = \Theta(B)\varepsilon_t.$$

ACF of an $MA(q)$

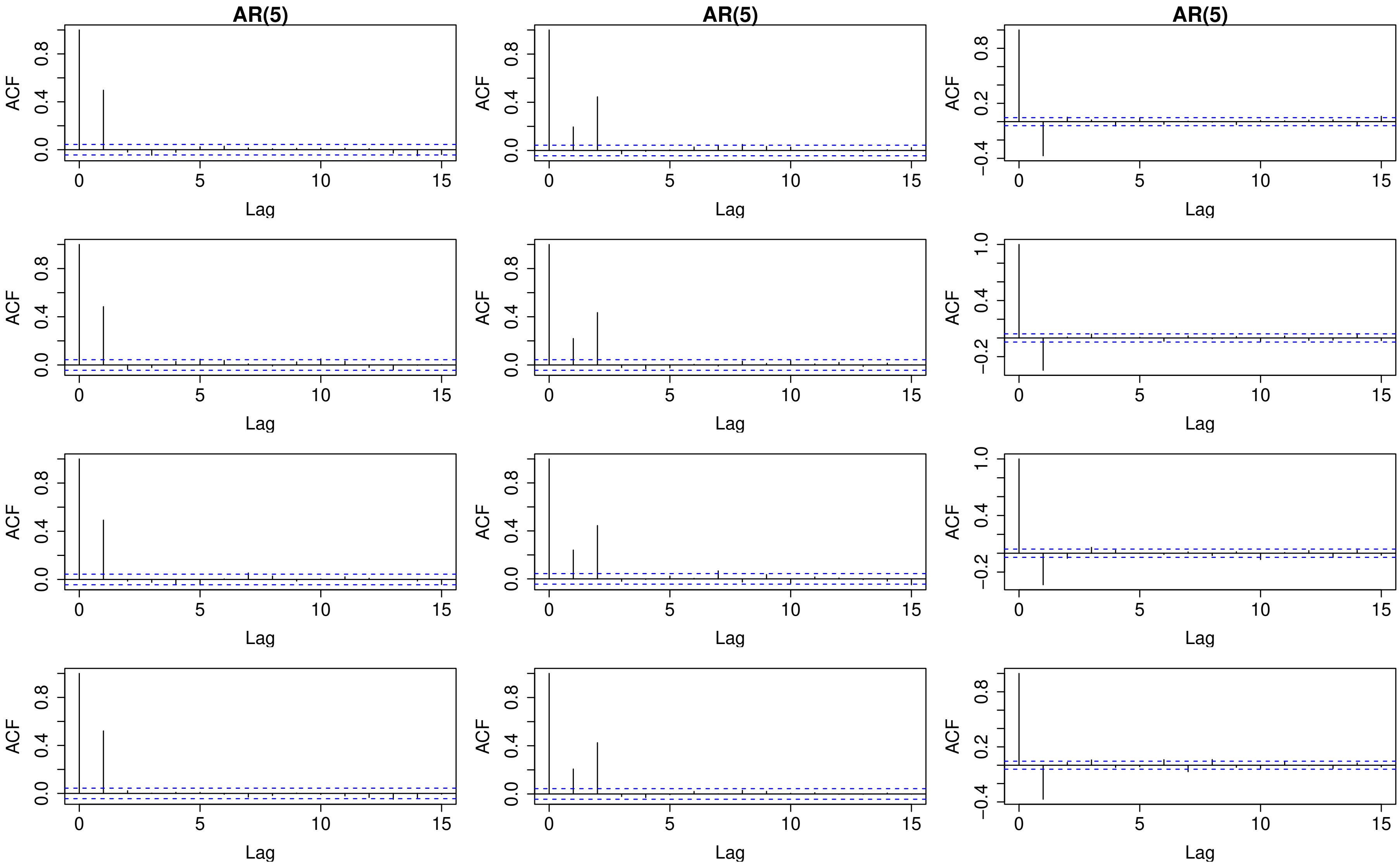


Figure 13: *ACF of 4 independent copies of a $MA(q)$ with, from left to right, $q = 1, 2, 5$.*

PACF of a $MA(q)$

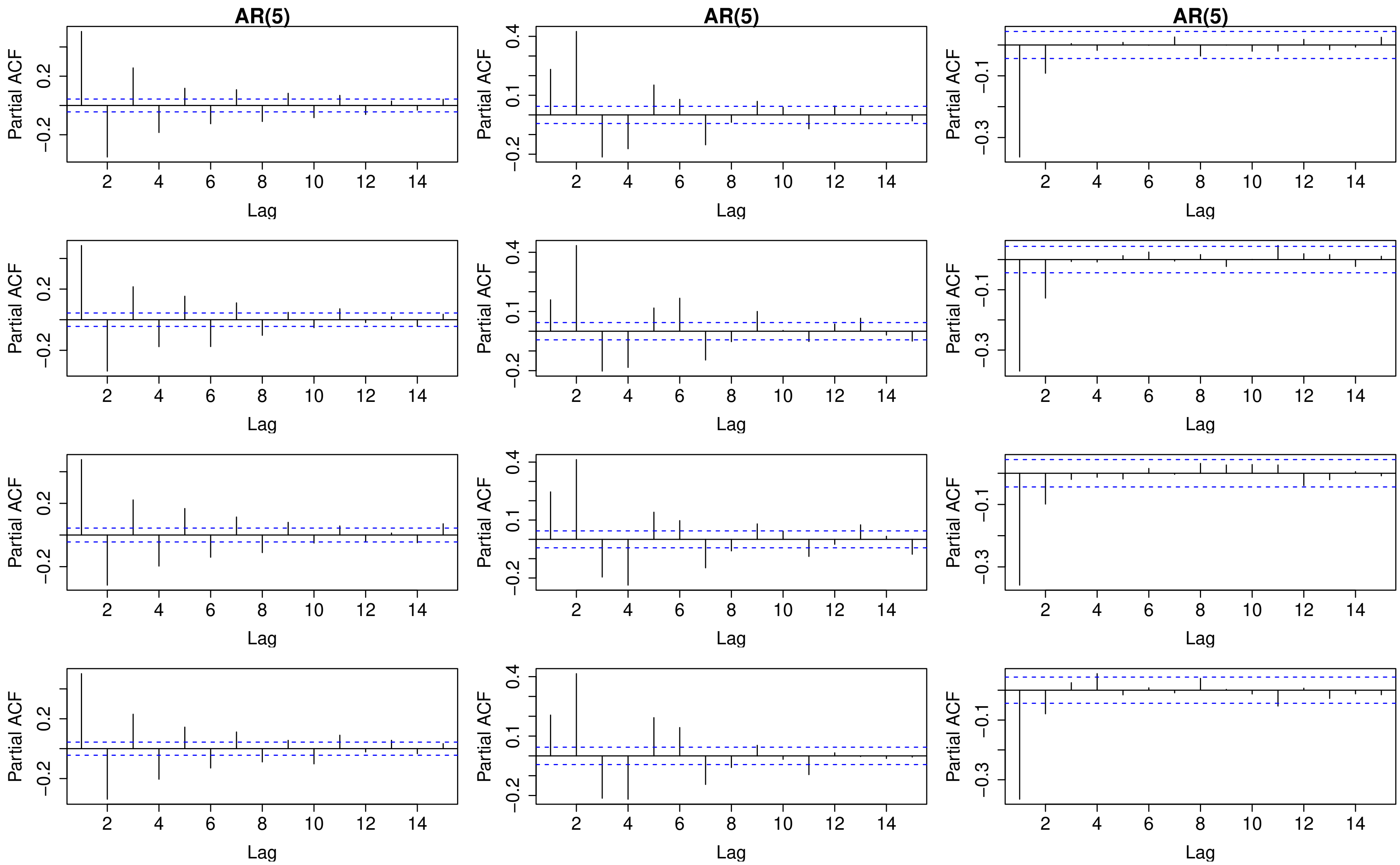


Figure 14: *PACF of 4 independent copies of a $MA(q)$ with, from left to right, $q = 1, 2, 5$.*

Towards $ARMA(p, q)$

- ARMA time series are widely used model for the following statement. . .

Towards $ARMA(p, q)$

- ARMA time series are widely used model for the following statement...

Let γ be any stationary autocovariance function such that $\lim_{\|h\| \to \infty} \gamma(h) \to 0.$ We can always build an $ARMA$ model whose autocovariance function is γ (admitted)

Definition 18. A time series $\{X_t : t \in \mathbb{Z}\}$ is a $ARMA(p,q)$, $q, q \in \mathbb{N}_*$, if it is stationnary and such that

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

or equivalently using our compact notation

$$\phi(B) X_t = \theta(B) \varepsilon_t.$$

Definition 18. A time series $\{X_t : t \in \mathbb{Z}\}$ is a $ARMA(p,q)$, $q, q \in \mathbb{N}_*$, if it is stationnary and such that

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

or equivalently using our compact notation

$$\phi(B) X_t = \theta(B) \varepsilon_t.$$

☞ Beware to the apparently complex model

$$\eta(B)\phi(B) X_t = \eta(B)\theta(B)\varepsilon_t.$$

which, after simplification by $\eta(B)$, leads to a simpler $ARMA$ model.

Illustration

- Consided the following $ARMA(1,1)$ model

$$X_t = 0.5X_{t-1} - 0.5\varepsilon_{t-1} + \varepsilon_t.$$

Illustration

- Consided the following $ARMA(1,1)$ model

$$X_t = 0.5X_{t-1} - 0.5\varepsilon_{t-1} + \varepsilon_t.$$

- One can show that

$$X_t - 0.5X_{t-1} = \varepsilon_t - 0.5\varepsilon_{t-1} \Longleftrightarrow \eta(B)X_t = \eta(B)\varepsilon_t,$$

  with $\eta(B) = 1 - 0.5B$.
- The above "ARMA" model is actually a white noise $X_t = \varepsilon_t$!!!

Illustration

- Consided the following $ARMA(1,1)$ model

$$X_t = 0.5X_{t-1} - 0.5\varepsilon_{t-1} + \varepsilon_t.$$

- One can show that

$$X_t - 0.5X_{t-1} = \varepsilon_t - 0.5\varepsilon_{t-1} \Longleftrightarrow \eta(B)X_t = \eta(B)\varepsilon_t,$$

  with $\eta(B) = 1 - 0.5B$.
- The above "ARMA" model is actually a white noise $X_t = \varepsilon_t$!!!

☞ We should check if no shared roots between polynomials $\phi(B)$ and $\theta(B)$.

ACF of an $ARMA(p, q)$

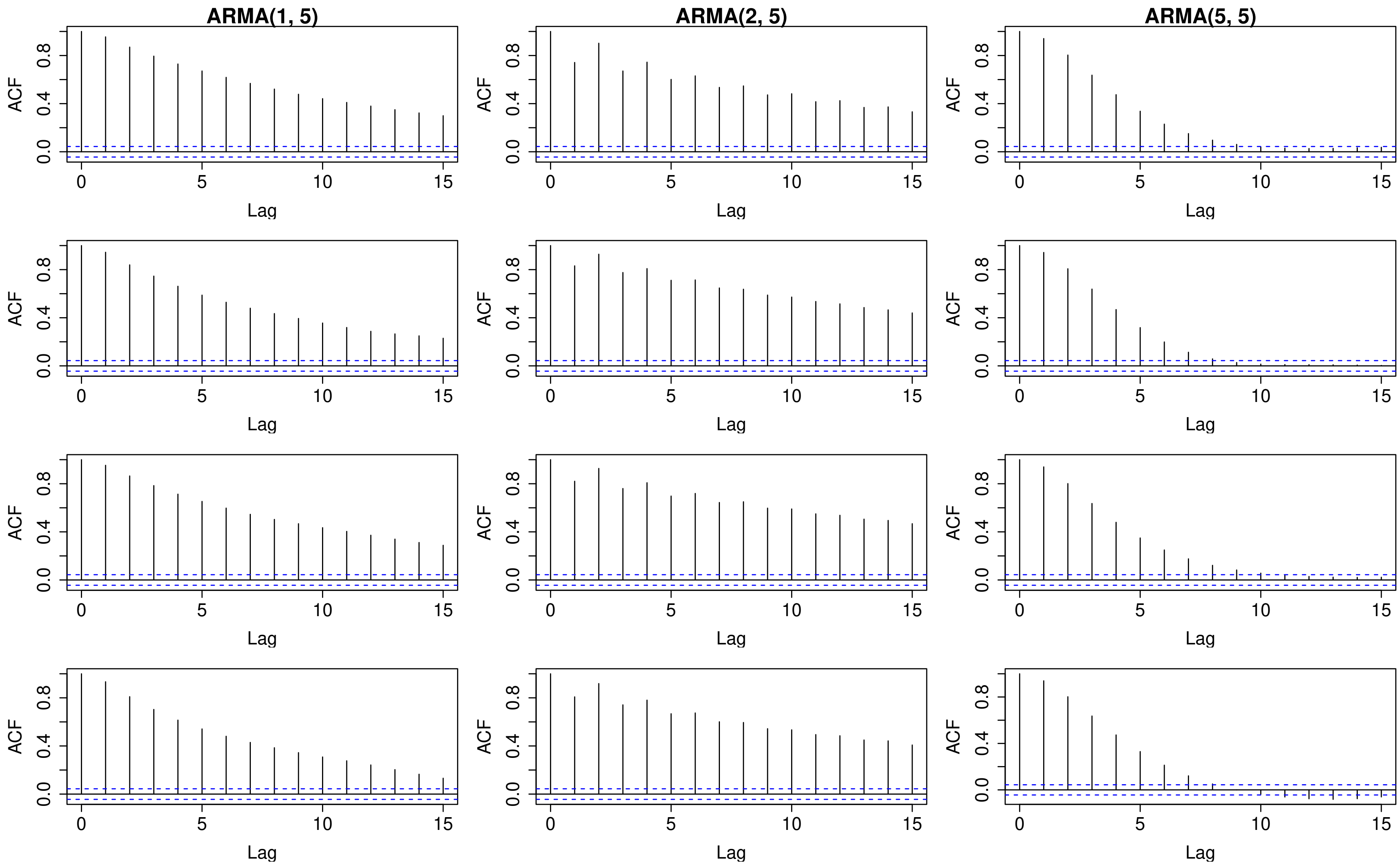


Figure 15: *ACF an $ARMA(p, q)$ with $p, q = 1, 2, 5$.*

PACF of an $ARMA(p, q)$

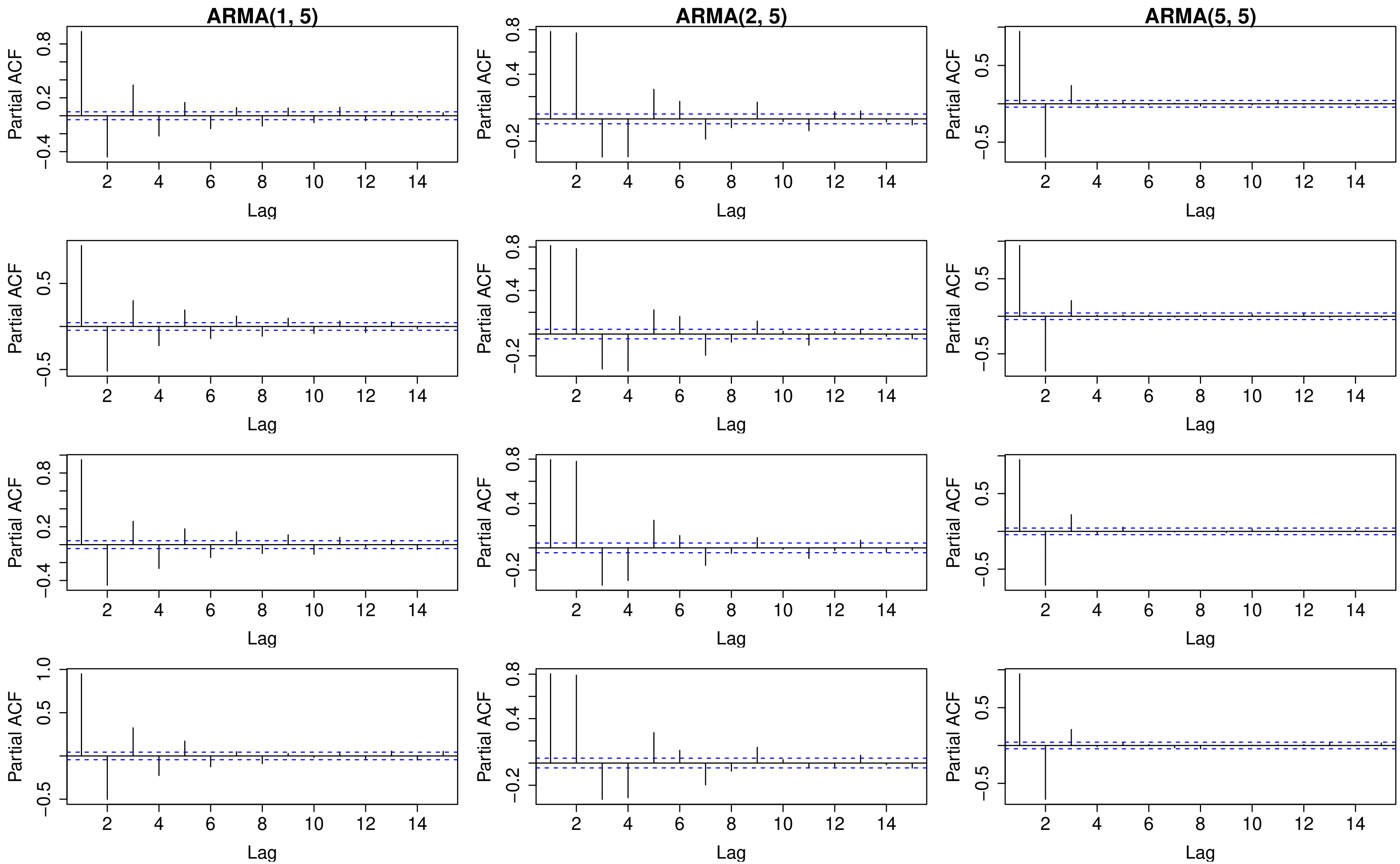


Figure 16: *PACF of an* $ARMA(p, q)$ *with* $p, q = 1, 2, 5$.

Table 1: *Identification of the order of pure $AR(p)$ or pure $MA(q)$ processes.*

	$AR(p)$	$MA(q)$	$ARMA(p,q)$
ACF	$\to 0$	Cutoff at lag q	$\to 0$
PACF	Cutoff at p	$\to 0$	$\to 0$

Definition 19. A time series $ARMA(p, q)$ $\{X_t \colon t \in \mathbb{Z}\}$ is causal if it can be written as

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} = \psi(B)\varepsilon_t,$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ and $\sum_{j=0}^{\infty} |\psi_j| < \infty$, $\psi_0 = 1$.

Definition 19. A time series $ARMA(p,q)$ $\{X_t \colon t \in \mathbb{Z}\}$ is causal if it can be written as

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} = \psi(B)\varepsilon_t,$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ and $\sum_{j=0}^{\infty} |\psi_j| < \infty$, $\psi_0 = 1$.

Proposition 3. *An $ARMA(p,q)$ time series is causal iff $\phi(z) \neq 0$ on $\{z \in \mathbb{C} \colon |z| \leq 1\}$. Stated differently, an $ARMA(p,q)$ is causal iff the roots of the polynomial $\phi(z)$ are outside of the complex unit circle.*
The coefficients of the polynomial ψ can be obtained from the following representation

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, \qquad |z| \leq 1.$$

Why causality?

- Consider the following AR(1) process: $X_{t+1} = \phi X_t + \varepsilon_{t+1}$, $|\phi| > 1$.
- Its causal representation is

$$X_{t+1} = \phi X_t + \varepsilon_{t+1} = \ldots = \sum_{j \geq 0} \phi^j \varepsilon_{t+1-j},$$

  but diverges (from an L_2 point of view).
- However we have

$$X_t = \phi^{-1} X_{t+1} - \varepsilon_{t+1} = \ldots = -\sum_{j \geq 0} \phi^{-j} \varepsilon_{t+j},$$

  which converges.
- The latter representation is useless since it requires to know the futre to forecast the present! Causality is therefore a sensible property.

Definition 20. An $ARMA(p,q)$ $\{X_t : t \in \mathbb{Z}\}$ time series is said inversible if it can be written as

$$\pi(B)X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} = \varepsilon_t,$$

where $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j$ and $\sum_{j=0}^{\infty} |\pi_j| < \infty$, $\pi_0 = 1$.

Definition 20. An $ARMA(p, q)$ $\{X_t \colon t \in \mathbb{Z}\}$ time series is said inversible if it can be written as

$$\pi(B)X_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} = \varepsilon_t,$$

where $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j$ and $\sum_{j=0}^{\infty} |\pi_j| < \infty$, $\pi_0 = 1$.

Proposition 4. *An $ARMA(p, q)$ time series is inversible iff $\theta(z) \neq 0$ on $\{z \in \mathbb{C} \colon |z| \leq 1\}$. Stated differently, an $ARMA(p, q)$ is inversible iff the roots of the polynomial $\theta(z)$ are outside of the complex unit circle.*
The coefficients of the polynomial π can be obtained from the following relation

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, \qquad |z| \leq 1.$$

Why inversibility?

- Consider the two following MA(1) time series:

$$X_t = \varepsilon_t + \frac{1}{5}\varepsilon_{t-1}, \qquad Y_t = \omega_t + 5\omega_{t-1},$$

  where $\varepsilon_t \overset{\text{iid}}{\sim} N(0,25)$ and $\omega_t \overset{\text{iid}}{\sim} N(0,1)$.
- These two time series yield to the same process (check it!)
- If we were able to observe the innovations ε_t and ω_t, we could decide which model is the right one.
- However since the latter are not observed, we should put some constraint to ensure identifiability.
- We thus restrict to inversible processes.

Exercise 2. Only one of these time series are inversible. Which one?

Consider more sophisticated models!

Definition 21. A time series $\{X_t \colon t \in \mathbb{Z}\}$ is an $ARIMA(p, d, q)$, $p, d, q \in \mathbb{N}$, if the differenciated $Y_t = (1 - B)^d X_t$ is an $ARMA(p, q)$, i.e., we have

$$\phi(B)(1 - B)^d X_t = \theta(B)\varepsilon_t.$$

- □ The $ARIMA$ processes extend the modeling to non stationary processes (with polynomial trends).

Example 1. The random walk $X_t = X_{t-1} + \varepsilon_t$ is actually an $ARIMA(0, 1, 0)$

Not enough????

Not enough????Even more sophisticated!!!

Definition 22. A time series $\{X_t : t \in \mathbb{Z}\}$ is an $ARIMA(p, d, q) \times (P, D, Q)_s$, $p, d, q, P, D, Q, s \in \mathbb{N}$, if it can be written as follows

$$\Phi(B^s)\phi(B)(1 - B)^d(1 - B^s)^D X_t = \Theta(B^s)\theta(B)\varepsilon_t.$$

We will say that the above time series is a (multiplicative) $SARIMA$, S refers to Seasonal $ARIMA$.

- $SARIMA$ times series allow to model non stationary time series having trends and seasonality.

Example

$$(1 - 0.8B^{12})X_t = (1 - 0.9B)\varepsilon_t,$$

Example

$$(1 - 0.8B^{12})X_t = (1 - 0.9B)\varepsilon_t, \qquad ARIMA(0,0,1) \times (1,0,0)_{12}.$$

Example

$$(1 - 0.8B^{12})X_t = (1 - 0.9B)\varepsilon_t, \qquad ARIMA(0,0,1) \times (1,0,0)_{12}.$$

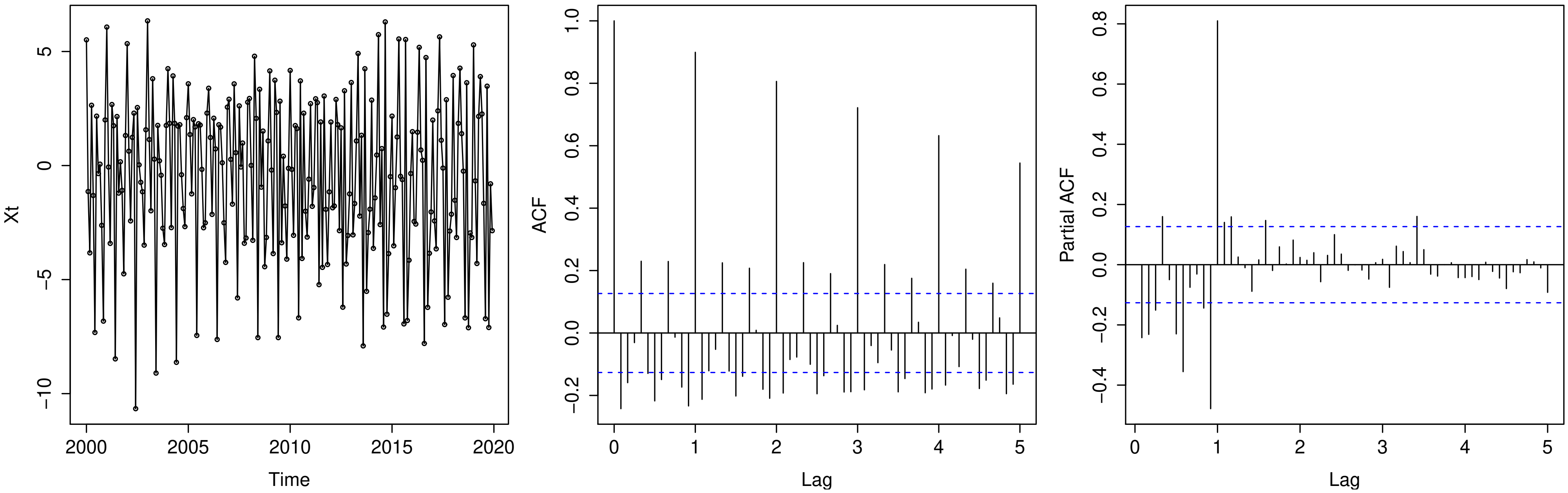


Figure 17: *ACF and PACF of the above $SARIMA$.*

Example

$$(1-0.8B^{12})X_t = (1-0.9B)\varepsilon_t, \qquad ARIMA(0,0,1)\times(1,0,0)_{12}.$$

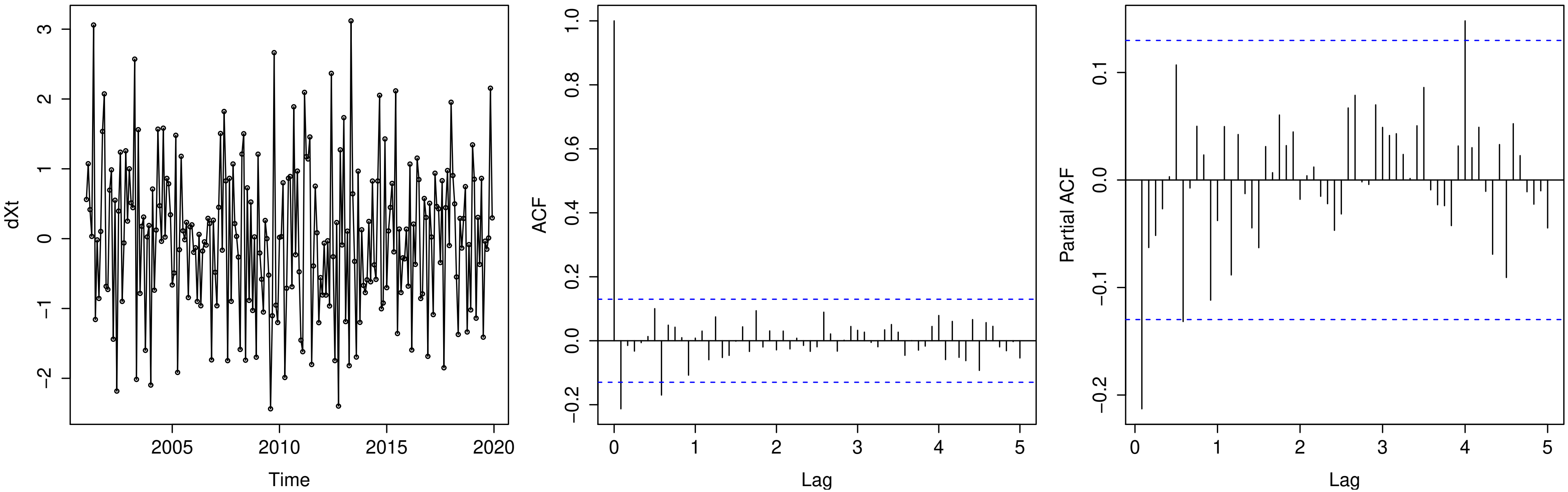


Figure 17: *ACF and PACF for the time series* $(1-B^{12})X_t$.

3. Spectral analysis

Idea

- Spectral analysis decomposes the time series into sinusoidal functions with (uncorrelated) random coefficients.
- It is a Fourier decomposition on random functions
- Such an approach is referred to spectral analysis to be opposed to l'temporal analysis so we have seen so far.
- The two approaches differ since:
  - temporal analysis is nothing but a regression on past values;
  - spectral analysis is nothing but a regression on sinusoidal functions.

Periodic time series

- Consider the following periodic time series

$$X_t = A\cos(2\pi\omega t) + B\sin(2\pi\omega t),$$

where A, B are uncorrelated centered and scaled random variables.

- It can be easily shown that

$$X_t = C\sin(2\pi\omega t + \phi), \qquad C = \sqrt{A^2 + B^2}, \quad \tan\phi = \frac{B}{A}.$$

- Hence

$$\mu(t) = \mathbb{E}[X_t] = 0, \qquad \gamma(t, t+h) = \cos(2\pi\omega h),$$

and the time series $\{X_t \colon t \in \mathbb{N}\}$ is stationary.

Autocovariance function of a periodic time series

Exercise 3. Consider the following periodic time series

$$X_t = \sum_{j=1}^{k} \left\{ A_j \cos(2\pi\omega_j t) + B_j \sin(2\pi\omega_j t) \right\},$$

where A_j, B_j are uncorrelated random variables with mean 0 and variance σ_j^2.
Compute its autocovariance function?

- We just show that
$$\gamma(h) = \sum_{j=1}^{k} \sigma_j^2 \cos(2\pi\omega_j h).$$
- In other words, the autocovariance function can be decomposed into a Fourrier series whose coefficients are the variances of the sinusoidal components.
- The spectral density is a continuous version of the above decomposition, i.e., a stochastic Fourrier transform[1]

[1] More rigorously we should rely on stochastic integrals which is far beyond the scope of this course.

Spectral density

Definition 23 (Spectral density)**.** The spectral density f of a stationary time series $\{X_t : t \in \mathbb{N}\}$ whose autocovariance function is γ and such that $\sum_{h\geq 0} |\gamma(h)| < \infty$ is

$$f(\omega) = \sum_{h\in\mathbb{Z}} \gamma(h) e^{-2i\pi\omega h}, \qquad \omega \in \mathbb{R}.$$

- □ It is indeed a density (not a proba one though) since for all $\omega \in \mathbb{R}$,

$$|f(\omega)| \leq \sum_{h\in\mathbb{Z}} |\gamma(h) e^{-2i\pi\omega h}| = \sum_{h\in\mathbb{Z}} |\gamma(h)| < \infty$$

- □ Further it is periodic as a consequence of $\omega \mapsto \exp(-2i\pi\omega h)$ being periodic with period 1.

Definition 23 (Spectral density)**.** The spectral density f of a stationary time series $\{X_t : t \in \mathbb{N}\}$ whose autocovariance function is γ and such that $\sum_{h \geq 0} |\gamma(h)| < \infty$ is

$$f(\omega) = \sum_{h \in \mathbb{Z}} \gamma(h) e^{-2i\pi\omega h}, \qquad \omega \in \mathbb{R}.$$

- ☐ It is indeed a density (not a proba one though) since for all $\omega \in \mathbb{R}$,

$$|f(\omega)| \leq \sum_{h \in \mathbb{Z}} |\gamma(h) e^{-2i\pi\omega h}| = \sum_{h \in \mathbb{Z}} |\gamma(h)| < \infty$$

- ☐ Further it is periodic as a consequence of $\omega \mapsto \exp(-2i\pi\omega h)$ being periodic with period 1.

☞ We can thus restrict our attention to the domain $[-0.5, 0.5]$.

- One can show that:
  - f is even;
  - $f(\omega) \geq 0$ for all $\omega \in \mathbb{R}$ (since γ is positive definite)
  - $\gamma(h) = \int_{-1/2}^{1/2} \exp(2i\pi\omega h) f(\omega) \mathsf{d}\omega$.

Exercise 4. Proof that

$$\gamma(0) = \mathsf{Var}(X_t) = \int_{-1/2}^{1/2} f(\omega) \mathsf{d}\omega.$$

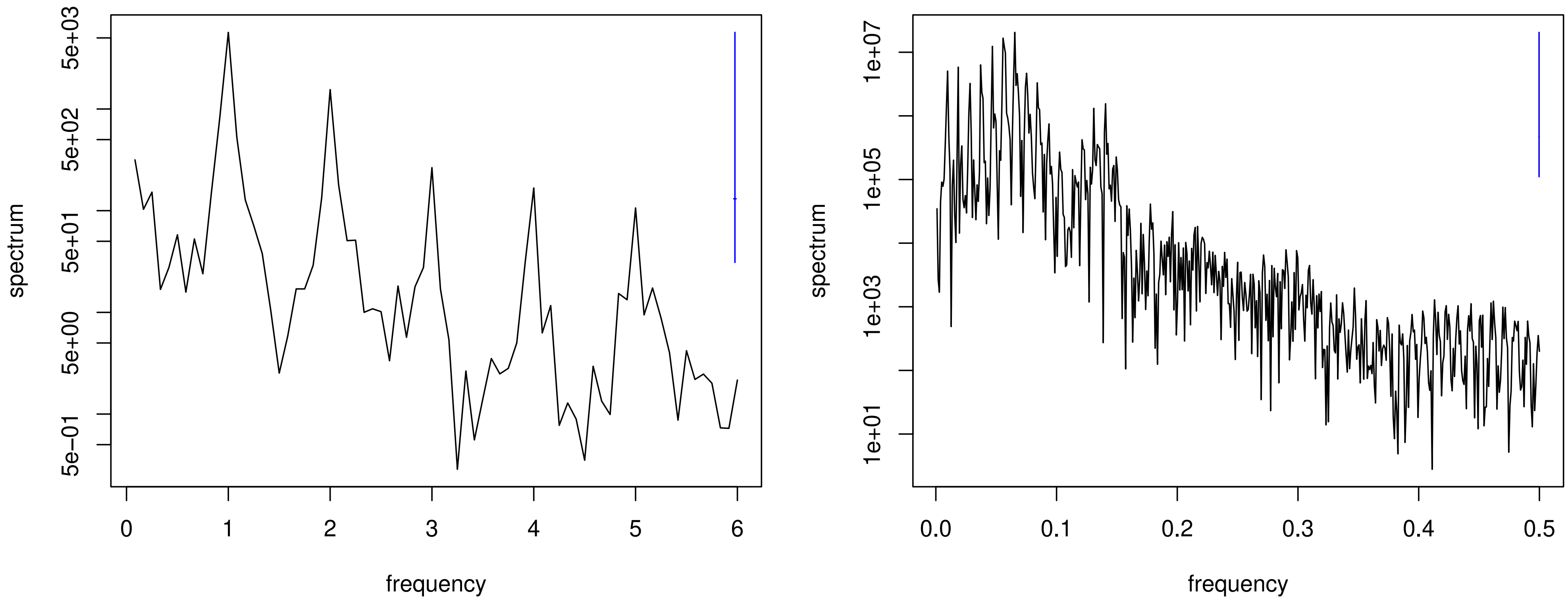


Figure 18: *Spectral density estimate for the international airline passengers (left) and "aaaaahhhh" (right) data.*

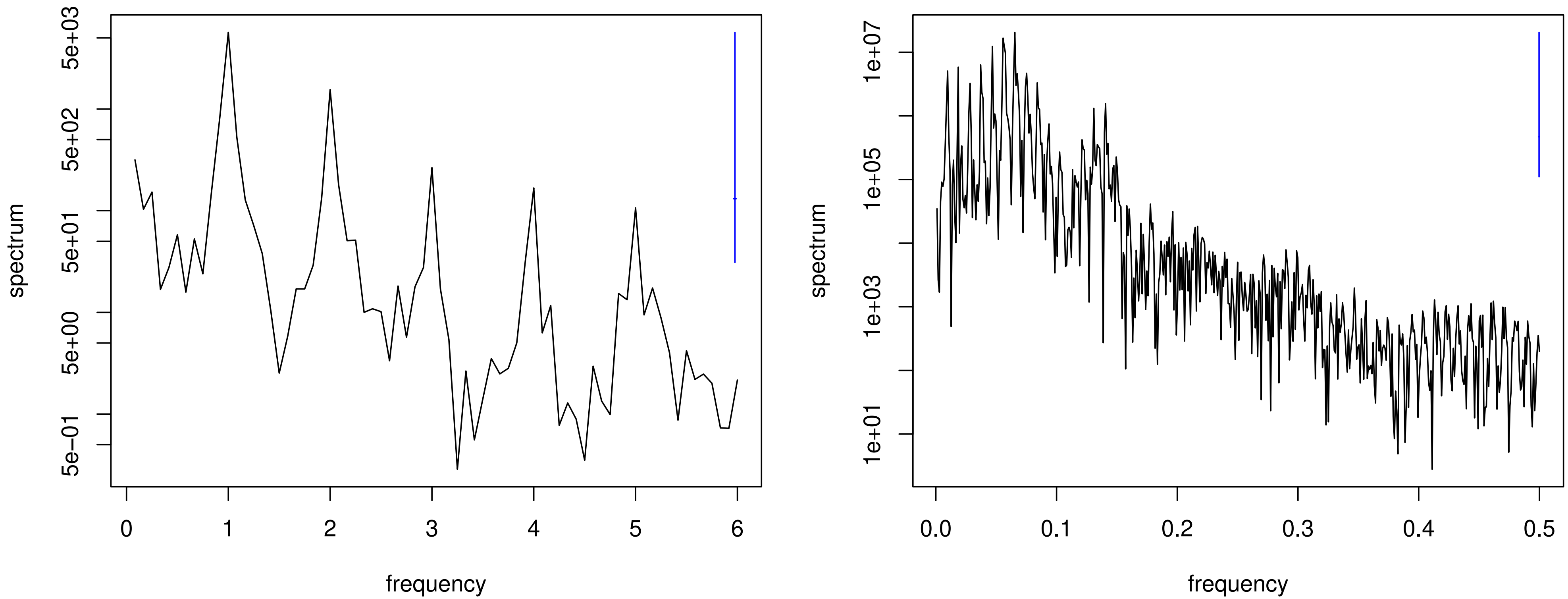


Figure 18: *Spectral density estimate for the international airline passengers (left) and "aaaaahhhh" (right) data.*

- Annual seasonality for the airline data set is clearly visible (as well as harmonics);
- For the "aaaaahhhh" time series, one can note a much smaller frequency.

Exercise 5. What is the spectral density of a white noise? What can you conclude?

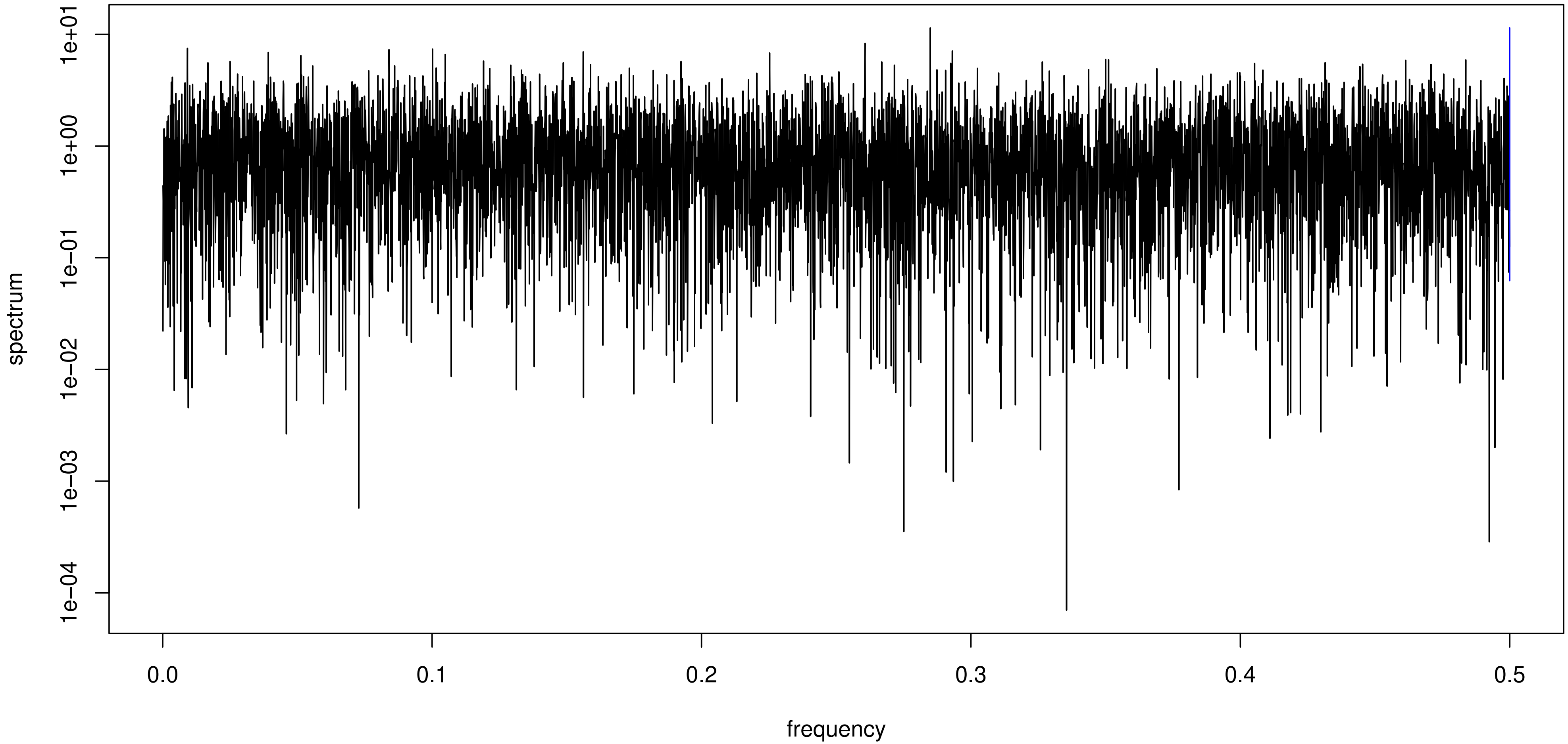


Figure 19: *Spectral density estimate of a white noise.*

A useful property

Proposition 5. *Consider the following time series*

$$Y_t = \sum_{j \in \mathbb{Z}} a_j X_{t-j},$$

where $\{X_t \colon t \in \mathbb{Z}\}$ is a time series with spectral density $f_X(\omega)$.
Then the spectral density for $\{Y_t \colon t \in \mathbb{Z}\}$ is

$$f_Y(\omega) = |A(\omega)|^2 f_X(\omega), \qquad A(\omega) = \sum_{j \in \mathbb{Z}} a_j \exp(-2i\pi\omega j).$$

Proof. Hint: What is the autocovariance function for $\{Y_t \colon t \in \mathbb{Z}\}$? □

Application

Exercise 6. Compute the spectral density of the ARMA process

$$\Phi(B)X_t = \Theta(B)\varepsilon_t.$$

4. Fitting

Maximum likelihood estimator

- There are plenty of methods to fit an $ARMA(p, q)$ process.
- The l'maximum likelihood estimator is a (very) popular option
- Why?

invariant 1 if we transform the data using a one-one mapping $Y = g(X)$, then $L(\theta; x) = L(\theta; y)$;

invariant 2 if we transform the parameters using a one-one mapping $\psi = \psi(\theta)$, then $f_*(x; \psi) = f_*(x; \psi(\theta)) = f(x; \theta)$ so that $L_*(\psi) = L(\theta)$ where $\hat{\psi} = \hat{\theta}$;

Efficient The Cramer–Rao bound is reached asymptotically $\Rightarrow$ Confidence intervals and related hypothesis tests based on the likelihood are asymptotically optimal.

Maximum likelihood estimator (reminder)

- Given a regular parametric statistical model, the maximum likelihood estimator $\hat{\theta}$ satisfies

$$\hat{\theta} \dot{\sim} N\left\{\theta_*, J(\hat{\theta})^{-1}\right\}, \qquad n \text{ grand},$$

where $J(\hat{\theta})$ is the observed Fisher information matrix, i.e., $J(\hat{\theta}) = -\nabla^2 \ell(\hat{\theta})$.

Maximum likelihood estimator (reminder)

- Given a regular parametric statistical model, the maximum likelihood estimator $\hat{\theta}$ satisfies

$$\hat{\theta} \dot{\sim} N\left\{\theta_*, J(\hat{\theta})^{-1}\right\}, \qquad n \text{ grand},$$

where $J(\hat{\theta})$ is the observed Fisher information matrix, i.e., $J(\hat{\theta}) = -\nabla^2 \ell(\hat{\theta})$.

- As a consequence, confidence intervals for $\theta_{*,r}$ are easily obtained, e.g., symmetric case,

$$\hat{\theta}_r \pm z_\alpha \sqrt{j_{rr}^{-1}},$$

where $j_{rr}^{(-1)}$ is the r-th diagonal element of the matrix $J(\hat{\theta})^{-1}$.

Definition 24. Consider the two following statistical models $\{f_A(x;\theta)\colon \theta \in \Theta\}$ and $\{f_B(x;\psi)\colon \psi \in \Psi\}$, $\Theta \subseteq \Psi$. We say that f_A is nested in f_B if there exists some values for some element of ψ such that, for all $\theta \in \Theta$, $f_A(x;\theta) = f_B(x;\psi)$.

Example 2. The statistical model $N(\mu, \sigma^2)$ is nested within an $AR(1)$ since the former is actually an $AR(1)$ with $\theta_1 = 0$, or an $AR(0)$.

Proposition 6. *Given two models A and B, A being nested within B, we can check if*

$$H_0\colon \textit{Model } A \textit{ is right} \quad H_1\colon \textit{Model } B \textit{ is correct}$$

using the likelihood ratio test statistics W that, under the null H_0, satisfies

$$W = 2\{\ell_B(\hat{\psi}) - \ell_A(\hat{\theta})\} \dot{\sim} \chi^2_p, \qquad n \textit{ large enough},$$

where $p = dim(\Psi) - dim(\Theta)$.

Case study: Beaver temperature

```
> beav2
  day time  temp activ
1 307  930 36.58     0
2 307  940 36.73     0
3 307  950 36.93     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
```

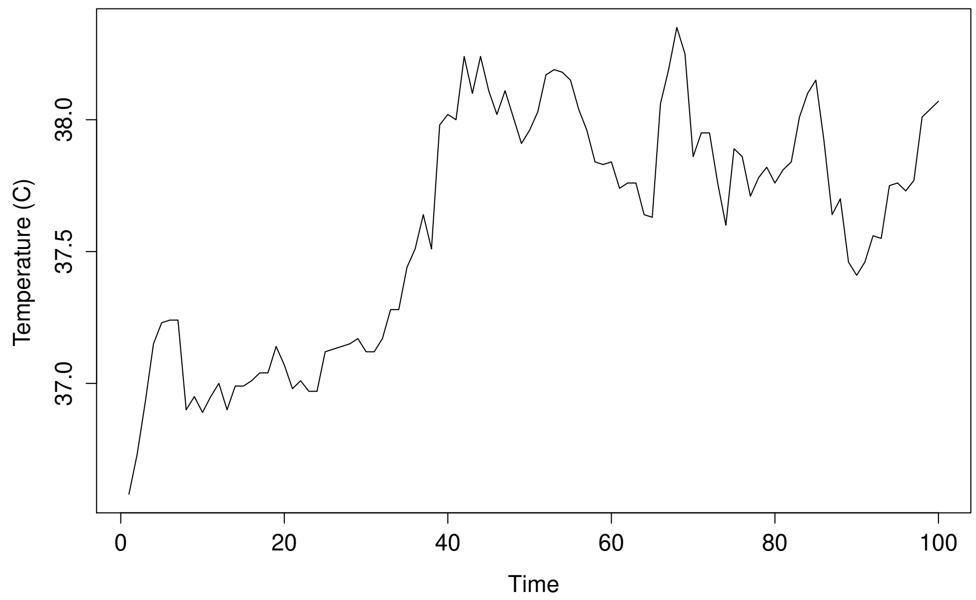


Figure 20: *Time series of the body temperature of a female beaver recorded each 10 minutes— dataset* `beav2` *of the* `MASS` *library.*

Modeling (Thanks Prof. Anthony Davison!!!)

- Model 1: $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$;
- Model 2: $X_1, \ldots, X_\gamma \overset{iid}{\sim} N(\mu, \sigma^2)$ indenpendently of $X_{\gamma+1}, \ldots, X_n \overset{iid}{\sim} N(\mu + \delta, \sigma^2)$, with $\gamma = 38$;
- Model 3: $X_1, \ldots, X_n \sim AR(1)$ with parameters μ, σ^2, ϕ_1 ;
- Model 4: $X_1, \ldots, X_n \sim AR(1)$ with parameters $\mu, \delta, \sigma^2, \phi_1$ and where the expectation is μ for the first 38 observations and $\mu + \delta$ for the others.

Modeling (Thanks Prof. Anthony Davison!!!)

- Model 1: $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$;
- Model 2: $X_1, \ldots, X_\gamma \overset{iid}{\sim} N(\mu, \sigma^2)$ indenpendently of $X_{\gamma+1}, \ldots, X_n \overset{iid}{\sim} N(\mu + \delta, \sigma^2)$, with $\gamma = 38$;
- Model 3: $X_1, \ldots, X_n \sim AR(1)$ with parameters μ, σ^2, ϕ_1 ;
- Model 4: $X_1, \ldots, X_n \sim AR(1)$ with parameters $\mu, \delta, \sigma^2, \phi_1$ and where the expectation is μ for the first 38 observations and $\mu + \delta$ for the others.

Remark. To compare each model performance, one must fit each of them on the same data set. It is problematic for $AR(1)$ models since it requires to know the distribution of Y_0. Several approaches are possible:

- We use the stationary distribution, i.e., $Y_1 \sim N\{\mu, \sigma^2/(1 - \phi_1^2)\}$;
- Imputation, i.e., we use an arbitrary (but sensible) value for Y_0, e.g., $Y_0 = \bar{Y}$;
- We just discard the contribution of Y_1 in the likelihood.

Modeling (Thanks Prof. Anthony Davison!!!)

- Model 1: $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$;
- Model 2: $X_1, \dots, X_\gamma \overset{iid}{\sim} N(\mu, \sigma^2)$ indenpendently of $X_{\gamma+1}, \dots, X_n \overset{iid}{\sim} N(\mu + \delta, \sigma^2)$, with $\gamma = 38$;
- Model 3: $X_1, \dots, X_n \sim AR(1)$ with parameters μ, σ^2, ϕ_1 ;
- Model 4: $X_1, \dots, X_n \sim AR(1)$ with parameters $\mu, \delta, \sigma^2, \phi_1$ and where the expectation is μ for the first 38 observations and $\mu + \delta$ for the others.

Remark. To compare each model performance, one must fit each of them on the same data set. It is problematic for $AR(1)$ models since it requires to know the distribution of Y_0. Several approaches are possible:

- We use the stationary distribution, i.e., $Y_1 \sim N\{\mu, \sigma^2/(1 - \phi_1^2)\}$;
- Imputation, i.e., we use an arbitrary (but sensible) value for Y_0, e.g., $Y_0 = \bar{Y}$;
- We just discard the contribution of Y_1 in the likelihood.

☞ For pedagogical purposes, we will use the 3rd option although the 1st approach is probably the best one.

Model 1

```
## Negative log-likelihood
nllik1 <- function(par, data){
    if (par[2] <= 0)
        return((1 - par[2]) * 1e6)

    -sum(dnorm(data$temp, par[1], par[2], log = TRUE))
}

## Initial values for theta
init1 <- c(mean(beav2$temp), sd(beav2$temp))

## Numerical optimization
fit1 <- nlm(nllik1, init1, hessian = TRUE, data = beav2)

## Estimates and standard errors
rbind(fit1$estimates, sqrt(diag(solve(fit1$hessian))))
```

Model 2

```
## Negative log-likelihood
nllik2 <- function(par, data){
    if (par[2] <= 0)
        return((1 - par[2]) * 1e6)

    -sum(dnorm(data$temp, par[1] + par[3] * data$activ, par[2], log = TRUE))
}

## Initial values for theta
init2 <- c(init1, 1)

## Numerical minimization
fit2 <- nlm(nllik2, init2, hessian = TRUE, data = beav2)

## Estimates and standard errors
rbind(fit2$estimates, sqrt(diag(solve(fit2$hessian))))

## Likelihood ratio test
W21 <- 2 * (fit1$minimum - fit2$minimum)
(p.val <- pchisq(W21, 1, lower.tail = FALSE))
```

Model 3

```
## Negative log-likelihood
nllik3 <- function(par, data){
    if (par[2] <= 0)
        return((1 - par[2]) * 1e6)

    if (abs(par[3]) >= 1)
        return(abs(par[3]) * 1e6)

    ##
    ## Exercise: Write the code
    ##
}

## Initial values for theta
init3 <- c(init1, 0.5)

## Numerical minimization
fit3 <- nlm(nllik3, init3, hessian = TRUE, data = beav2)

## Estimates and standard errors
rbind(fit3$estimates, sqrt(diag(solve(fit3$hessian))))

## Likelihood ratio test
W31 <- 2 * (fit1$minimum - fit3$minimum); (p.val <- pchisq(W31, 1, lower.tail = FALSE))
```

Model 4

```
## Negative log-likelihood
nllik4 <- function(par, data){
    if (par[2] <= 0)
        return((1 - par[2]) * 1e6)

    if (abs(par[3]) >= 1)
        return(abs(par[3]) * 1e6)

    mu <- par[1] + par[4] * data$activ[-100] +
        par[3] * (data$temp[-100] - par[1] - par[4] * data$activ[-100])

    -sum(dnorm(data$temp[-1], mu, par[2], log = TRUE))
}
## Initial values for theta
init4 <- c(init1, 0.5, 0.4)

## Numerical minimization
fit4 <- nlm(nllik4, init4, hessian = TRUE, data = beav2)

## Estimates and standard errors
rbind(fit4$estimates, sqrt(diag(solve(fit4$hessian))))

## Likelihood ratio test
W43 <- 2 * (fit3$minimum - fit4$minimum); (p.val <- pchisq(W43, 1, lower.tail = FALSE))
```

Model fitting summary

Model	# parameters	$\ell(\hat{\theta})$	AIC
1	2	-60.82	125.6
2	3	13.74	-21.5
3	3	61.42	-116.9
4	4	62.39	-116.8

Parameter	Model 1	Model 2	Model 3	Model 4
μ	37.6 (0.04)	37.1 (0.03)	37.8 (0.22)	37.36 (0.19)
σ	0.44 (0.03)	0.21 (0.01)	0.13 (0.01)	0.13 (0.01)
δ	—	0.81 (0.04)	—	0.55 (0.22)
ϕ_1	—	—	0.93 (0.03)	0.86 (0.06)

Model fitting summary

Model	# parameters	$\ell(\hat{\theta})$	AIC
1	2	-60.82	125.6
2	3	13.74	-21.5
3	3	61.42	-116.9
4	4	62.39	-116.8

Parameter	Model 1	Model 2	Model 3	Model 4
μ	37.6 (0.04)	37.1 (0.03)	37.8 (0.22)	37.36 (0.19)
σ	0.44 (0.03)	0.21 (0.01)	0.13 (0.01)	0.13 (0.01)
δ	—	0.81 (0.04)	—	0.55 (0.22)
ϕ_1	—	—	0.93 (0.03)	0.86 (0.06)

Note

- □ the significant increase of the standard errors for μ and δ when $\phi_1 \neq 0$.
- □ the significant decrease of σ when $\delta \neq 0$ or $\phi_1 \neq 0$.

Residual analysis

For Model 3, the (standardized) residuals are given by

$$r_t := \frac{x_t - \hat{x}_t}{\hat{\sigma}} = \frac{x_t - \hat{\mu} - \hat{\phi}_1(x_{t-1} - \hat{\mu})}{\hat{\sigma}}, \qquad t = 2, \dots, n,$$

and, if the model is appropriate, should behave as (Gaussian) white noise.

For Model 3, the (standardized) residuals are given by

$$r_t := \frac{x_t - \hat{x}_t}{\hat{\sigma}} = \frac{x_t - \hat{\mu} - \hat{\phi}_1(x_{t-1} - \hat{\mu})}{\hat{\sigma}}, \qquad t = 2, \ldots, n,$$

and, if the model is appropriate, should behave as (Gaussian) white noise.

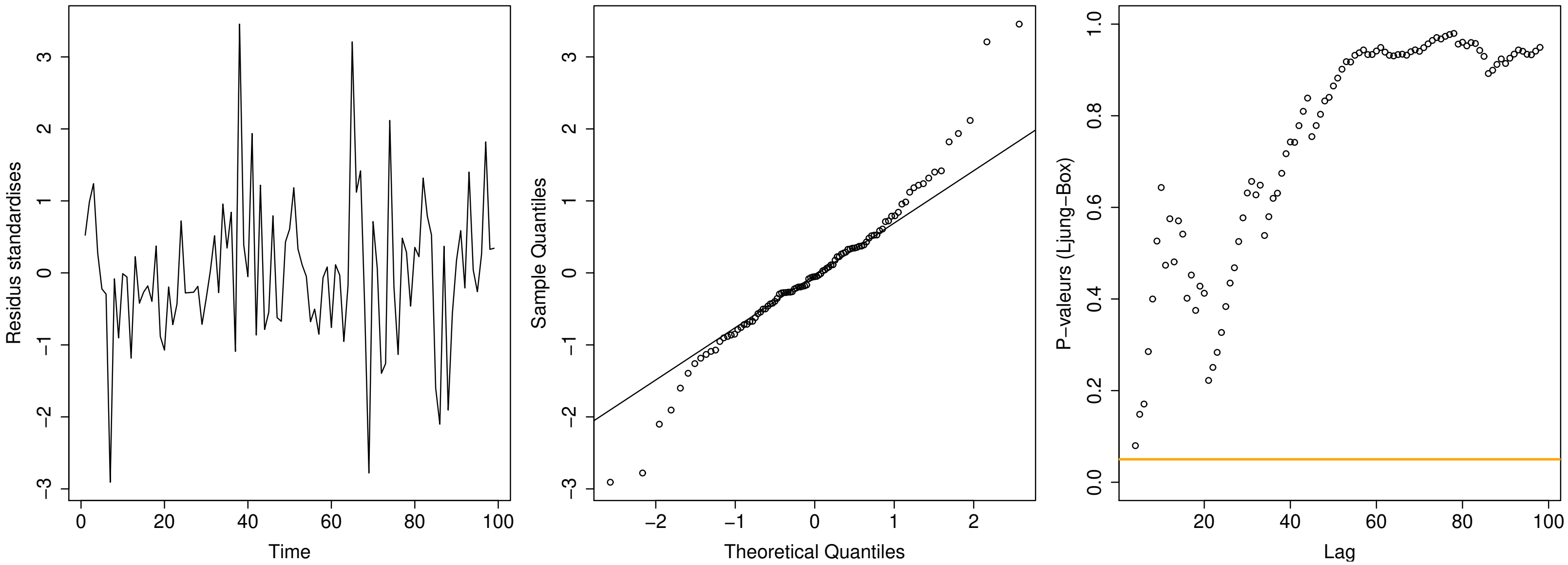


Figure 21: *Residuals analysis. From left to right: time series of residuals; Normal QQ-plot; p-values for the Ljung–Box test for varying values of* m*.*

Residual analysis

For Model 3, the (standardized) residuals are given by

$$r_t := \frac{x_t - \hat{x}_t}{\hat{\sigma}} = \frac{x_t - \hat{\mu} - \hat{\phi}_1(x_{t-1} - \hat{\mu})}{\hat{\sigma}}, \qquad t = 2, \ldots, n,$$

and, if the model is appropriate, should behave as (Gaussian) white noise.

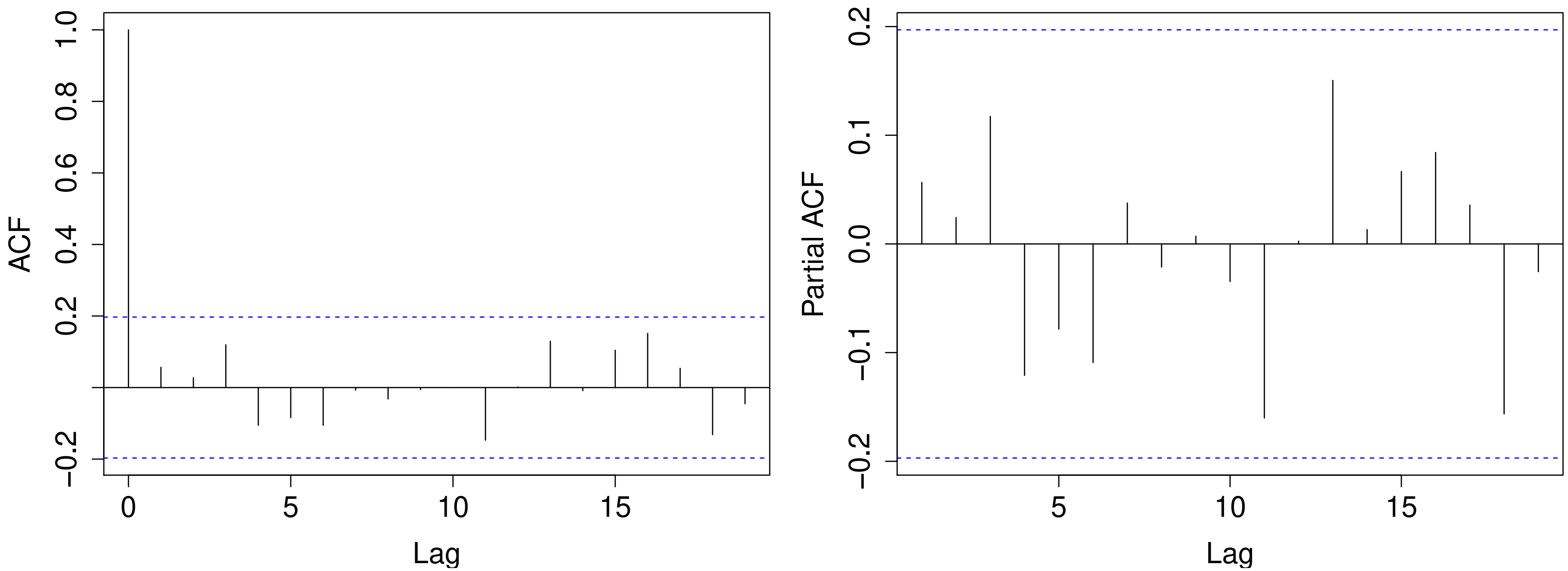


Figure 21: *ACF and PACF of residuals.*

Modelling using a $SARIMA$ is typically a 4 steps procedure (that we may repeat)

- transformation of the time series to stabilize the variance if needed
- identification of the order p, d, q, P, D, Q and s;
- estimation of the parameters ϕ, θ, Φ and Θ;
- model checking of the fitted model.

Modelling using a $SARIMA$ is typically a 4 steps procedure (that we may repeat)

- transformation of the time series to stabilize the variance if needed
- identification of the order p, d, q, P, D, Q and s;
- estimation of the parameters ϕ, θ, Φ and Θ;
- model checking of the fitted model.

Once a relevant model has been fitted, the next step is usually forecasting.

Order identification

Choosing d:

- Graphical inspection of the overall behaviour of the time series (looking for trends and / or seasonality);
- If there is a trend, differenciate the series, typically $d = 1, 2$ and $D = 0, 1$ are enough.

Choosing p and q (and similarly for P and Q but at lags $k \times s$)

- Inspection of ACF and PACF of the differenced series
- Cut off of the ACF at lag q suggests a $MA(q)$;
- Cut off of the PACF at lag p suggests an $AR(p)$;
- No cut off of both ACF/PACF suggests an $ARMA$, typically with $p, q \leq 2$.
- Very slow decreasing of both ACF / PACF suggests to differentiate a bit more the time series.

Order identification

Choosing d:

- Graphical inspection of the overall behaviour of the time series (looking for trends and / or seasonality);
- If there is a trend, differenciate the series, typically $d = 1, 2$ and $D = 0, 1$ are enough.

Choosing p and q (and similarly for P and Q but at lags $k \times s$)

- Inspection of ACF and PACF of the differenced series
- Cut off of the ACF at lag q suggests a $MA(q)$;
- Cut off of the PACF at lag p suggests an $AR(p)$;
- No cut off of both ACF/PACF suggests an $ARMA$, typically with $p, q \leq 2$.
- Very slow decreasing of both ACF / PACF suggests to differentiate a bit more the time series.

☞ In any case, we should opt for parcimonious models, i.e., Occam's razor.

Estimation

- Fitting a SARIMA model is easily done using `R` and the function `arima` or, even better, `sarima` of the package `astsa`.

```
> arima(lh, c(1, 0, 1))

Coefficients:
         ar1     ma1  intercept
      0.4522  0.1982     2.4101
s.e.  0.1769  0.1705     0.1358
```

```
> sarima(lh, 1, 0, 1)

Coefficients:
         ar1     ma1   xmean
      0.4522  0.1982  2.4101
s.e.  0.1769  0.1705  0.1358
```

Model selection / goodness of fit

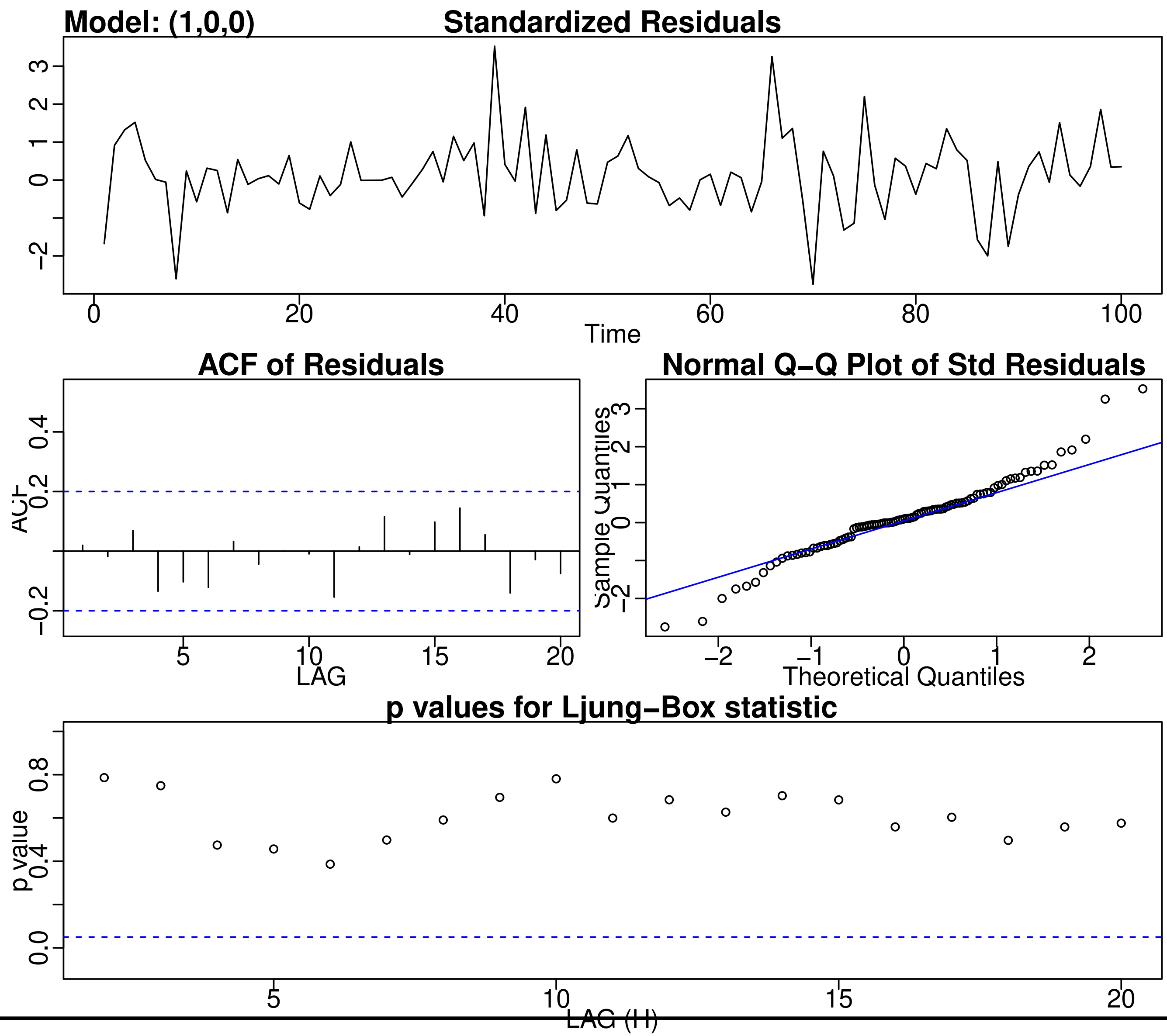
- The best model will be identified using AIC or, whenever possible, using likelihood ratio test.
- Once the "best" model is obtained, we should analyze residuals which, for an $ARMA(p,q)$, are given by

$$r_t = (x_t - \hat{\mu}) - \sum_{j=1}^{p} \hat{\phi}_j (x_{t-j} - \hat{\mu}) - \sum_{j=1}^{q} \hat{\theta}_j r_{t-j},$$

  where $r_1 = \cdots r_p = 0$.
- If the residual analysis does not show any problem, we can proceed to forecasting.

Residual analysis for the beaver temperature time series

```
> sarima(beaver2$temp,
         1, 0, 0)
```



5. Forecasting

Framework

- Most often the aim of modelling a time series is to forecast according to a given model.
- Hence we will now suppose that we have a sensible model for which parameters were previously fitted.
- Write $X_{n+h}^{(n)}$ the forecast at lag h based on the past observations $x_1, \dots, x_n$.
- The estimator minimizing the mean squared error is, as usual,

$$X_{n+h}^{(n)} = \mathbb{E}(X_{n+h} \mid X_1, \dots, X_n).$$

- Here we will focus on linear estimators (which in the Gaussian case are optimal), i.e.,

$$X_{n+h}^{(n)} = \beta_0 + \sum_{j=1}^{n} \beta_j x_{n+1-j}, \qquad \beta_j \in \mathbb{R}.$$

- Beware the coefficients β_j depend on h and n.

Proposition 7. *Having observed* $x_1, \ldots, x_n$*, the best linear estimator* $X_{n+h}^{(n)} = \beta_0 + \sum_j \beta_j x_{n+1-j}$ *satisfies*

$$\mathbb{E}\{(X_{n+h} - X_{n+h}^{(n)}) X_{n+1-k}\} = 0, \qquad k = 1, \ldots, n.$$

Proof. Consider the least square problem and... □

□ Looking at the above expression when $h = 1$, we haven for $k = 1, \ldots, n$,

$$\sum_j \beta_j \gamma(k - j) = \gamma(k) \Longleftrightarrow \Gamma \boldsymbol{\beta} = \boldsymbol{\gamma}.$$

Forecasting at lag $h > 1$

- Obviously one can forecast at higher lags.
- Most estimators are obtained using an iterative scheme.
- I list here the most 2 popular ones:
  - Durbin–Levinson algorithm;
  - innovation algorithm proposed by Brockwell and Davis.
- I will not cover these algorithm but note that softwares perform forecasting on these two!

Forecasting the international airline passenger data set

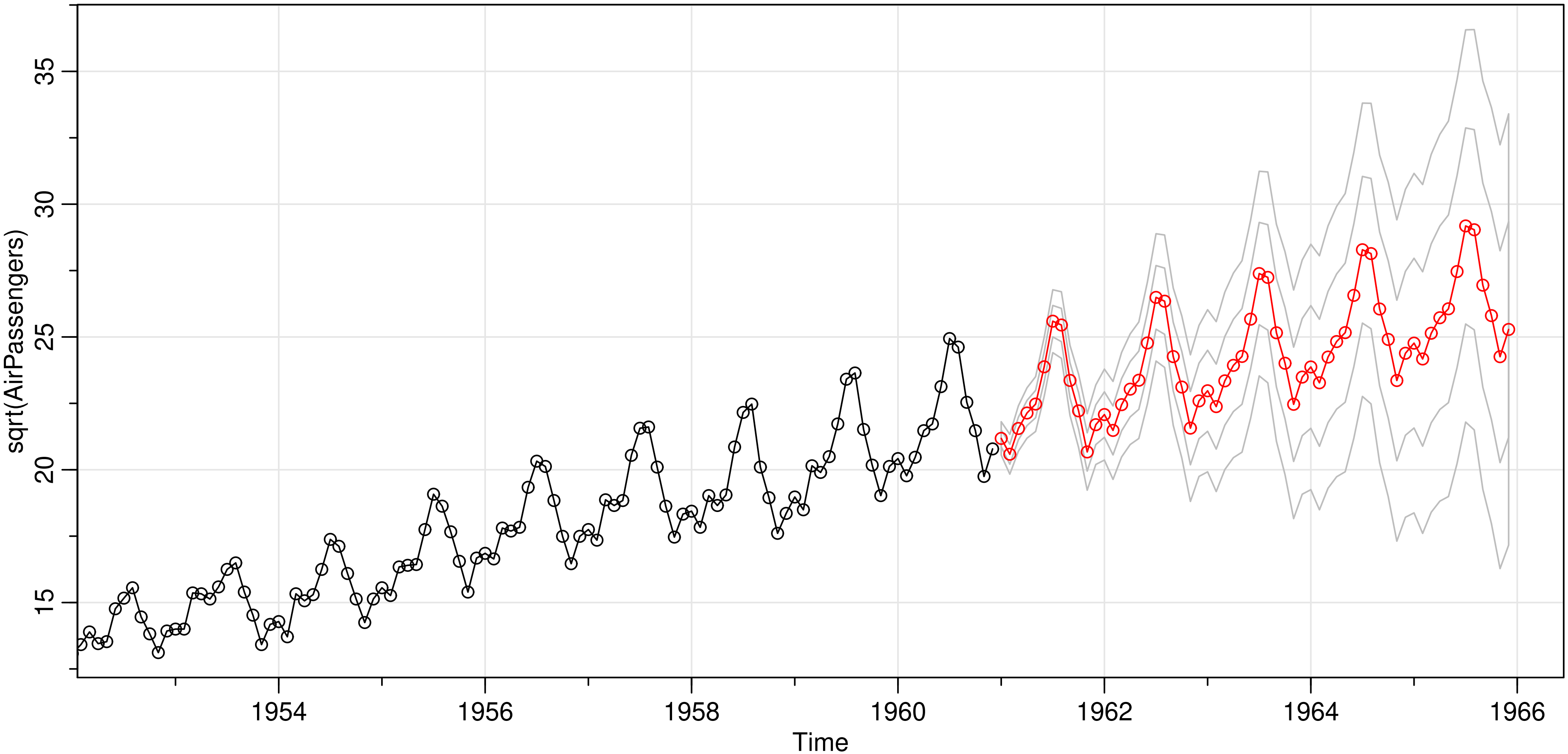


Figure 22: *Forecasting using a fitted $SARIMA$ for the next 5 years.*

STL decompostion (in 2 slides!)

- Sometimes it is necessary to model the time series in a two steps procedure:
  1. remove trend and seasonality
  2. fit a stationary time series model on the residuals
- More formally we write
$$Y_t = T(t) + S(t) + X_t,$$
  where $T(t)$ denotes trend, $S(t)$ denotes seasonality and X_t is a stationary time series, e.g., ARMA.
- The STL decomposition (Seasonal and Trend decomposition by Loess) is a widely used choice to estimate the functions T and S.

STL decomposition on the international airline passenger data set

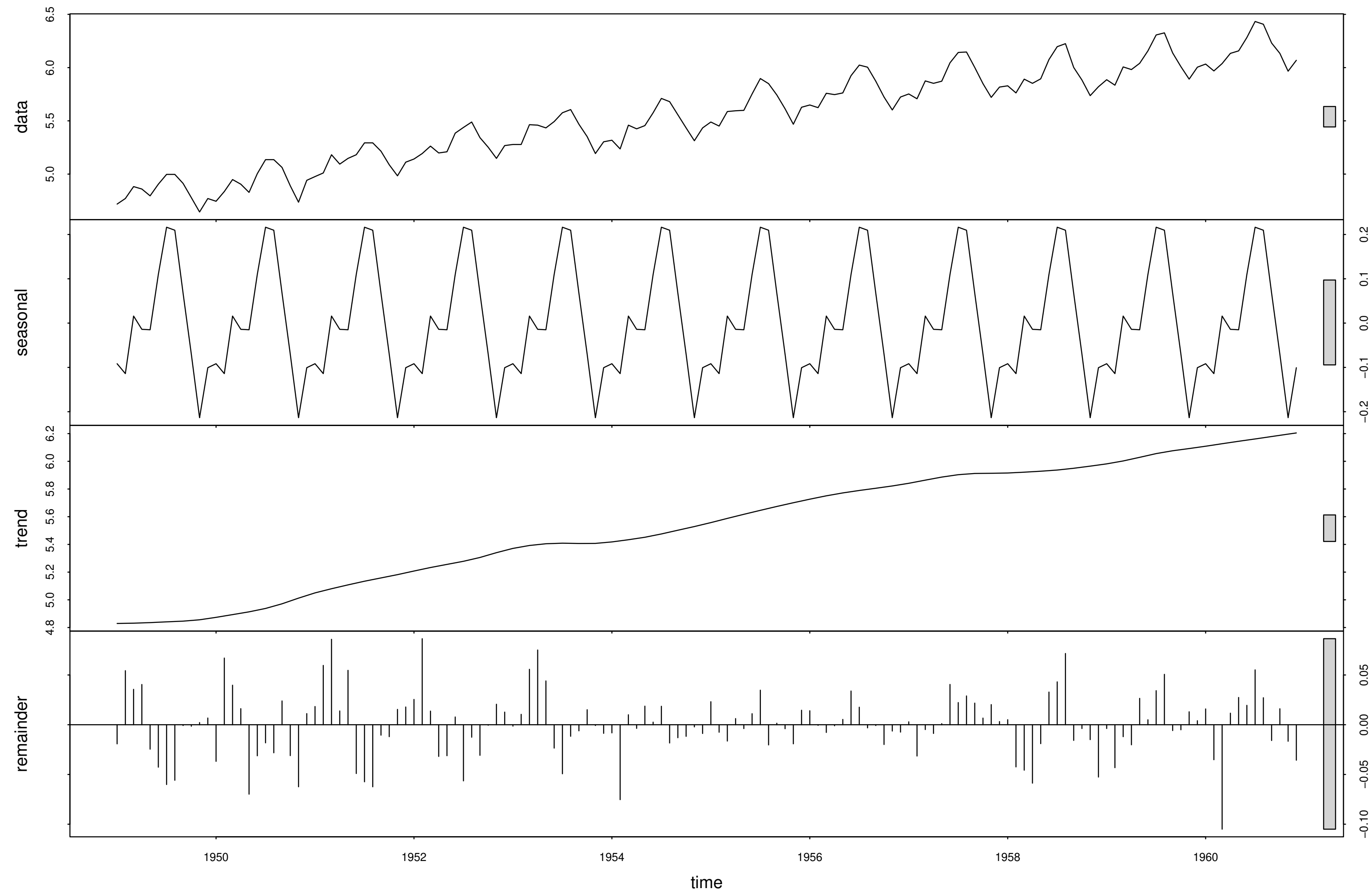


Figure 23: *STL decomposition on the logarithm of the data.*

Not enough time...

We were not able to talk about:

- Heteroscedastic models, e.g., $ARCH, GARCH$.
- Multivariate time series